
Dynamic Prediction, Mediation and Communication for Survival Outcomes, with Applications to Cystic Fibrosis

Kamaryn T. Tanner

Thesis submitted in accordance with the requirements for
the degree of Doctor of Philosophy of the
University of London
15 July 2021

LONDON
SCHOOL *of*
HYGIENE
& TROPICAL
MEDICINE



Department of Medical Statistics
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine

This thesis was self-funded with the exception of Chapter 5 which was funded by a Vertex Pharmaceuticals, Inc. Circle of Care award, awarded to the Cystic Fibrosis Trust and Ruth Keogh.

Declaration of Authorship

I have read and understood the School's definition of plagiarism and cheating given in the Research Degrees Handbook.

I, Kamaryn T. Tanner, confirm that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text and the appendix "Statement of Joint Work", and that this work has not been submitted for any other degree or professional qualification. I have acknowledged all results and quotations from the published and unpublished work of other people.

I have exercised reasonable care to ensure that the work is original and does not to the best of my knowledge break any UK law or infringe any third party's copyright or other intellectual property right.

Date: 15 July 2021

Name: Kamaryn T. Tanner

Acknowledgements

I wish to thank my supervisors Ruth Keogh, Linda Sharples and Rhian Daniel for their support and guidance. I feel that anything I say here will be inadequate to properly describe how much they have helped me. They kept me focused on the goal of completing this thesis and were always happy to give me the benefit of their experience. Just as valuable, however, were the times we spent laughing together and sharing stories about our lives. I genuinely looked forward to the weekly meetings with them and feel so grateful for their friendship.

Thanks also to the people in the Department of Medical Statistics at LSHTM for welcoming me into their community. I'd like to specifically acknowledge the help I received from my advisory committee: Karla Diaz-Ordaz, Nick Jewell and Stijn Vansteelandt.

I also appreciate the helpful suggestions received from the examiners at my up-grading: Bianca De Stavola and Geert Molenberghs.

In the CF community, I'm grateful for all of the interactions I've had with the CF Trust team, people with CF and CF clinicians, especially Rebecca Cosgriff, Tom Smith and Diana Bilton. I'm also indebted to the CF Trust for allowing me to use the UK Cystic Fibrosis Registry data throughout this thesis. Thanks also to the people with CF and their families who allowed their data to be captured in this registry and who shared their stories with us.

Special thanks to Cris and Nina for supporting me in going back to school and building a life in London. I know it wasn't always easy. Thank you Daniel, Isabel and Elena for cheering me on from New York.

Getting a PhD at this stage of life is not the traditional path. Thanks to everyone who believed in me. I am so fortunate to have had this opportunity.

Abbreviations and Notation

Abbreviation, Symbol	Description
<i>Part I: Dynamic Prediction of Survival</i>	
CF	Cystic fibrosis
FEV1%	Forced expiratory volume in 1 second as percent of predicted
IPCW	Inverse probability of censoring weighted
$i = 1, \dots, n$	i indexes individuals
C_i	censoring time for individual i
T_i^*	event time for individual i
T_i	observed time, equal to $\min(T_i^*, C_i)$
δ_i	indicator of whether the individual experienced an event ($\delta_i = 1$) or was censored ($\delta_i = 0$)
t_{hor}	fixed time horizon for prediction
$\pi_i(t_{hor} s)$	probability for individual i of survival to time t_{hor} conditional on survival to time s
X_i	set of time-fixed covariates for individual i
$\mathcal{Y}_i(s)$	longitudinal covariate history up to time s for individual i
$y_{ki}(t)$	value of the k^{th} longitudinal outcome for individual i at time t
$m_{ki}(t)$	true but unobserved value of longitudinal outcome k for individual i at time t
$h(t)$	hazard at time t
$h_{0,s}(t)$	baseline hazard at time t after landmark time s
$\tilde{y}(s)$	value of time-varying predictor(s) at landmark time s . If y_{ki} was measured at time s then $\tilde{y}_{ki}(s) = y_{ki}(s)$
$(a_1, a_2]$	notation describing a discrete-time interval beginning just after time a_1 and continuing up to and including time a_2
P_{li}	probability of individual i having an event in the discrete time period $(a_{l-1}, a_l]$ conditional on i being event-free up to time a_{l-1}
S_{li}	probability of survival to time a_l for individual i
$G(T_i)$	Kaplan-Meier estimate of the censoring distribution at individual i 's event time, T_i

Part II: Communication of Survival Predictions

PWCF	person/people with cystic fibrosis
HCP	health care professional

Part III: Mediation Analysis for Survival Outcomes

CFRD	Cystic fibrosis-related diabetes
BMI	Body mass index
DE	Direct effect, the effect of the exposure on the outcome not via the mediator
IE	Indirect effect, the effect of the exposure on the outcome via the mediator
TE	Total effect of the exposure on the outcome
$\lambda_{gh}(t)$	transition intensity from state g to state h
β_{gh}	regression parameters from transition from state g to h
$P_{gh}(u, t)$	probability of transitioning from state g to state h by time t given that they were in state g at time u
\mathcal{H}_u	event history up to time u
R	time of intermediate event (here, diagnosis of CFRD)
T	time of final event (here, death or transplant)
A	exposure
Y	outcome
M	mediator
M_0	baseline mediator measurement
L	time-varying confounder
\bar{M}_s, \bar{L}_s	history of the mediator, time-varying confounder up to time s
$Y_i(a)$	potential outcome of individual i if they had exposure $A = a$
$M_i(a)$	value of M if individual i had exposure $A = a$
$Y_i(a, M_i(a^*))$	potential outcome of individual i if they had $A = a$ and the mediator was set to the level it would have been if the exposure $A = a^*$
A^M	effect of the exposure on the mediator process (method of Aalen)
A^D	effect of the exposure directly on survival (method of Aalen)
$\lambda_i(t)$	hazard of having an event for individual i at time t
$\alpha_A, \alpha_M, \alpha_{Z_0}$	regression parameters in mediation analysis hazard model
β_A, β_{Z_0}	regression parameters in mediation analysis mediator model
$S_{A(a), M(a^*)}(t)$	survival probability at time t if the exposure $A = a$ and the mediator was set to level it would have been if the exposure $A = a^*$
$[t]$	visit time at or before time t
k	indexes visit times 1,2,3
Y_k	indicator of survival time greater than k ; $I(T > k)$
μ_M, Σ_M	mean and variance matrix for generation of random intercept and slope for simulated mediator values
U_0	unmeasured baseline confounder
κ, b	shape and scale parameters for Weibull distribution

$\alpha_A^s, \alpha_{M_{[t]}}^s, \alpha_{Z_0}^s$	parameter values used to generate simulated hazards
$\beta_A^s, \beta_L^s, \beta_{Z_0}^s$	parameter values used to generate simulated mediator data
$\psi_A^s, \psi_M^s, \psi_{Z_0}^s$	parameter values used to generate simulated time-varying confounder data
θ	true value of an estimand
$\hat{\theta}$	estimated value of an estimand
n_{sim}	number of simulated datasets

Contents

Introduction	2
1 Introduction	2
1.1 Overview and Aims	2
1.2 Cystic fibrosis	3
1.3 Survival prediction models in cystic fibrosis	5
1.4 Mediation analysis in cystic fibrosis	7
1.5 The UK Cystic Fibrosis Registry	8
1.6 Outline	8
I Dynamic Prediction of Survival	10
2 Dynamic Prediction Techniques: Review of Existing Methods and Introduction of a Machine Learning Approach	11
2.1 Introduction	11
2.2 Dynamic prediction	13
2.3 Standard dynamic prediction techniques	15
2.3.1 Joint modelling	15
2.3.2 Landmarking	18
2.3.3 Comparison of joint modelling and landmarking	19
2.4 Dynamic prediction using a machine learning ensemble	21
2.4.1 Discrete-time survival analysis	21
2.4.2 Super Learner	24
2.4.3 Super Learner landmark approach for dynamic prediction	27
2.5 Assessment of predictive performance	28
2.5.1 Validation procedure	28
2.5.2 Validation metrics	31
2.6 Discussion	36
3 Dynamic Survival Prediction Using the UK CF Registry	37
3.1 Introduction	37

3.2	Study population	37
3.3	Outcomes and predictors	38
3.4	Implementation of dynamic survival prediction methods	44
3.4.1	Creation of training and test datasets	44
3.4.2	Implementation of the joint model	44
3.4.3	Implementation of the traditional landmarking method	45
3.4.4	Implementation of the Super Learner landmarking method	47
3.4.5	Selection of algorithms and models for Super Learner	48
3.4.6	Performance measures	49
3.4.7	Software	51
3.5	Results	51
3.5.1	Cross-validated performance using the training dataset	51
3.5.2	Performance using the test dataset	59
3.5.3	Practical considerations	63
3.6	Discussion	63
4	Simulation Study: Comparison of Dynamic Prediction Methods	69
4.1	Introduction and aims	69
4.2	Data-generating mechanisms	70
4.3	Methods	73
4.4	Performance measures	73
4.4.1	Software	74
4.5	Results	74
4.6	Discussion	80
II	Communication of Survival Predictions	85
5	Communication of Survival Predictions	86
5.1	Introduction	86
5.2	Presentation of life expectancy information	90
5.2.1	Graphical displays	90
5.2.2	Uncertainty	92
5.2.3	Preservation of hope	93
5.3	Methods	93
5.3.1	Design and development of prototype life expectancy presentation	93
5.3.2	Design of semi-structured interview, recruitment and analysis	96
5.3.3	Development of web app	101
5.4	Results	102
5.4.1	Respondents	102

5.4.2	Feedback from people with CF	102
5.4.3	Feedback from health care professionals	107
5.5	Translation of results into life expectancy web app	111
5.6	Discussion	115
5.7	Contributions	118
5.8	Funding	119

III Mediation Analysis for Survival Outcomes 120

6 Cystic Fibrosis-Related Diabetes 121

6.1	Introduction	121
6.2	Cystic fibrosis-related diabetes	122
6.3	Cystic fibrosis-related diabetes and survival	123
6.4	Descriptive analysis of CFRD and survival in the UK CF Registry .	127
6.4.1	Study population	127
6.4.2	Methods and implementation	128
6.4.3	Results	132
6.5	Discussion	143

7 Investigation of Mediators of the Effect of CFRD on Survival 146

7.1	Introduction	146
7.2	Mediation	146
7.2.1	Overview of traditional mediation methods	146
7.2.2	Causal Estimands for Mediation	148
7.2.3	Mediation analysis for time-to-event outcomes	150
7.3	The Method of Aalen et al. (2020)	152
7.3.1	Overview	152
7.3.2	Setting and Estimands	152
7.3.3	Estimation	155
7.3.4	Assumptions	156
7.4	The Method of Vansteelandt et al. (2019)	156
7.4.1	Overview	156
7.4.2	Setting and Estimands	157
7.4.3	Estimation	159
7.4.4	Assumptions	161
7.5	Analysis of the UK CF Registry	161
7.5.1	Overview	161
7.5.2	Study population	162
7.5.3	Mediators, confounders and outcome	162
7.5.4	Implementation of the mediation methods	164

7.5.5	Results	168
7.6	Discussion	174
8	Simulation Study: Comparison of Mediation Methods for Survival Outcomes and Time-Updated Mediators	181
8.1	Introduction and aims	181
8.2	Data-generating mechanisms	181
8.2.1	Unmeasured confounding	185
8.2.2	Multiplicative hazards for event time generation	185
8.2.3	Baseline mediator affects the exposure	186
8.2.4	Time-varying confounders present	187
8.2.5	Infrequent mediator measurements	188
8.3	Methods	189
8.4	Estimands	191
8.5	Performance measures	192
8.5.1	Generation of the truth	193
8.6	Results	196
8.6.1	Overview and baseline scenario	196
8.6.2	Unmeasured confounding	199
8.6.3	Multiplicative hazards	199
8.6.4	Baseline mediator affects exposure	200
8.6.5	Time-varying confounders present	201
8.6.6	Infrequent mediator measurements	202
8.7	Discussion	203
	Discussion	209
9	Discussion	209
9.1	Key Findings	209
9.1.1	Dynamic prediction of survival	209
9.1.2	Communication of survival predictions	210
9.1.3	Mediation analysis for survival outcomes	211
9.2	Future Work	211
9.2.1	Dynamic prediction of survival	211
9.2.2	Communication of survival predictions	213
9.2.3	Mediation analysis for survival outcomes	214
9.3	Conclusions	214
A	Preparation of a Discrete-Time Landmark Super Dataset	216

B Super Learner Ensemble for Simulation Study	217
C Dynamic Survival Prediction Simulation: Reference Scenario	219
D Interview Topic Guide	221
E Mediation Analysis: BMI	228
F Mediation Simulation Parameters Used	230
G Mediation Simulation Results	234
H Statement of Joint Work	255

List of Tables

2.1	Algorithms and models used in the Super Learner ensemble	25
3.1	Deaths and lung transplants by calendar year	39
3.2	Predictors used in the final models of the 12 studies reviewed	40
3.3	Covariates from the UK Cystic Fibrosis Registry dataset used in the prediction models	42
3.4	Descriptive statistics for the study population at landmark ages 25, 35 and 45	43
3.5	The number of rows in the super dataset for varying numbers of discrete intervals.	47
3.6	Landmark super dataset before and after discretisation	48
3.7	Details of preliminary investigations into algorithm-hyperparameter combinations for Super Learner.	50
3.8	Brier Score for 2-year and 5-year dynamic survival prediction	56
3.9	Comparison of predictors between test and training datasets	59
3.10	Final prediction model	62
5.1	Personas used in development of prototype life expectancy presentation.	96
5.2	ID and time of diagnosis for the seven persons with CF interviewed.	102
5.3	Representative quotes from people with CF	104
5.4	Representative quotes from health care professionals	110
6.1	Sample person-years calculation	129
6.2	Descriptive characteristics for the incident cohort	134
6.3	Cox regression analyses of risk factors associated with death or lung transplant	136
6.4	Hazard ratio estimates for the three semi-parametric multi-state models	139
7.1	Data used in mediation analyses	165

8.1	Parameter values used in the mediator model and hazard model for the baseline scenario	184
8.2	Parameter values used for the scenarios with time-varying confounders	189
8.3	Listing of all simulation scenarios	190
8.4	Performance summary for methods of Vansteelandt and Aalen for mediation analysis	196
8.5	The empirical standard error and relative efficiency for the baseline scenario simulation analyses	197
8.6	The true effect and estimated absolute bias for the six simulation scenarios with unmeasured confounding	200
8.7	Estimates of indirect effect and proportion mediated with and without adjustment of M_0	205
B.1	Algorithm/hyperparameter combinations used in simulation study	218
F.1	Parameter values used to generate simulated data for the baseline scenario	231
F.2	Parameters used to generate simulated data	232
G.1	Simulation results for the baseline scenario	235
G.2	Simulation results for scenario U1	236
G.3	Simulation results for scenario U2	237
G.4	Simulation results for scenario U3	238
G.5	Simulation results for scenario U4	239
G.6	Simulation results for scenario U5	240
G.7	Simulation results for scenario U6	241
G.8	Simulation results for scenario W1	242
G.9	Simulation results for scenario W2	243
G.10	Simulation results for scenario W3	244
G.11	Simulation results for scenario B1	245
G.12	Simulation results for scenario B2	246
G.13	Simulation results for scenario L1	247
G.14	Simulation results for scenario L2	248
G.15	Simulation results for scenario L3	249
G.16	Simulation results for scenario L4	250
G.17	Simulation results for scenario L5	251
G.18	Simulation results for scenario L6	252
G.19	Simulation results for scenario F1	253
G.20	Simulation results for scenario F2	254

List of Figures

2.1	Dynamic prediction survival curves	14
2.2	Relationship between the longitudinal process and the hazard function in a joint model	16
2.3	The creation of sliding landmark datasets	20
2.4	Graphical representation of the discretisation of landmark data	23
2.5	Ten-fold cross-validation	30
2.6	Sample calibration plots	33
3.1	Number of annual reviews for individuals in the study population	38
3.2	FEV1% trajectories	41
3.3	Creation of training and test datasets	45
3.4	Observed and predicted FEV1% trajectories	46
3.5	Performance evaluation of the three candidate joint models	52
3.6	Composition of the Super Learner prediction function	53
3.7	Brier score by algorithm	54
3.8	Reduction in prediction error using cross-validation on the training dataset	55
3.9	Discriminative ability using cross-validation on the training dataset	57
3.10	Calibration plots	58
3.11	Reduction in prediction error using the test dataset	60
3.12	Discriminative ability using the test dataset	61
3.13	Brier score for SL landmark analyses with different loss functions	66
4.1	Transformed longitudinal variable versus the originally simulated longitudinal variable	72
4.2	Results for scenario 1/transformation A (Brier score)	76
4.3	Results for scenario 1/transformation A (C-index)	77
4.4	Results for scenario 2/transformation A	78
4.5	Results for scenario 1/transformation B	79
4.6	Results for scenario 1/transformation A (Cox landmarking with interactions)	81
4.7	Results for scenario 1/transformation A (no LMEM)	82

5.1	Collage of four publicly available online prognostic calculators . . .	88
5.2	Median predicted survival age, reproduced from the UK Cystic Fibrosis Registry Annual Data Report 2019	89
5.3	Prototype life expectancy presentation	97
5.3	Prototype life expectancy presentation (cont'd)	98
5.3	Prototype life expectancy presentation (cont'd)	99
5.3	Prototype life expectancy presentation (cont'd)	100
5.4	Prototype web app, landing page and overview page	112
5.5	Prototype web app, by age and by age and sex pages	114
5.6	Ideas for future communication work	118
6.1	Physiological processes associated with development of CFRD . . .	123
6.2	Classic illness-death model	130
6.3	Prevalence of CFRD in 2017	132
6.4	Estimated mortality/transplant rates for adults with CF	135
6.5	Multi-state model for the CFRD setting	137
6.6	Predicted transition probabilities	138
6.7	Predicted probability of transition to death or transplant	141
6.8	Predicted transition probabilities for the transition from no CFRD to death/transplant and from CFRD to death/transplant	142
7.1	DAG illustrating simple mediation	147
7.2	DAG showing single mediator plus confounders	150
7.3	Method of Aalen data generating mechanism	153
7.4	Dynamic path analysis	155
7.5	Method of Vansteelandt data generating mechanism	158
7.6	Path-specific effects in method of Vansteelandt	158
7.7	BMI trajectories	163
7.8	Sample timeline for data collection	164
7.9	Construction of the analysis dataset	166
7.10	Results from method of Aalen and method of Vansteelandt	169
7.11	Effect estimates using the method of Aalen with FEV1% as the candidate mediator	170
7.12	Survival curves estimated using the method of Vansteelandt with FEV1% as the candidate mediator	171
7.13	Effect estimates using the method of Vansteelandt with FEV1% as the candidate mediator	173
7.14	Effect estimates with IV days as the candidate mediator	175
8.1	Illustration of the data-generating mechanism	182
8.2	DAG illustrating unmeasured confounding scenarios	185

8.3	DAG illustrating scenarios where M_0 affects A	187
8.4	DAG illustrating scenarios with time-varying confounding	188
8.5	Survival curves for truth data	194
8.6	Diagnostic plots used to verify generated truth data	195
8.7	Effect estimates from the method of Vansteelandt for the Baseline- DE+IE scenario	198
8.8	Values of the time-varying confounder L for 60 simulated individuals	201
8.9	Results from 1,000 simulation datasets of scenario F1-DE+IE . . .	202
8.10	Comparison of effect estimates with and without individuals who did not survive to the first mediator measurement	206
C.1	Results for reference scenario/no transformation	220
E.1	Estimates of indirect effect of CFRD on mortality via BMI	229

Abstract

Patients with chronic diseases and their clinicians want accurate and up-to-date information about risk, prognosis, and survival. The overall aim of this thesis is to advance statistical methods available for providing such information. The methods are motivated by analysis of cystic fibrosis (CF), a genetic life-shortening disease, and illustrated using longitudinal data from the UK CF Registry.

First, dynamic models, that update predicted survival probabilities as new measurements become available, are studied. Although machine learning methods are established for prediction problems, they have not been widely used in dynamic survival prediction. Here, the combination of a machine learning ensemble with the landmarking approach is developed. Predictive performance of this method is compared to that of the most commonly-used statistical techniques: joint modelling and landmarking. A simulation study investigates cases where a machine learning ensemble may improve predictive accuracy.

This thesis then provides a review of literature on communicating survival predictions, focusing on preferred graphical formats, comprehension by a broad audience, and best practices in survival communication. Based on this literature and semi-structured interviews conducted by qualitative research partners, an on-line tool was created. This provides life expectancy information sensitively and according to an individual's characteristics.

In the final part of the thesis, CF-related diabetes (CFRD), a common comorbidity of CF, and its role in survival are investigated. Using multi-state models, the relationship between CFRD and survival is described. Mechanisms through which CFRD affects survival are explored using two methods that can accommodate longitudinal mediators for survival outcomes. Each method is applied to a stacked dataset, constructed in similar fashion to a landmark dataset, designed to maximally use the longitudinal registry data. A simulation study investigates the sensitivity of these two methods to model misspecification and data availability issues.

Introduction

Chapter 1

Introduction

1.1 Overview and Aims

Statistical models in clinical research aim to provide public health bodies, clinicians and patients with more information about illness or disease and shed light on how to better manage the condition at a population level and/or at an individual level. When the research question is about the time to an event, specialised models are required, in particular to accommodate censored or truncated follow-up, and there is an extensive literature on such methods. However, analysis of time-to event data continues to be an active area of research, and there remain many unknowns about how best to answer different types of research questions, with a key challenge being the integration of longitudinal data on covariates. Statistical models for survival outcomes are the focus of this thesis and, using disease registry data, I aim to investigate and develop techniques for prediction, causal effect estimation and communication of analysis results to a broad audience.

Disease registries provide an invaluable source of data for addressing a range of questions relevant in clinical research. A disease registry goes beyond a hospital-specific or locally established cohort of patients to provide a large sample size on a diverse population. Disease registries typically include longitudinal measurements made on patients over long periods of follow-up, alongside time-to-event data. This facilitates investigations of time-to-event outcomes, e.g. mortality, as well as investigations about the natural history of the disease, and how different longitudinal processes may be linked. Because they include a diverse population, i.e. not only those predisposed to participating in research or those who meet eligibility criteria, disease registries can help identify risk factors for mortality or comorbidity.

One of the great strengths of disease registry data, longitudinal data collection, offers great opportunity for clinical prediction models. Dynamic survival prediction models have the advantage of allowing predictions to be updated over time as

new information/data becomes available. However, there remain questions about how the data can be best used for dynamic prediction, and opportunities exist for further development of statistical methodology. The two established techniques for dynamic survival prediction, landmarking [Anderson et al., 1983, van Houwelingen, 2007, van Houwelingen and Putter, 2012] and joint modelling [Hogan and Laird, 1997, Tsiatis and Davidian, 2004, Rizopoulos, 2012], differ in their computational complexity and use of the longitudinal data. The application of machine learning techniques to dynamic survival prediction is relatively new. Its implementation is complicated in settings with both time-varying data and a time-to-event outcome, however, it has great potential for prediction without the constraints of an assumed data generating process.

Longitudinal data also enable investigations into the causal effects of exposures on time-to-event outcomes, including studies of long-term effects and how these may be mediated through intermediate longitudinal processes. These enable investigations into causal mechanisms that would not be feasible from randomized controlled trials, for example. However, causal modelling using longitudinal data invokes many challenges, and causal inference methods that enable estimation of how effects of exposures on time-to-event outcomes are mediated through intermediate longitudinal processes are still in their infancy.

The overall aim of this thesis is to develop and study methods for the analysis of survival outcomes using observational data and I focus on three areas: dynamic survival prediction models, communication of survival predictions, and mediation analysis for survival outcomes. The analyses are motivated and illustrated in the context of cystic fibrosis (CF) and make use of registry data from the UK Cystic Fibrosis Registry. A general review of CF and of the registry dataset is provided in the next sections.

1.2 Cystic fibrosis

Cystic fibrosis (CF) is a genetic condition caused by a mutation in both copies of the gene coding for the cystic fibrosis transmembrane conductance regulator protein, the CFTR gene. There are over 2,000 known mutations of this gene and a person's genotype can impact the severity of the disease as well as the treatments available [Elborn, 2016]. The protein encoded by this gene controls the movement of water into and out of cells; when it is defective due to mutations, cell surfaces of the gastrointestinal (GI) tract, lungs and other organs are not properly hydrated. In the lungs, the result is a thick mucus and frequent infections. Like other obstructive lung diseases, CF makes it difficult to completely exhale, meaning that air is left in the lungs at the end of each breath. In the GI tract, CF sufferers are unable to absorb and digest food normally due to mucus accumulation blocking

the transport of pancreatic enzymes. This can lead to poor nutritional status, a need to take digestive enzyme medication, and the need to consume more calories [Horsley et al., 2015]. Additionally, CF is associated with a range of comorbidities affecting a variety of organs and systems from gastro-oesophageal reflux disease to male infertility to cystic fibrosis-related diabetes (CFRD) [Ronan et al., 2017].

The incidence of CF in Caucasians is approximately 1 in 2500–4500 births [Fanen et al., 2014]. Since 2007, when national newborn screening for CF was implemented in the UK, the majority of CF diagnoses are made shortly after birth using a heel prick test. This screening is essential in reducing disease burden and avoiding missed diagnoses [Elborn, 2016]. In milder cases or asymptomatic childhoods, diagnosis may not be made until adulthood. For these people, diagnosis can be difficult because of the heterogeneity in disease expression or patients having a rare/unknown CFTR mutation [Simmonds, 2013]. Currently, more than 10,500 people in the UK, more than 30,000 in the US, and approximately 100,000 worldwide are affected by CF [UK Cystic Fibrosis Registry, 2020, Cystic Fibrosis Foundation, 2018, Cystic Fibrosis Trust, 2021].

Despite there being a single gene mutation that is both a necessary and sufficient cause of CF, the nature of the mutations alone does not explain the spectrum of disease seen [Simmonds, 2013]. Two individuals with identical CFTR mutations may have very different symptoms and lung function due to modifier genes, treatment adherence, age, and choices about exercise, smoking and diet [Fanen et al., 2014]. Mutations are grouped into functional classes based on their disease-causing mechanism [Rowe et al., 2005]. “Classic” CF is generally associated with mutations in functional classes I–III, pancreatic insufficiency, and a more severe disease that impacts multiple organs [Quon and Aitken, 2012, Simmonds, 2013, Fanen et al., 2014]. The most common mutation, known as Phe508del or F508del is a class II mutation found on at least one copy of the CFTR gene in 90% of people with CF worldwide [Boyle and De Boeck, 2013]. In England, 50.8% are F508del homozygous meaning they have two copies of the F508del mutation [UK Cystic Fibrosis Registry, 2019]. “Non-classic” or atypical CF is less common and is often characterised by pancreatic sufficiency, milder lung disease and possibly later diagnosis [Simmonds, 2013, Fanen et al., 2014].

The burden of treatment for people with CF is high and the disease is managed on a day-to-day basis with physiotherapy, exercise and a combination of medications including antibiotics, steroids and enzyme capsules. In a study of 204 adults with CF in the US, the median number of medications taken per day was 7 (range 0 – 20) and the mean number of minutes spent on treatments was 108 (standard deviation 58 minutes) [Sawicki et al., 2009]. In recent years, precision medicines (e.g. ivacaftor, lumacaftor/ivacaftor, and elexacaftor/tezacaftor/ivacaftor) have been developed for people with specific CFTR gene mutations and several trials have

shown these drugs lead to improvements in lung function and a number of other outcomes [Whiting et al., 2014, Wainwright et al., 2015, Ratjen et al., 2017]. In addition to these daily treatment regimens, people with CF repeatedly suffer from pulmonary exacerbations, which often result in admittance to hospital for intravenous antibiotics [Elborn, 2016]. More serious episodes may additionally require non-invasive ventilation. Pulmonary exacerbations mark a decline in health and are characterised by one or more symptoms such as: reduced lung function, weight loss, shortness of breath, respiratory infection, cough or fatigue [Stanford et al., 2021]. Waters et al. [2012] found that around half of the observed decline in lung function for CF patients was associated with exacerbations.

Recurring infection, inflammation and airway obstruction over time result in progressive damage to lung tissue, manifested by irreversible decline in lung function. At this stage, lung transplantation is an option. Due to the riskiness of the procedure and shortage of donor organs, patients are only referred for transplantation when their CF has become severe, they are at maximal medical therapy, and their FEV1% is below 30% or in a rapid, irreversible decline [Cardiothoracic Advisory Group on behalf of NHSBT, 2018]. Post-transplant, the patient is burdened with side effects from immunosuppressive drugs and possible complications such as primary graft dysfunction, increased risk of infection and acute rejection [Horsley et al., 2015].

While standards of care have improved, there is currently no cure for CF and most people with CF die from progressive respiratory disease [Fanen et al., 2014]. In the late 1960s, the median survival age was less than 20 in the UK and even in the mid 1980s, 5% of babies diagnosed with CF in the UK died within their first year of life [Dodge et al., 2007]. Today, most people born with CF will live well into adulthood. The median predicted life expectancy is 49 for babies born in 2019 in the UK [UK Cystic Fibrosis Registry, 2020] and 48 for babies born in 2019 in the US [Cystic Fibrosis Foundation, 2020].

1.3 Survival prediction models in cystic fibrosis

The earliest work in survival prediction in the context of CF made use of the Cox proportional hazards model. Hayllar et al. [1997] recognised that the ability to predict survival could help optimise the planning of lung transplantation and Aurora et al. [2000] studied people already referred for lung transplant assessment. In contrast to these relatively small sample studies, Liou et al. [2001] and Mayer-Hamblett et al. [2002] used multivariate logistic regression and registry data on over 5,000 and 14,000 patients in the US, respectively. Mayer-Hamblett et al. [2002] studied 2-year survival with a focus on selection of lung transplant candidates. The objectives of Liou et al. [2001] were broader and, in addition to lung transplant

selection, included informing counselling, evaluation of therapies and assisting with future research design. The multivariate Cox regression model of George et al. [2011] used an expanded set of predictors relative to its predecessors with the hypothesis that use of therapies such as long-term oxygen, non-invasive ventilation and recombinant DNase were predictive of survival. In Buzzetti et al. [2012], the Liou et al. [2001] model was validated on an Italian CF patient population of approximately 900 people. Clinically, this population differed in numerous respects to the US study population and 5-year survival was not accurately predicted. The authors then used multivariate logistic regression to create a model based on only 4 characteristics for 5-year survival prediction. Liou et al. [2020] performed a validation of Liou et al. [2001] in the US and found that while the original model continued to have good discriminative ability, performance could be enhanced by adjusting the intercept for improved mortality rates. Nkam et al. [2017] created a model for 3-year survival prediction using the French CF registry with the goal of updating prognostic modelling in light of increased life expectancy and a reduction in paediatric mortality.

While the majority of the early models for survival in CF utilised logistic regression or Cox regression with baseline values of the predictors, recent models have applied more sophisticated techniques. For example, the Aaron et al. [2015] predictive model using Canadian registry data was based on threshold regression. They hypothesize that two components explain a person’s health: normal disease progression and “shocks” caused by pulmonary exacerbations. Stanojevic et al. [2019] extended this model to include 2-year survival predictions as well as 1-year predictions and performed external validation of the model using the UK CF Registry dataset. Keogh et al. [2019b] applied a landmarking technique to UK CF Registry data to create a dynamic survival prediction model capable of predicting 2- to 10-year survival for an individual based on 16 predictors, some time-varying. Machine learning techniques were introduced to CF survival modelling in Alaa and van der Schaar [2018] and Lee et al. [2019] using large algorithmic frameworks designed to be operated as a black box. While Alaa and van der Schaar [2018] focused on static prediction of 3-year mortality, the method of Lee et al. [2019] was designed to incorporate longitudinal data for producing a dynamic survival prediction. The use of registry data with longitudinal information on thousands of people has helped facilitate the development of increasingly complex survival models and this will be discussed in depth in Part I of this thesis.

The ultimate goal of clinical prediction models is to create a model that is useful to clinicians and/or to patients. Clinical prediction models may be used to inform clinicians and their patients about what to expect in terms of their survival, disease progression, or recurrence. Time-to-event predictions could allow clinicians to evaluate different treatment regimes or care plans and may also be

helpful in answering patient questions. Another application for these models is to provide information directly to patients to help them in their medical decision-making. Taking the next step from academic research to use in practice requires solving technology hurdles as well as addressing implementation questions around the sensitivity of the information and the ability of people to understand the provided interface. There is no universal guidance available for translation of a predictive algorithm to a point-of-care tool or to a publicly available application. Each situation must be investigated and considered in light of the disease, the care setting, the emotional impact of the information and the desire/need for the information. I describe an example in the setting of CF in Part II of this thesis.

1.4 Mediation analysis in cystic fibrosis

Some of the central questions of interest in CF surround understanding of disease mechanisms that lead to lung function decline and lowered life expectancy. Clinical and laboratory research led to the finding that the CFTR protein functions as an ion channel and that mutations in the CFTR gene lead to problems with the function of this protein. However, questions still remain about the mechanisms involved in the cycle of infection and inflammation experienced by people with CF.

Statistical research using mediation analysis applied to observational data has also contributed to our understanding of the mechanisms involved in CF. For example, Collaco et al. [2016] used mediation analysis to partially explain the observed positive association between ambient temperature and lung function in people with CF. They found that this association was mediated by three respiratory pathogens that each individually accounted for between 13% and 31% of the association seen. Using a product of coefficients mediation analysis method, Schlüter et al. [2019] estimated the direct effect of CF on birth weight and the indirect effect of CF on birth weight via gestational age; they concluded that 40% of the total effect was mediated through gestational age. Using joint modelling of lung function and survival instead of a formal mediation analysis, Taylor-Robinson et al. [2020] estimated that lung function mediates 37% of the association between sex and survival in people with CF. These few examples illustrate the great potential statistical mediation models have to test hypotheses about the mechanisms for effects present in CF. Defining and estimating direct and indirect effects presents challenges in a setting with a longitudinal mediator and time-to-event outcome, but recent methods have been developed to enable this. In part III of this thesis, I will focus on two of these methods to investigate the mechanisms involved in the effect of CFRD on mortality.

1.5 The UK Cystic Fibrosis Registry

The UK Cystic Fibrosis Registry, administered by the Cystic Fibrosis Trust, contains more than 100,000 annual review records from just over 12,000 patients [Taylor-Robinson et al., 2017]. This represents over 99% of people with CF in the UK, so that results of analysis of this registry will be representative of the UK and similar CF populations. Each individual or their parent / guardian provides written consent to their care team for their data to be collected. The registry data includes demographic information (age, height, weight, education, employment, etc.), genotype, measures of lung health (forced expiratory volume, chronic infections by pathogen type), prevalence/incidence of complications and treatment information. Although not all data are complete for all patients, the registry contains information on hundreds of variables.

While some encounter-based data may be collected in the registry, this research exclusively uses data systematically collected at annual reviews because it is more complete. Health care for people with CF, including their annual review, takes place at specialist CF centres. Annual reviews are routine monitoring visits where patients have detailed discussions about their health and undergo a battery of tests including: spirometry, blood tests, glucose tolerance test, exercise test and a chest x-ray. Patients may also review their physiotherapy routine, diet, fitness regimen and mental health [NHS Oxford University Hospitals, 2019]. In addition to data arising from these test results, summary data from the past year is also added to the registry, e.g. aggregated data on respiratory infections from the previous year. The registry dataset is typical of complex longitudinal data sets in that it contains right-censored, left-truncated time-to-event data, a mixture of binary, continuous and categorical variables measured at baseline and/or at regular intervals, and some degree of measurement error and missingness. All of these characteristics require methodological choices to be made when conducting analyses.

1.6 Outline

This thesis contains three parts. Dynamic survival prediction is the topic of Part I. A method that combines landmarking formulated in discrete-time with a machine learning ensemble is proposed and compared empirically with traditional landmarking and joint modelling using data from the UK CF Registry. I also conduct a simulation study to explore when the machine learning ensemble approach may outperform a traditional landmarking analysis. Part II focuses on the communication and presentation of survival predictions to both people with cystic fibrosis and to clinicians. I developed a prototype life expectancy presentation that was trialled using semi-structured interviews. The aim was to understand the

desire for and impact of personalised life expectancy information and to determine best presentation methods. Part III moves from prediction of survival to studying causal mechanisms. I investigate mechanisms through which CFRD may affect survival using two recently proposed methods for mediation analysis in the setting of a time-to-event outcome and time-updated mediator. As these methods are both new, they have been implemented in limited settings and there is no prior work comparing them or providing suggestions on how to choose between them. To better understand the strengths and weaknesses of these mediation analysis techniques and to assist in interpreting our results, I conduct a simulation study to evaluate potential sources of bias in each method. Finally, the findings and contributions of this work are summarised in a discussion. Areas for future work are also highlighted. A series of appendices are included that provide additional detail and code for many analyses described in this thesis is available on github at <https://github.com/KamTan>.

Part I

Dynamic Prediction of Survival

Chapter 2

Dynamic Prediction Techniques: Review of Existing Methods and Introduction of a Machine Learning Approach

2.1 Introduction

Predictive models for time-to-event outcomes are used widely in medicine to identify individuals at elevated risk, to inform treatment strategies and to update patients about their prognosis. For people with life-shortening conditions, it is natural for them to want to know their short-term and long-term survival prospects as time goes on. Answering these questions requires going from a static prediction at diagnosis time to predictions that can be updated dynamically as new information becomes available. Data obtained longitudinally through electronic health records, patient registries and established cohorts, have brought opportunities to develop dynamic prediction models for large numbers of individuals. Recent examples of dynamic prediction models focused on: 10-year cardiovascular disease risk based on electronic health records [Paige et al., 2018], survival for people with cystic fibrosis using registry data [Keogh et al., 2019b], intervention-free survival for patients with aortic stenosis based on a cohort [Andrinopoulou et al., 2015] and survival based on breast cancer recurrence data in a French cohort [Lafourcade et al., 2018].

Two techniques for dynamic survival prediction with longitudinal data dominate the statistical literature: joint modelling and landmarking. Joint models model the relationship between longitudinal covariates and the time-to-event process. The most commonly used class of joint models is based on shared random

effects [Tsiatis and Davidian, 2004, Rizopoulos, 2012] but the longitudinal process may also be linked to the survival process via shared latent classes [Proust-Lima et al., 2014, Hickey et al., 2016]. In the landmarking approach a survival model is fitted from a series of time origins (landmarks) as a function of predictors measured up to the landmark time. Joint models are flexible and provide consistent predictions when correctly specified (i.e. the prediction at time $t + \Delta t$ can be derived from the prediction at t via the distribution of the longitudinal outcome between t and $t + \Delta t$) [Jewell and Nielsen, 1993, Rizopoulos et al., 2017], and development of statistical software has made the analysis feasible [Rizopoulos, 2010, 2016, Hickey et al., 2016, Philipson et al., 2018]. However, this approach can be computationally complex, particularly for large datasets and multiple longitudinal predictors [Rizopoulos et al., 2017]. In contrast, landmarking is straightforward to implement and computationally simple, but uses less information from the longitudinal covariate(s) [van Houwelingen and Putter, 2012, Rizopoulos et al., 2017].

Machine learning algorithms are gaining in popularity as tools for clinical prediction. Non-parametric machine learning methods assume no knowledge about the data generating process and include artificial neural networks, support vector machines, decision trees and their relatives, random forests, boosting and bagging [Breiman, 2001b]. Statistical models such as generalised additive models and penalised regression are often included in the broader category of statistical learning which encompasses a broad range of methods for learning from data [James et al., 2013a]. Ensemble approaches combine several algorithms which can include methods of both of the above types. In this thesis, I will include these ensemble techniques in the definition of machine learning even though some of the component algorithms are parametric statistical models. Recent applications of machine learning methods for clinical prediction include: assessment of delirium risk [Wong et al., 2018], 3-year survival for cystic fibrosis patients [Alaa and van der Schaar, 2018], mortality in coronary artery disease [Steele et al., 2018], and time to revision surgery after knee replacement [Aram et al., 2018]. These studies and others have compared machine learning methods to traditional methods and found mixed results with some finding superior performance for parametric statistical models and others finding improved results with various machine learning algorithms. A review of comparisons of machine learning and logistic regression for clinical prediction found no benefit to machine learning in comparisons the authors deemed to be unbiased [Christodoulou et al., 2019].

Most machine learning examples in the survival literature are designed to make a static survival prediction using baseline covariates, and there is a gap in knowledge about how such methods can be applied in the context of dynamic survival prediction. In this chapter I develop a machine learning approach for dynamic prediction of time-to-event outcomes using longitudinal data. In particular, I describe

how the landmarking concept can be used in conjunction with machine learning. The focus is on the Super Learner ensemble, which combines a set of user-specified algorithms, and for which software for implementation is available in R [van der Laan et al., 2007]. We exploit the fact that many machine learning algorithms have been designed for binary outcomes, and show how dynamic prediction can be performed via application of the methods across a series of short discrete-time intervals [Polley and van der Laan, 2011]. In Chapter 3, the methodology is motivated and illustrated using longitudinal data from the UK Cystic Fibrosis Registry. The machine learning approach is compared empirically with a traditional landmarking analysis based on Cox regression and with joint modelling. I discuss how predictive performance, discrimination and calibration of the different approaches is evaluated using cross-validation and through application to a holdout dataset.

2.2 Dynamic prediction

Static prediction models provide information about survival from the time origin using baseline covariates measured at or before this origin. The time origin could be a specific age if age is used as the time scale or the date of diagnosis or treatment if the time scale represents time on study. In many contexts it will be important to provide predictions from different time origins, for example to enable predictions for people who have lived longer. This may involve developing different models or may be incorporated into a combined model. In a prediction model for individuals diagnosed with a particular condition, it will be important to update predictions to condition on time since diagnosis. In the motivating example in CF, to be relevant for the subject, the model at a minimum needs to be updated by conditioning on the patient's current age.

In many datasets arising from patient monitoring, time-varying covariates will be part of the available dataset. For example, in CF, many important health indicators are captured annually such as lung function, presence of respiratory infections and body mass index. When making predictions from new time origins, we can incorporate these time-updated measures of predictors. An extended Cox model could be fitted to the longitudinal data by simply reformatting it into long format where each subject has one record per value of the time-varying covariate and start- and stop-time information corresponding to the measurement times. While this model will provide estimates of hazard ratios, it cannot be used for prediction because the distribution of future values of the time-varying covariates would be required.

In contrast to static prediction, dynamic models enable predictions to be updated at new times s conditional on new measurements (if available) and on the patient having survived to time s [van Houwelingen and Putter, 2012, Proust-Lima

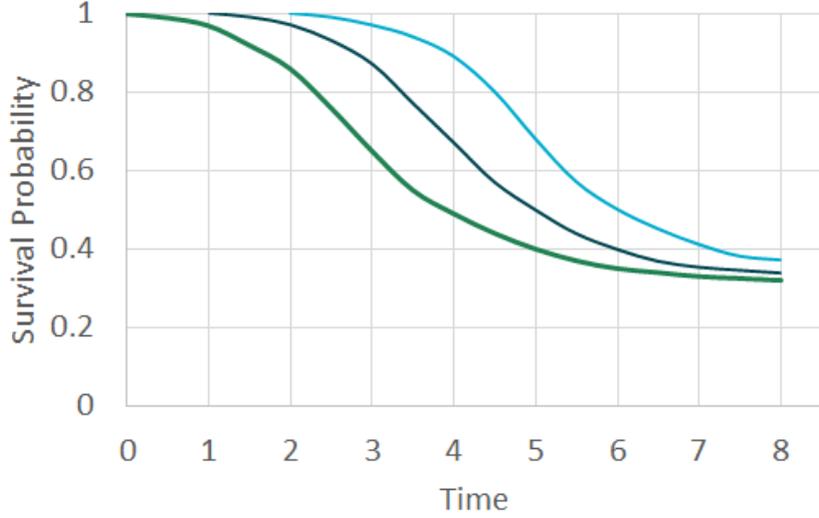


Figure 2.1: Using a dynamic prediction model, survival curves can be obtained for time $t=0$, but they can also be updated at time $t=1, 2$, etc.

and Blanche, 2016]. The aim of dynamic prediction is to estimate the probability of survival to some time horizon t_{hor} conditional on survival to time s , $s < t_{hor}$ and conditional on covariates measured up to time s . More generally, a survivor curve showing predictions conditional on survival to times s may be obtained as shown in Figure 2.1.

Let T_i^* and C_i denote respectively the event time and the censoring time for an individual i , ($i = 1, \dots, n$). The observed time is $T_i = \min(T_i^*, C_i)$ and $\delta_i = I(T_i^* \leq C_i)$ is an indicator of whether the individual experienced an event ($\delta_i = 1$) or was censored ($\delta_i = 0$). Let X_i denote a set of time-fixed covariates and $\mathcal{Y}_i(s)$ denote the longitudinal covariate history up to time s . The conditional survival probability of interest is

$$\pi_i(t_{hor} | s) = \Pr(T_i^* > t_{hor} | T_i^* > s, \mathcal{Y}_i(s), X_i) \quad (2.1)$$

As time passes, dynamic predictions are updated by adding the latest measurements at time s and calculating $\pi_i(t_{hor} | s)$. Special modelling techniques such as joint modelling or landmarking are required for dynamic prediction with updated longitudinal variables and I provide an overview of these methods in the next section.

2.3 Standard dynamic prediction techniques

2.3.1 Joint modelling

A joint model consists of a model for the survival process, a model for the true longitudinal process(es) (i.e. the time-varying predictors), and a specified link between them [Tsiatis and Davidian, 2004, Rizopoulos, 2012]. The most commonly used class of joint models is based on shared random effects. The original work on joint models focused on a single longitudinal process modelled using a linear mixed effects model [Hogan and Laird, 1997, Tsiatis and Davidian, 2004, Rizopoulos, 2011] but extensions have been proposed to accommodate multiple longitudinal variables [Lin et al., 2002, Chi and Ibrahim, 2006, Rizopoulos, 2016], and the focus is on this more general setting here. Consider k longitudinal outcomes and let $Y_{ki}(t)$ denote the value of the k^{th} outcome for individual i at time t . Because $Y_{ki}(t)$ is observed with measurement error, let $m_{ki}(t)$ represent the true but unobserved value of the longitudinal outcomes. A multivariate linear mixed effects model for the longitudinal outcomes is

$$Y_{ki}(t) = m_{ki}(t) + \epsilon_{ki}(t) \quad (2.2)$$

$$= W_{ki}^\top(t)\beta_k + Z_{ki}^\top(t)b_{ki} + \epsilon_{ki}(t) \quad (2.3)$$

where $W_{ki}(t)$ is the design matrix for the fixed effects β_k , $Z_{ki}(t)$ is the design matrix for random effects b_{ki} and $\epsilon_{ki}(t)$ are independent normally distributed errors conditional on the model covariates and random effects. The random effects b_{ki} are assumed to be independent of the error term and are assumed to follow a multivariate normal distribution with mean 0 and covariance matrix D . As described by Rizopoulos [2016], this model for the longitudinal outcomes may be extended to accommodate binary or categorical measures using different link functions.

The model for the longitudinal measures is linked to the survival process through a hazard model, with the hazard at time t assumed to depend on some function of the $\{m_{ki}(u) : u < t\}$, the unobserved true longitudinal outcome values up to time t , and a vector of time-fixed covariates, X_i . A simple form for the hazard model assumes that the hazard at time t depends on current values of $m_{ki}(t)$:

$$h_i(t) = h_0(t) \exp \left\{ \gamma^\top X_i + \sum_k \alpha_k m_{ki}(t) \right\} \quad (2.4)$$

where $h_0(t)$ is the baseline hazard at time t . Options for specification of the baseline hazard include leaving it unspecified (as in Cox regression), modelling it as a Weibull or other distribution, or using a flexible form such as B-splines or a piecewise constant function [Rizopoulos, 2011]. The model in equation 2.4 is a

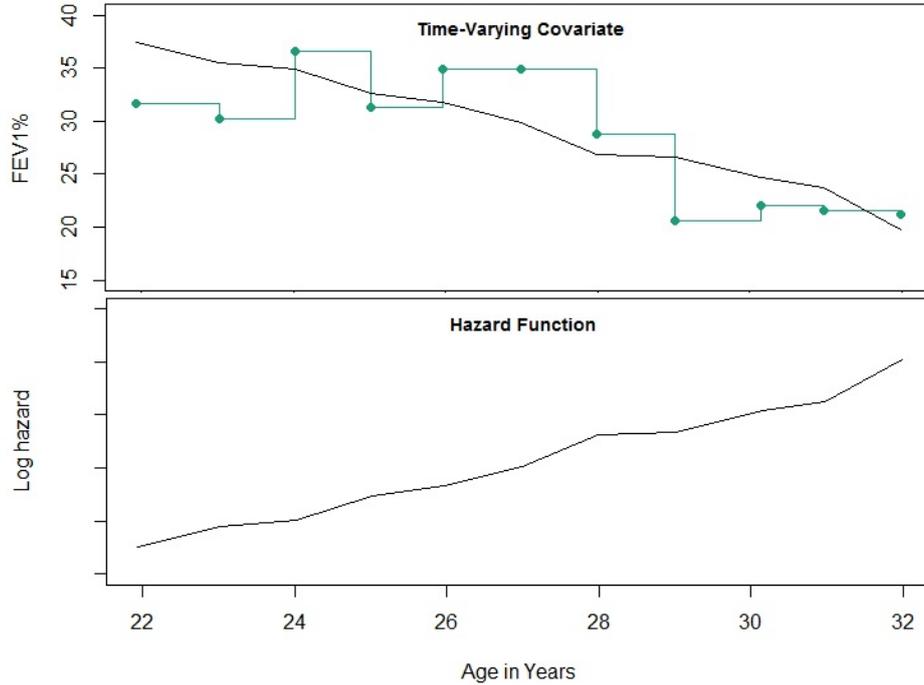


Figure 2.2: An example in the context of CF of the relationship between the longitudinal process and the hazard function in a joint model. The top panel shows observed values, $Y_i(t)$, of the time-varying predictor FEV1% as green dots. The green lines joining these dots illustrate how $Y_i(t)$ is approximated using an extended Cox model. The black line represents the joint model approximation of the true but unobserved values of FEV1%, $m_i(t)$. The bottom panel shows the jointly modelled hazard function over time.

proportional hazards model but extensions have been developed for stratified survival models and accelerated failure time models [Rizopoulos, 2012]. By replacing X_i with a time-varying covariate vector, the model above can be extended to incorporate other exogenous time-varying predictors. In this case, data would be formatted with start and stop times for each time-varying covariate. Figure 2.2 illustrates the relationship between the longitudinal outcome and the hazard in a joint model specified in this way. In the top panel, the black line represents the jointly modelled true longitudinal process, $m_i(t)$, and the corresponding hazard is plotted in the lower panel. The green dots in the upper panel represent the observed values of the time-varying predictor FEV1%, $Y_i(t)$.

A joint model formulated as in equation 2.4 assumes that the hazard of an event is dependent on the current value of the longitudinal outcome(s) but there may be reason to suspect that the risk depends on other aspects of the longitudinal

history. Two alternative parametrisations are: (1) allow for the hazard of an event to depend on the current value and the rate of change (slope) of the longitudinal outcomes at time t and (2) allow for the hazard of an event to depend on the cumulative effects of the longitudinal outcome histories. Following Rizopoulos [2012], these are both possible through a modification of the survival sub-model in 2.4. To incorporate the slope of the longitudinal trajectory, a new term is added to the hazard model:

$$h_i(t) = h_0(t) \exp \left\{ \gamma^\top X_i + \sum_k (\alpha_{1k} m_{ki}(t) + \alpha_{2k} m'_{ki}(t)) \right\} \quad (2.5)$$

where $m'_{ki}(t)$ is the derivative of the true value of longitudinal outcome k with respect to t and α_{2k} measures the association between the slope of outcome k and the risk of an event for constant $m_{ki}(t)$. The cumulative effects parametrisation is accomplished by including a term to capture the area under the longitudinal trajectory up to time t ,

$$h_i(t) = h_0(t) \exp \left\{ \gamma^\top X_i + \sum_k \alpha_k \int_0^t m_{ki}(s) ds \right\} \quad (2.6)$$

where the integral of the longitudinal trajectory measures the area and, therefore, represents the cumulative effects. This model may be further extended with a weighting function to allow more recent observations to have a greater weight in the measure. Additionally, all three of these survival sub-models can be extended to incorporate exogenous time-dependent predictors that are not incorporated as outcomes in the longitudinal model.

Estimation of the joint model parameters can be performed by maximum likelihood or by Bayesian Markov chain Monte Carlo. Individual predicted survival probabilities ($\pi_i(t_{hor} | s)$ in equation 2.1) are Monte Carlo estimates based on the posterior predictive distribution of the survival process [Rizopoulos, 2016]. Numerous software packages are available in R for implementing joint modelling including but not limited to: **JM** [Rizopoulos, 2010], **JMbayes** [Rizopoulos, 2016], **joineR** [Philipson et al., 2018], **joineRML** [Hickey et al., 2018b], **1cmm** [Proust-Lima et al., 2017] and **rstanarm** [Brilleman et al., 2018]. With large datasets such as those based on electronic health records or multiple longitudinal outcomes, computation is challenging and may even be infeasible [Paige et al., 2018]. This is because maximum likelihood estimation typically requires numerical integration and Bayesian approaches require an impractically large number of parameters to be sampled.

Joint latent class models are another type of joint model that link the longitudinal process to the survival process via shared latent classes. These models posit

that the heterogeneous population of patients is made up of homogeneous latent subgroups sharing the same longitudinal process characteristics and the same hazard of having an event [Proust-Lima et al., 2014]. A critical assumption within this framework is that, conditional on the covariates, the hazard of the event within a latent class is independent of the longitudinal measurement [Proust-Lima et al., 2014]. Joint latent class models allow a more flexible link between the longitudinal and survival processes than random effects models and do not require a homogeneous population [Proust-Lima and Blanche, 2016]. On the other hand, the required independence assumption may not hold and the models can be difficult to fit due to local optima and lack of convergence [Rizopoulos, 2011, Proust-Lima et al., 2017, Hickey et al., 2018a]. For these reasons, only shared random effects joint models will be considered here.

2.3.2 Landmarking

Landmarking was first described as a technique for estimating associations between time-dependent covariates and the hazard using Cox models as the foundation, as an alternative to Cox models with time-updated covariates [Anderson et al., 1983]. In the landmarking approach the dynamic prediction at a given landmark time is based on a model fitted only to those patients still at risk at the landmark time. Following van Houwelingen [2007], consider a landmark time s and a clinically relevant prediction time period, v , predictions of survival to time $t_{hor} = s + v$, conditional on survival and predictors up to time s , can be obtained based on the Cox proportional hazards model,

$$h_{s,i}(t \mid X_i, \mathcal{Y}_i(s), s, v) = h_{0,s}(t \mid s) \exp \{ \gamma_s^\top X_i + \alpha_s Y_i(s) \}, s < t \leq s + v \quad (2.7)$$

where $h_{s,i}(t \mid X_i, \mathcal{Y}_i(s), s, v)$ is the hazard at landmark time s and $Y_i(s)$ is the value of the time-dependent predictors at s . This analysis, called sliding landmarking, uses separate datasets, one per landmark time s . Each sliding landmark dataset is created by including only individuals at risk at time s and administratively censoring at $s + v$.

The basic landmarking approach incorporates longitudinal predictors in the sliding landmark dataset by taking the most recently measured value prior to landmark time s . To avoid loss of longitudinal information due to this “last observation carried forward” method, predictions of longitudinal measurements based on mixed modelling can be used [Paige et al., 2018]. Using these predicted values allows us to obtain estimates of Y_i at time s . Either way, the predictors appear in the landmark dataset as time-fixed covariates at time s . The proportional hazards model in (2.7) can be fitted separately for each landmark time $s = \{s_1, \dots, s_L\}$

leading to L separate models, however, this is likely to be inefficient. van Houwelingen [2007] proposed fitting a combined model across multiple landmark times with efficiency gained through assumptions such as common log hazard ratio parameters across landmark times. An example of a combined model, sometimes called a supermodel, is

$$h_i(t | X_i, \mathcal{Y}_i(s), s, v) = h_{0,s}(t | s) \exp \{ \gamma^\top X_i + \alpha Y_i(s) \} \quad (2.8)$$

where $h_{0,s}(t)$ is the baseline hazard at landmark time s . This differs from the model of equation 2.7 in that the baseline hazard is stratified by landmark time but the hazard ratios for both X_i and $Y_i(s)$ are common across landmark times. The model may be expanded to account for time-varying effects by letting the parameters depend on s and I refer to van Houwelingen and Putter [2012] for other extensions to the supermodel. The supermodel in (2.8) can be fitted to a dataset formed by vertically stacking the data created at each landmark time, the sliding landmark data sets, into one landmark super dataset with the baseline hazard stratified by landmark time s . Figure 2.3 illustrates the process of creating sliding landmark datasets from survival data.

Estimates of the predicted survival probabilities are obtained from the supermodel of equation 2.8 using

$$\pi_i(s + v | s) = \exp \{ -H_{0,s}(s + v | s) \exp(\gamma^\top X_i + \alpha Y_i(s)) \} \quad (2.9)$$

and Breslow’s estimate of the cumulative baseline hazard functions, $H_{0,s}(s + v | s)$. Landmarking is a less computationally intensive approach compared with joint modelling and can be implemented with standard software. A closely related paradigm is the partly conditional survival model which also uses landmark times but the time scale is reset to the time since measurement of the longitudinal variable [Zheng and Heagerty, 2005].

2.3.3 Comparison of joint modelling and landmarking

Joint modelling and landmarking take two different approaches to dynamic prediction and each has unique advantages and disadvantages. While joint modelling may be computationally complex, landmarking may not fully use the longitudinal information. That the landmarking model is not fitted using a full likelihood has also been cited as a drawback [van Houwelingen and Putter, 2012]. In both landmarking and joint modelling the survival model is typically a proportional hazards model [Ferrer et al., 2018]. Recent papers have compared the two approaches using simulation studies. Rizopoulos et al. [2017] demonstrated that joint models tend to outperform landmarking when the effect of time is correctly specified in the longitudinal sub-model. As misspecification increases, the differences become

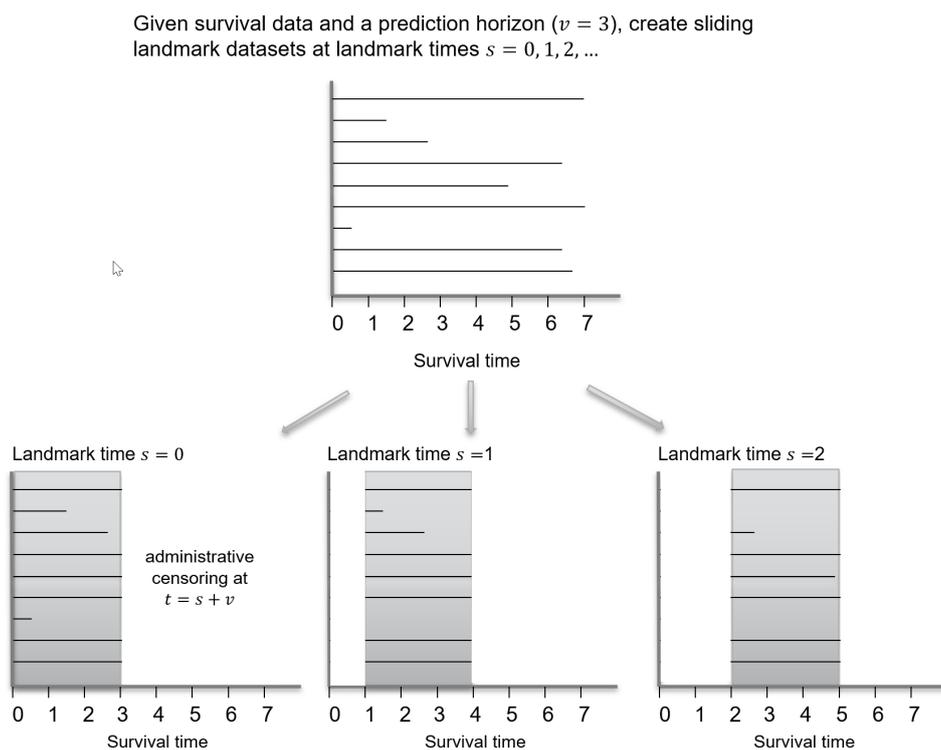


Figure 2.3: The creation of sliding landmark datasets. For each landmark time, a dataset is created with all those at risk at the landmark time and administrative censoring at time equal to the landmark time plus the prediction horizon. The procedure is repeated for each landmark time of interest and these may be vertically stacked to create a landmark super dataset.

smaller and landmarking can even outperform joint modelling. In contrast, a misspecified association structure between the longitudinal and survival processes did not substantially affect the relative performance of the two analysis methods. Ferrer et al. [2018] agreed that joint models outperform landmarking for a correctly specified joint model but noted that landmarking is less sensitive to a misspecified longitudinal process. Suresh et al. [2017] found that joint modelling provided better performance than landmarking for an illness-death model but the difference was quite small. Maziarz et al. [2017] reported that partly conditional models (landmarking-style models) offered comparable performance to joint models and were more computationally efficient.

2.4 Dynamic prediction using a machine learning ensemble

2.4.1 Discrete-time survival analysis

Most machine learning algorithms were originally conceived for predicting binary or continuous outcomes. Their use for prediction of time-to-event outcomes, particularly in the presence of right-censoring, generally requires either modification of the algorithm or manipulation of the data into a form suitable for applying techniques for binary outcomes. Some machine learning algorithms have been adapted for right-censored data and these include survival trees [Segal, 1988, 1997, Bou-Hamad et al., 2011], random survival forests [Ishwaran et al., 2008, 2014], support vector machines [Van Belle et al., 2011] and artificial neural networks [Ripley et al., 2004]. However, extensions to accommodate time-dependent measures of predictors are still limited and I am aware of only one, *Dynamic-Deep Hit*, a custom deep learning approach designed to learn survival distributions given longitudinal data, created for dynamic prediction [Lee et al., 2019].

In this work an approach is taken that enables us to exploit a large library of machine learning algorithms that are capable of estimating the conditional probability of a binary outcome without one-by-one modification of the algorithms for right-censored data. For this, the data are transformed into a discrete-time format such that each individual has a record in each of a series of short time intervals at which they are at risk.

I first outline the discrete-time approach in general terms, starting with the simplified setting of a single time origin, before extending to the dynamic setting. For discrete-time survival analysis, the follow-up time is divided into a sequence of d adjoining time periods from landmark time $s = a_0$ to the end of the prediction horizon $s + v = a_d$: $(a_0, a_1], (a_1, a_2], \dots, (a_{d-1}, a_d]$. In discrete-time, the “hazard” in period $(a_{l-1}, a_l]$ is the conditional probability of an individual having an event in

time period $(a_{l-1}, a_l]$ given that the individual was event-free up to time a_{l-1} and given time-fixed covariates X_i and the values of the time-dependent covariates at the time origin $\mathcal{Y}_i(a_0)$ [Allison, 1982]:

$$P_{li}(X_i, \mathcal{Y}_i(a_0)) = \Pr\{a_{l-1} < T \leq a_l \mid T > a_{l-1}, X_i, \mathcal{Y}_i(a_0)\} \quad (2.10)$$

Because these are conditional probabilities, the survival probability is given by:

$$S_{li}(X_i, \mathcal{Y}_i(a_0)) = \Pr(T > a_l \mid X_i, \mathcal{Y}_i(a_0)) = \prod_{j=1}^l (1 - P_{ji}(X_i, \mathcal{Y}_i(a_0))) \quad (2.11)$$

This discrete-time formulation is general and any method for computing the probability of a binary event could be applied. I refer the reader to Tutz and Schmid [2016] for a thorough overview of discrete-time survival analysis. A standard fully parametric statistical approach to estimating the conditional probabilities in (2.10) is through a logistic regression model [Cox, 1972, D’Agostino et al., 1990]:

$$\text{logit } P_{li}(X_i, \mathcal{Y}_i(a_0)) = \theta_l + \gamma^\top X_i + \alpha Y_i(a_0) \quad (2.12)$$

where θ_l ($l = 1, 2, \dots, l$) is a set of parameters capturing the baseline hazard in each discrete interval. The model can be fitted in a pooled way across all discrete time periods. In the formulation in (2.12), the coefficients for X_i and $Y_i(a_0)$ are assumed constant across time periods (i.e. proportional hazards), but more generally they could be allowed to be time-dependent. The baseline hazard may be constrained to some particular shape, restricted via groupings, or, in the most general case, left to take on any shape by allowing a separate parameter in each time period [Singer and Willett, 1993]. Non-parametric machine learning approaches may also be used to estimate the conditional probabilities in (2.10) [Malley et al., 2012]; model-free estimates of the conditional probabilities of the binary response for each discrete interval are estimated and, as in (2.11), are multiplied to obtain the predicted v -year survival probability.

Extension to dynamic prediction

The discrete-time analysis outlined above can be extended to the dynamic prediction setting with time-dependent covariates by adapting landmarking for discrete-time. The discrete-time equivalent of the landmark supermodel for dynamic prediction in equation (2.8) is:

$$\text{logit } P_{li}(X_i, \mathcal{Y}_i(s), s) = \theta_{s,l} + \gamma^\top X_i + \alpha Y_i(s) \quad (2.13)$$

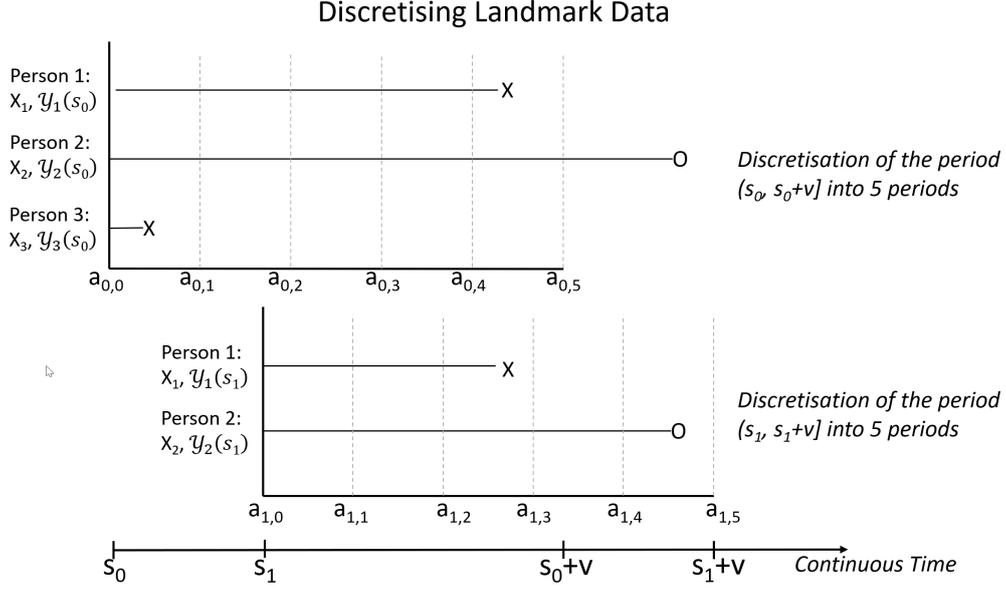


Figure 2.4: Graphical representation of the discretisation of landmark data. Event and censoring times are indicated by X and O symbols, respectively. From landmark times s_0, s_1 , the prediction period is divided up into 5 discrete intervals and the event status of each person in each period is recorded. For landmark times s_0 and s_1 , Person 1 will have an event indicator equal to one in time periods $(a_{0,4}, a_{0,5}]$ and $(a_{1,2}, a_{1,3}]$, respectively, and an event indicator equal to zero for all other periods that they were at risk. Because Person 2 is at risk of an event at the start of all 5 periods for both s_0 and s_1 , they will have a record for each interval with an event indicator equal to zero. Person 3 will only have one record, the record corresponding to the event in period $(a_{0,0}, a_{0,1}]$.

Using the relationship from (2.11), the corresponding predicted v -year survival probability is:

$$\pi_i(t_{hor} = s + v \mid X_i, \mathcal{Y}_i(s), s) = \prod_{l=1}^d (1 - P_{s,li}(X_i, \mathcal{Y}_i(s), s)) \quad (2.14)$$

where d is the number of discrete intervals between s and $s + v$.

The landmark dataset for each landmark time s_1, \dots, s_L is separately discretised and values of time-dependent covariates at the landmark time are used. Figure 2.4 illustrates the discretisation of a landmark super dataset. Step-by-step instructions for the preparation of a discrete-time landmark super dataset can be found in Appendix A. The analyses using the Super Learner ensemble described below are all based on the application of models or machine learning algorithms for binary outcomes to the discretised landmark super dataset.

2.4.2 Super Learner

The discrete-time hazard in equation (2.10) can be estimated using fully parametric models (e.g. logistic regression), semi-parametric methods (e.g. generalized additive models), or machine learning algorithms that estimate the conditional probability of a binary outcome [Malley et al., 2012]. The performance of different machine learning algorithms and statistical models will differ across applications, depending on the features of the dataset and the interrelations between covariates and between the covariates and the outcome [van der Laan and Rose, 2011]. Ensemble learning, using a combination of different algorithms or statistical models rather than just one, is designed to remove the dilemma of how to choose a single method. In an ensemble learning system, a library of independently fitted (or “trained”) algorithms are each used to predict the target – here, the conditional survival probability – then the ensemble combines the predictions from the component algorithms based on pre-defined rules or the results of another algorithm.

The Super Learner is a machine learning ensemble developed by van der Laan et al. [2007] that has been implemented in R and SAS. It is in the class of ensembles described as “stacking” algorithms, which employ a separate learner to optimally combine the predictions from the library of algorithms [Wolpert, 1992, LeDell, 2016]. The Super Learner is underpinned by theory showing that, given a bounded loss function, it will perform asymptotically as well as the best individual algorithm and asymptotically as well as the (unknown) optimal combination of learners [van der Laan and Dudoit, 2003, van der Laan et al., 2007].

Here I outline the general use of the Super Learner, as described by Polley et al. [2011], before extending to the discrete-time survival setting [Polley and van der Laan, 2011]. Finally, I describe my extension to the dynamic discrete-time survival setting. The inputs to the Super Learner algorithm are the data on outcomes and predictors, a user-specified list of algorithms (or “learners”) to use, and a loss function for quantifying the prediction error. Table 2.1 provides a brief description of each algorithm and model used in the application to CF data, which are: random forest, gradient boosting, support vector machine, generalised linear models (GLM), lasso, elastic-net GLM, Bayesian GLM, and generalised additive models (GAMs). These algorithms were selected to provide error “diversity” while being accessible to statisticians and analysts. Additional algorithms are available and, in theory, almost any algorithm could be used within the Super Learner framework by creating a wrapper function to facilitate data exchange between the code for the algorithm and the Super Learner framework. When choosing a list of algorithms, a diverse set of algorithms should be selected for optimal performance [Brown et al., 2005]. Candidate learners may include both different algorithms and multiple versions of the same algorithm with different tuning parameters or different subsets of data [Brown et al., 2005, Polley et al., 2011, LeDell, 2016].

Table 2.1: Algorithms and models used in the Super Learner ensemble. For each one, the name of the R package used, a description, and references are provided. See James et al. [2013a] for coding examples in R for many of these algorithms.

Algorithm <i>R Package</i> (Reference)	Description
Random forest <i>ranger</i> [Ho, 1995, Breiman, 2001a, Wright and Ziegler, 2017]	Multiple trees built on multiple bootstrapped training samples. At each step, a random sample of predictors (approximately the square root of the number of predictors) is considered, so that correlation between trees is reduced. Random forests may identify interactions between predictors.
Gradient boosting <i>xgboost</i> [Freund and Schapire, 1999, Friedman, 2001, Chen et al., 2018]	Builds multiple trees sequentially. Learns slowly by using residuals of the current model to fit a shallow tree. Incorporation of a shrinkage parameter further slows the updating of the current model.
Support vector machine <i>e1071</i> [Cortes and Vapnik, 1995, Meyer et al., 2018]	A binary classification algorithm that allows for non-linear boundaries of separation between two classes. The covariate space is expanded using kernels for efficient computation.
Generalized linear model <i>stats</i> [McCullagh and Nelder, 1989, R Core Team, 2020]	A class of statistical models that allow for binary, categorical and count data outcomes. Logistic regression is a specific type of GLM for binary data using a logit link between the outcome and the linear predictor.
Generalised additive model <i>gam</i> [Hastie and Tibshirani, 1990, Hastie, 2018]	Extends a regression problem with multiple predictors by allowing for a nonlinear relationship between each predictor and the outcome using smoothing splines or local regression. GAMs facilitate the modelling of nonlinear relationships.
Lasso and elastic-net GLM <i>glmnet</i> [Tibshirani, 1996, Friedman et al., 2010]	Fits a GLM via penalised likelihood which shrinks the coefficient estimates towards zero. The lasso uses an L_1 penalty. An L_2 penalty is known as a ridge regression and the elastic-net penalty allows for a combination of L_1 and L_2 penalties. These algorithms may improve predictive ability by decreasing variance at the expense of some bias.
Big lasso <i>biglasso</i> [Zeng and Breheny, 2017]	An algorithm for penalised regression that allows for multi-gigabyte data by avoiding storage of the dataset in memory and implementing new screening rules.
Bayes GLM <i>arm</i> [Gelman and Su, 2018]	GLMs fitted using Bayesian techniques. Requires specification of prior distributions for the regression coefficients.

The Super Learner involves fitting each of the specified algorithms to the data in a cross-validation procedure, and deriving the optimal combination of algorithms to minimize the prediction error. The procedure is described in detail below. The prediction error can be quantified in different ways, giving rise to different so-called “loss functions”. A commonly used loss function for regression problems is the squared error loss, which is defined as the squared difference between the actual outcome and the predicted outcome. Loss functions are discussed further in the next section in the context of applying the Super Learner to discrete-time survival analysis.

Given data with a binary outcome and selection of a squared error loss function, the Super Learner algorithm uses the following steps [Polley et al., 2011]:

1. The data are split randomly into V folds (e.g. $V = 10$). By removing each fold in turn as a test set and retaining the remaining $V - 1$ folds as a training set, this provides V training sets and V corresponding test sets. Note that each of the n individuals will be included in exactly one of the test sets.
2. Each of the Q individual algorithms is fitted to each of the V training sets and predicted outcomes are obtained from each algorithm for each individual in the corresponding V test sets. Let Ψ be the $Q \times n$ matrix of predicted conditional survival probabilities where Ψ_{ij} is the predicted outcome for individual i using algorithm j .
3. A vector of length Q of optimal weights, ω , for each algorithm is determined by finding the values of ω that minimise the expected value of the selected loss function \mathcal{L} over the n individuals. By restricting to a convex combination of ω and selecting a squared error loss function, the minimisation problem to determine the optimal weights is formulated as a non-negative least squares problem:

$$\begin{aligned} \min \quad & \frac{1}{n}(\omega\Psi - \Pi)^2 \\ \text{s.t.} \quad & \omega \geq 0, \quad \sum \omega = 1 \end{aligned} \tag{2.15}$$

In practice, we first solve for the non-negative weights and then rescale so they sum to one.

4. The final predictions are then obtained by fitting each learner to the complete data and using the weights ω , from the previous step, to combine them. In other words, the predicted conditional survival probability for individual i is a weighted combination of the predicted conditional survival probabilities from each individual algorithm.

2.4.3 Super Learner landmark approach for dynamic prediction

To use the Super Learner ensemble for dynamic survival prediction allowing for time-dependent predictors, we formulate the problem as one of predicting binary outcomes in short time periods defined through the discrete-time landmark super data set-up outlined in section 2.4.1. We also must select a loss function to measure the predictive performance and determine the final Super Learner prediction function.

In a standard setting of predicting binary outcomes, common choices for loss functions include the squared error loss function and the negative log loss function [Polley et al., 2011]. For a discrete-time survival analysis, we define these loss functions with respect to the conditional hazard in a given interval, (equation 2.10) which equals the conditional probability of having an event in that interval [Polley and van der Laan, 2011]. Because our interest is in measuring the loss on the conditional survival function, we use a squared error loss function that accommodates censoring via inverse probability of censoring weights (IPCW). I supplemented Super Learner with such an IPCW squared error loss function on the v -year survival probability. It uses IPCWs to adjust the squared error loss for the information lost due to censoring, by re-weighting the individuals who were not censored. This loss function is the squared difference between actual v -year survival and predicted v -year survival, multiplied by an IPCW and then summed across all observations present in each landmark time risk set R_s :

$$\begin{aligned} \mathcal{L}_{v,IPCW} = & \\ & \sum_{s=s_1, \dots, s_L} \frac{1}{n(s)} \sum_{i \in R_s} \{ (0 - \hat{\pi}_i(s+v | X_i, \mathcal{Y}_i(s), s))^2 I(T_i \leq s+v, \delta_i = 1) (1/\hat{G}_s(T_i)) \\ & + (1 - \hat{\pi}_i(s+v | X_i, \mathcal{Y}_i(s), s))^2 I(T_i > s+v) (1/\hat{G}_s(s+v)) \} \end{aligned} \quad (2.16)$$

Here, $\hat{G}_s(t)$ represents the Kaplan-Meier estimate of the censoring distribution at time t . It is estimated separately for each $s = s_1 \dots, s_L$ using data from risk set R_s . This equation shows that we are computing a sum across landmark times in which each individual contributes one value per discrete interval per landmark time that he or she is at risk. The first term captures the loss for those who had an event prior to $s+v$ (actual v -year survival probability = 0) and the second term captures the loss for those known to have survived past $s+v$ (actual v -year survival probability = 1). Because this loss function compares actual and predicted v -year survival it is a more appropriate measure of performance than the squared error loss on the conditional hazard. We use the v -year IPCW squared error loss to determine the weights in the final Super Learner prediction function by minimising

$\mathcal{L}_{v,IPCW}$ for an algorithm-weighted combination subject to the algorithm weights being non-negative. Details on the implementation using R and availability of sample code are given in section 3.4.7.

2.5 Assessment of predictive performance

A crucial aspect of the development of any prediction model is assessment of its predictive performance. To judge validity, we may ask the questions: Is there evidence suggesting that the predictions produced by the model are adequate for its intended use? [Altman and Royston, 2000] Is the model valuable for the specified set of patients? [van Houwelingen, 2000] Does the model work acceptably for “new” patients, i.e. patients not present in the dataset used to fit the model? [Altman and Royston, 2000, Altman et al., 2009] Because the goals of a predictive model are to inform patients and their families about prognosis and assist clinicians in decision-making, it is only valuable if it can successfully predict on new data – data collected at new time points or on new individuals. A structured process for validation of the model will yield information about its ultimate value.

There are two distinct aspects of model validation: the *procedure* for validating the model and the *metrics* used to validate the model. The validation procedure relates to the choice of data used to fit the model versus the data used to validate it and may include internal, temporal and external validation, as well as bootstrapping, cross-validation and test datasets. The chosen procedure is then applied to the model by implementing a set of performance measures typically covering the areas of calibration, discrimination and overall performance.

2.5.1 Validation procedure

Altman and Royston [2000] outline a validation procedure consisting of three progressively more rigorous tests beginning with internal validation, continuing to temporal validation and ending with external validation. Each of these is described in the next sections.

Internal validation

Internal validation is essentially validation using the original dataset. “Apparent validation” is fitting a model to the entire dataset and then measuring performance of the in-sample predictions made on that same dataset [Steyerberg and Harrell, 2016]. There are three fundamental problems with this approach. First, it is measuring something irrelevant to our research goal [James et al., 2013b]. The survival of the patients in the original dataset is known; predicting their survival

does not achieve the goal of informing patients or helping them to make decisions. A useful model will accurately predict survival in data that has not been seen. Second, the error rate measured in an apparent validation is usually over-optimistic [Efron, 1986]. The final problem with apparent validation is related to the first; because it is measuring an irrelevant quantity, it cannot be used to accurately compare different models. In fact, apparent validation will always select the most flexible model as the best model, the one with the lowest predictive error [James et al., 2013b]. This is because more flexible models tend to fit the noise in the given dataset which is rewarded when looking at in-sample predictions but may lead to poor performance on new data. James et al. [2013b] show that for any dataset and any set of statistical methods, as model flexibility increases, the mean-squared error of in-sample predictions will decrease but the mean-squared error of out-of-sample predictions may not.

Fortunately, other methods exist for internal validation. The simplest approach is to randomly split the original dataset into two parts, one for model development and one for model validation. Suppose the original dataset is split so that 50% of the observations are used to develop the model and 50% are used to test it. One criticism of this approach is that because the model is developed on only half of the data, it won't perform as well as a model developed with a larger number of observations [Steyerberg and Harrell, 2016]. Also, the validation results are highly subject to variance in composition of the two datasets, i.e. a different split of the data could lead to very different performance [James et al., 2013c]. The method recommended by Steyerberg et al. [2001] for predictive logistic regression models is bootstrapping; using a standard bootstrap procedure, models can be developed in a bootstrap sample and then tested against the original dataset.

A final option for internal validation is cross-validation. Cross-validation is a resampling method designed to estimate the predictive model's performance on unseen data. The dataset is randomly divided into k folds, where typical values of k are 5 or 10. One fold is used as the validation data and the model is trained on the remaining $k - 1$ folds. Predictions are made on the validation data and the prediction error is measured. The procedure is then repeated k times so that each fold is used once for validation and the error is averaged over the k folds. Figure 2.5 illustrates the 10-fold cross-validation process. Cross-validation is computationally expensive because the model is being fit k times instead of once. Another reason for limiting the value of k is the bias-variance trade-off. While $k = n$, known as leave-one-out cross-validation, offers the greatest bias reduction, it suffers from greater variance as the measured prediction error between folds will be highly correlated; $k=5$ and $k=10$ have been shown to balance the trade-off between bias and variance well [James et al., 2013c].

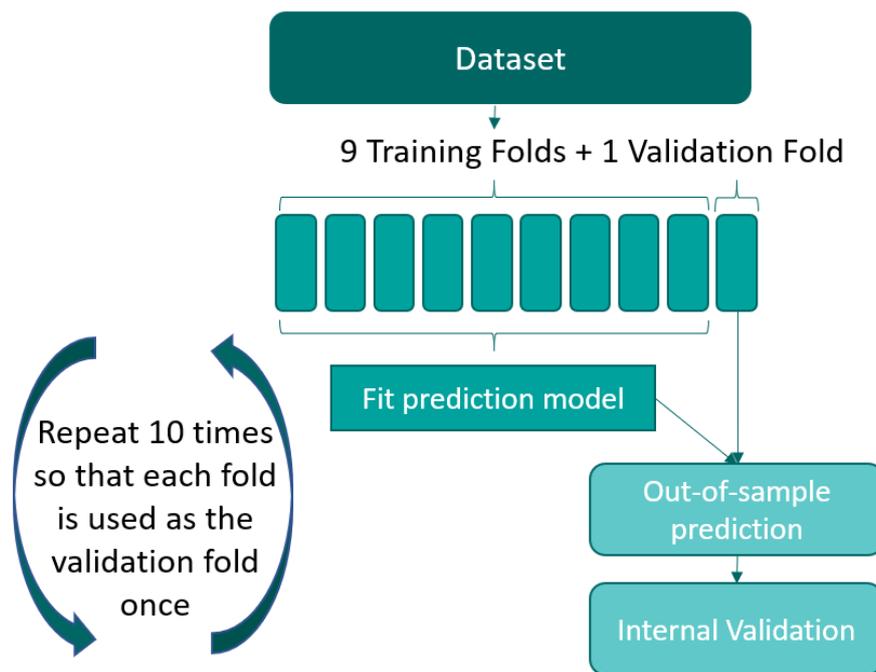


Figure 2.5: Ten-fold cross-validation. The dataset is randomly split into 10 folds. Each fold is removed once to serve as the test data for a model fit on the remaining 9 folds. Performance is measured by averaging across the 10 folds.

Temporal validation

Temporal validation is the evaluation of the model on future subjects from the same centres/regions/countries as the original dataset. In theory, this could be accomplished by simply splitting the original dataset by observation date but it is considered more stringent than internal validation because the validation dataset occurs at a different time than the dataset used for model development and is therefore independent in time [Altman and Royston, 2000].

External validation

What sets external validation apart from the previous two methods is that it seeks to assess the generalisability of the model to new populations. This is generally implemented using data not available at the time of model development. Examples could include data from other centres, countries or regions or data from different patient populations such as adults versus children [Debray et al., 2015]. When the external dataset is very similar to the original dataset, as assessed by comparing descriptive data, the external validation may be measuring reproducibility rather than generalisability [Steyerberg and Harrell, 2016]. Debray et al. [2015] proposed a framework for interpreting the results of an external validation study. They recommend first quantifying the similarity of the original and external datasets by looking at subject characteristics and predictor effects and then assessing performance in both datasets to understand how they differ. This information is then used to assess whether the validation is measuring the performance of new data from the same population or new data from a related yet different population.

The validation procedure should be created with respect to the research question. If the goal is a predictive model of survival in adults in France, there is no need for an external validation procedure using a dataset of children or a dataset of adults in New Zealand, but a validation using different hospitals in France would be valuable.

2.5.2 Validation metrics

Having reviewed validation procedures, I now turn to the second aspect of model validation, choice of metrics for validation. There are three areas of predictive performance to be quantified: calibration, discrimination and overall performance [Steyerberg, 2009]. In a survival setting with censored data, evaluation along these three dimensions is more complicated and there are multiple measures available depending on the assumptions that can be made about the censoring process. Each is briefly reviewed in the following sections.

Calibration

Calibration captures how well the predicted outcomes match the observed outcomes and is a measure of prediction accuracy. For example, if we predict an 80% probability of surviving 5 years for a group of patients, we should observe approximately 80% of patients in that group surviving to 5 years or more. Calibration is often assessed graphically by splitting the observations up into risk groups, obtaining the expected survival in each group using the model and then comparing this to the observed Kaplan-Meier survival in each risk group [Royston and Altman, 2013]. When plotted, the points from a perfectly calibrated model will form a 45-degree line reflecting the model’s ability to accurately predict at all levels of risk.

Figure 2.6 shows the calibration plots from three hypothetical predictive models for 3-year survival. For this example, individuals were divided into 10 risk groups based on predicted probability of an event within 3 years. In each graphic, the 3-year predicted survival probability versus the 3-year Kaplan-Meier survival probability for each group is shown as a green circle with a line connecting them. A grey 45-degree line is provided for reference. Model 1 on the left reflects a well-calibrated model with the calibration line closely matching the 45-degree line. The plot for Model 2 in the centre indicates a well-calibrated model for lower survival probabilities but poor calibration at higher survival probabilities. Above a 60% survival probability, Model 2 under-predicts the survival probability. The rightmost graphic, Model 3, depicts the case of miscalibration in the large where survival probabilities are systematically over-predicted, regardless of level [Royston and Altman, 2013].

Discrimination

In a binary setting, discrimination refers to the model’s ability to discriminate between subjects with and without the event and it is typically assessed by computing the area under the Receiver Operating Characteristic (ROC) curve. To plot the ROC curve, the sensitivity and specificity are calculated. Sensitivity is defined as $\Pr(\hat{q}_i > c \mid \Upsilon_i = 1)$ and specificity is defined as $\Pr(\hat{q}_i \leq c \mid \Upsilon_i = 0)$ where \hat{q}_i is the predicted probability of individual i having an event, Υ_i is the binary outcome and c is a cut-off level for classifying the prediction as having or not having the event. The ROC curve is then drawn by plotting sensitivity versus 1–specificity for the complete range of values of c . An area under the curve of 1.0 indicates perfect discriminative ability whereas a value of 0.5 signals no discriminative ability.

In a survival context, measures of discrimination must also be able to account for censoring and the fact that a person who survives one month and a person who survives five years have provided very different information. A predictive survival

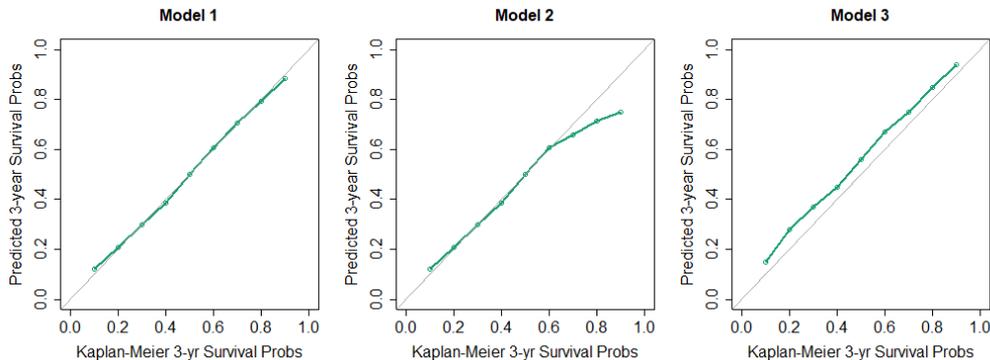


Figure 2.6: Sample calibration plots. Subjects were stratified into 10 groups based on 3-year survival probability and green points show predicted versus Kaplan-Meier survival probabilities. (Left) A well-calibrated model approximates the line $y = x$. (Centre) A model may be well-calibrated at some survival probabilities but poorly-calibrated at others as shown here by a bend in the line away from $y = x$ at higher survival probabilities. (Right) A model may show systematic errors in prediction of survival probabilities when the calibration plot is entirely above or below the $y = x$ line.

model should be able to discriminate between those who survived a short time and those who survived a long time [Pencina et al., 2012]. Traditional sensitivity and specificity measures have been extended for the censored time-to-event data case by Heagerty and Zheng [2005], enabling the creation of a time-dependent ROC curve. Instead of an individual having a binary outcome, in their model, survival time is considered to be a time-varying binary outcome. They then develop different measures of sensitivity and specificity depending on how “cases” (incident or cumulative) and “controls” (static or dynamic) are defined. Further extensions for longitudinal data have been proposed. For example, Zheng and Heagerty [2007] calculate sensitivity and specificity using a prediction rule based on the most recent longitudinal measurement; Rizopoulos [2011] uses the ROC framework adapted for joint modelling to evaluate the discriminative ability of the longitudinal predictor.

Discrimination may also be measured by the concordance statistic (C-statistic) which, for binary outcome analyses, is the area under the ROC curve. It is defined as the probability that in a given pair of subjects where one has the event and one does not, the one having the event was predicted to have the event with greater probability than the other [Hanley and McNeil, 1982]. For survival data, comparisons can be made between pairs of individuals only when one is known to survive longer than the other. So, for example, two censored individuals cannot be compared nor can two individuals be compared when one was censored prior

to the other one having an event. A pair is concordant if the individual with the longer observed survival time T was given a higher predicted survival probability π . Harrell's C-index uses actual event times to compare concordant and discordant pairs for prediction of events prior to some time of interest, τ , and is defined as:

$$C_{Harrell}(\tau) = \frac{\sum_{i \neq j} \{I(\pi_i > \pi_j) I(T_i < T_j, T_i < \tau) I(\delta_i = 1)\}}{\sum_{i \neq j} \{I(T_i < T_j, T_i < \tau) I(\delta_i = 1)\}} \quad (2.17)$$

where i and j index unique individuals. However, Harrell's C-index has been shown to be biased in the presence of censoring [Harrell et al., 1982, Pencina and D'Agostino, 2004, Pencina et al., 2012].

Uno's C-index attempts to overcome this bias by means of inverse probability of censoring weighting [Uno et al., 2011]. Uno's C-index is defined as:

$$C_{Uno}(\tau) = \frac{\sum_{i \neq j} \{I(\pi_i > \pi_j) I(T_i < T_j, T_i < \tau) I(\delta_i = 1) G(T_i)^{-2}\}}{\sum_{i \neq j} \{I(T_i < T_j, T_i < \tau) I(\delta_i = 1) G(T_i)^{-2}\}} \quad (2.18)$$

where $G(T_i)$ represents the Kaplan-Meier estimate of the censoring distribution at individual i 's event time, T_i . Pencina et al. [2012] showed that it is unaffected by the censoring distribution when censoring can be said to be independent of the covariates.

Gerds et al. [2013] extended the C-index further for the case when censoring is not independent of the covariates. Their proposed truncated C-index captures the ability of the model to discriminate or rank event times occurring prior to the truncation time t . Uno's C-index is a special case of the truncated C-index which is defined as:

$$C_{truncated}(t) = \frac{\frac{1}{m^2} \sum_{i,j} \{I(\pi_i > \pi_j) I(T_i < T_j, T_i \leq t) I(\delta_i = 1) W_{ij}(T_i)^{-1}\}}{\sum_{i \neq j} \{I(T_i < T_j, T_i \leq t) I(\delta_i = 1) W_{ij}(T_i)^{-1}\}} \quad (2.19)$$

In this equation the inverse probability of censoring weight, $W_{ij}(T_i)$, is defined as $G(T_i | X_j)G(T_i- | X_i)$ where $G(T_i | x) = P(C_i > t | x)$ and $G(T_i- | x) = P(C_i \geq t | x)$.

Overall performance measures

Overall performance measures evaluate a distance between predicted and actual outcomes with a smaller distance indicating greater predictive accuracy. Standard goodness-of-fit measures for continuous outcomes like R^2 fit into this category but it is important to distinguish the evaluation of performance using the same data, as in goodness-of-fit, from the evaluation of performance using new data or cross-validation [Steyerberg et al., 2010]. Numerous R^2 -style measures have been

proposed for survival data and good reviews are found in Choodari-Oskooei et al. [2012], Rahman et al. [2017]. I will focus on two related performance measures, the Brier score and the prediction error curve. The Brier score measures the mean-squared error for the special case where the actual outcomes are binary and the predicted outcomes are a probability. Ignoring censoring, the Brier score at time t^* is defined as:

$$BS(t^*) = \frac{1}{n} \sum_{i=1}^n \{I(T_i > t^*) - \hat{\pi}(t^* | X_i)\}^2 \quad (2.20)$$

where $I(T_i > t^*)$ is 1 if subject i 's event time is greater than t^* , 0 otherwise and $\hat{\pi}(t^* | X_i)$ is the predicted survival probability of subject i with covariates X_i at time t^* [Graf et al., 1999]. If nothing were known about survival and all predicted event-free probabilities were $\frac{1}{2}$, the Brier score would equal 0.25. To incorporate censoring, the observations are weighted according to their inverse probability of censoring using Kaplan-Meier estimates of censoring probabilities [Graf et al., 1999].

$$BS_{IPCW}(t^*) = \frac{1}{n} \sum_{i=1}^n \{(0 - \hat{\pi}(t^* | X_i))^2 I(T_i \leq t^*, \delta_i = 1)(1/\hat{G}(T_i)) \\ + (1 - \hat{\pi}(t^* | X_i))^2 I(T_i > t^*)(1/\hat{G}(t^*))\} \quad (2.21)$$

Again, $\hat{G}(t)$ represents the Kaplan-Meier estimate of the censoring distribution at time t . The first term of equation 2.21 captures the contribution from the estimated predictions for those who had an event prior to t^* while the second term includes those who were neither censored nor had an event prior to t^* . Subjects who were censored prior to t^* only contribute to the IPCWs.

The output of this calculation is a number representing the predictive accuracy of the model at time t^* . One criticism of the Brier score is that the resulting score is difficult to interpret unless it is presented relative to something else. van Houwelingen and Putter [2012] suggest presenting a percent reduction in prediction error which is calculated as the percentage change in Brier score between the predictive model of interest and a Kaplan-Meier model.

Because dynamic prediction involves obtaining predictions from multiple time origins (landmark times), the model must be assessed at multiple time points. One option is the integrated Brier score which averages the Brier score over time via integration. However, this single number does not provide information about how predictive accuracy changes across the landmark times. Another option is to produce a prediction error curve showing the predictive accuracy measured by the Brier score or by percent error reduction, at each time point [Gerds and

Schumacher, 2007]. Note that overall performance measures such as the mean-square error or Brier score can also be thought of as loss functions that measure the cost of the predicted values being different from the observed values. Loss functions are discussed in section 2.4.3.

2.6 Discussion

In this chapter, I have provided an overview and comparison of the two dominant techniques for dynamic survival prediction: joint modelling and landmarking. I then described a new framework for dynamic survival prediction using a machine learning ensemble, the Super Learner. In this approach, discrete-time survival analysis is combined with landmarking to allow the use of any model or algorithm capable of predicting the conditional probability of a binary outcome. Predictions from multiple methods may be combined via the machine learning ensemble to provide the best predictive performance. Procedures and metrics for validating prediction models were also reviewed. In the next chapter, I empirically compare the three prediction methods discussed using data from the UK CF Registry. Specifically, calibration plots, the truncated C-index and the Brier score are used in both a cross-validation procedure and with predictions on a test dataset. These validation methods are also used in a simulation study in Chapter 4.

Chapter 3

Dynamic Survival Prediction Using the UK CF Registry

3.1 Introduction

The focus of this chapter is on the application of the techniques described in the previous chapter to dynamic prediction of survival in CF. Using data from the UK CF Registry, I first provide an overview of the study population and the outcomes and predictors used in the analysis. This is followed by detailed information on how the three dynamic prediction techniques are implemented to produce 2-year and 5-year dynamic survival predictions for people with CF. In particular, information is provided about the construction of the discretised stacked landmark dataset used by the Super Learner landmark approach. Results for calibration, discrimination and predictive performance for each of the three methods are presented with the aim of providing a thorough empirical comparison of the methodologies. The chapter concludes with a discussion.

3.2 Study population

The study sample consisted of all individuals in the UK CF Registry who had an annual review between 1/1/2005 and 31/12/2015 and were 16 years of age or older at the time of the review. Further, any individual without genotype information and without at least one measurement of forced expiratory volume in 1 second as percentage of predicted (FEV1%), forced vital capacity as percentage of predicted (FVC%), body mass index (BMI), and weight from one annual review were omitted (3% of individuals). The resulting dataset consisted of 6,363 unique individuals with 43,880 annual review records. 962 (15%) of these had the composite event of either death or lung transplant.

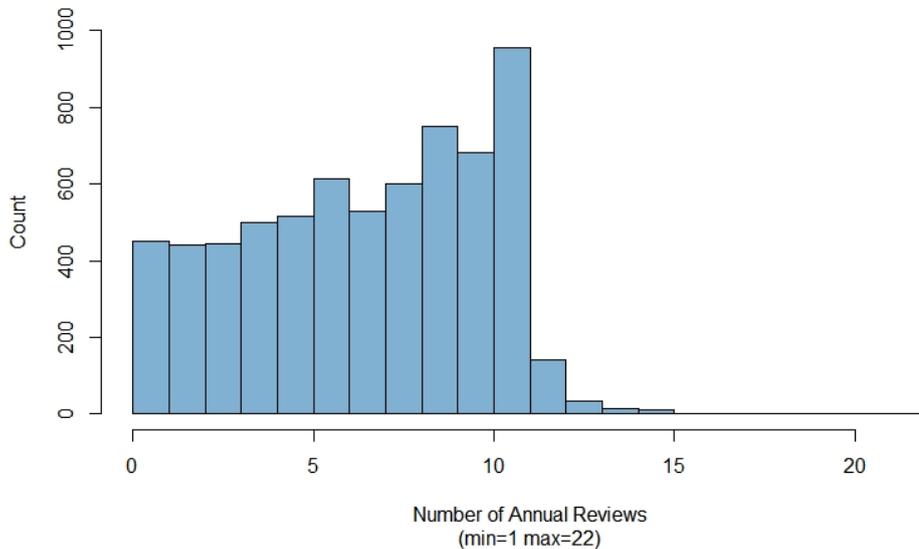


Figure 3.1: A histogram showing the number of individuals in the study population (2005 - 2015) having 1, 2, ..., 22 annual reviews in the dataset.

The number of annual reviews recorded in the dataset for each subject indicates the number of longitudinal observations available for that subject. Figure 3.1 summarises the number of annual reviews in the dataset for individuals in the study population. Seven percent of individuals had only a single visit recorded in the registry while 55% had seven visits or more recorded. 955 individuals (15%) had eleven separate longitudinal measurements in the dataset, roughly equating to one for each year in the sample.

3.3 Outcomes and predictors

The key outcomes for the analysis of dynamic survival prediction are 2-year and 5-year survival for individuals aged 20 – 50 years. Predictions were made from age 20 to confine the study to the adult population as the natural history of CF is different in children. Predictions longer than 5 years or for ages greater than 50 years were not attempted because at older ages, the small number of individuals in the dataset may lead to a model that is not well-calibrated [Keogh et al., 2019b]. Survival is taken to be a composite outcome of all-cause death or lung transplantation. This composite outcome was chosen because lung transplantation occurs when patients are considered to have short life expectancy and there are different risk factors for

Table 3.1: Information on number of deaths and lung transplants by calendar year in the study population. The number of patients represents the number of patients who had an annual review in the given calendar year.

<i>Calendar Year</i>	<i>No. patients</i>	<i>No. deaths</i>	<i>No. lung transplants</i>	<i>% Events that were transplants</i>
2005	2870	88	22	20%
2006	2406	68	13	16%
2007	2913	77	20	21%
2008	3210	71	24	26%
2009	3765	85	23	22%
2010	4128	92	30	25%
2011	4550	77	43	37%
2012	4667	101	35	26%
2013	4808	98	47	33%
2014	5107	105	35	26%
2015	5167	181	14	7%

death post-transplant. Post-transplant, the patient is burdened with side effects from immunosuppressive drugs and possible complications such as primary graft dysfunction, increased risk of infection and acute rejection for which data was not available [Horsley et al., 2015]. Additionally, post-transplant, FEV1% can increase from below 30% to over 75%, making post-transplant predictions not comparable to pre-transplant ones. [Ochman et al., 2019]. Table 3.1 provides information about the number of deaths and lung transplants for each calendar year in the dataset. From 2005 to 2011, lung transplants accounted for 23% of all events, where an event is defined as death or lung transplantation.

Revisiting the 12 CF survival prediction models reviewed in the Introduction (section 1.3), Table 3.2 lists the predictors used in the final models of each and the number of models that include each predictor. The only predictor present in every study’s final model was FEV1%, either at baseline or as a time-updated covariate. Age was the second most commonly used predictor and was either incorporated as a predictor in the model, used to restrict the study population, or used as the time scale in the survival analysis. Gender, respiratory infections, and comorbidities were also included in many models. In this thesis, we restrict attention to three baseline covariates and 16 time-dependent covariates found to be important in the prediction of survival for people with CF in recent studies [Aaron et al., 2015, Alaa and van der Schaar, 2018, Keogh et al., 2019b]. A complete listing can be found in Table 3.3. Table 3.4 provides a summary of the predictors for individuals in the

Table 3.2: Predictors used in the final models of the 12 studies reviewed. Predictors are grouped by category and the count indicates the number of studies that included each predictor in their final model.

Category / Predictor	Count	Category / Predictor	Count
<i>Demographics</i>		<i>Other health</i>	
Age	10	Weight	5
Gender	8	Height	5
Genotype	4	BMI	5
Age at diagnosis	3	Resting heart rate	1
		Oxygen saturation	1
<i>Lung function</i>		Albumin	1
FVC%	4	Haemoglobin	1
FEV1%	12	White blood cells	1
Change in FEV1%	2		
<i>Respiratory Infections</i>		<i>Treatment / Therapy</i>	
Pulmonary exacerbations	3	Hospital IV days	3
<i>Pseudomonas aeruginosa</i>	6	Home IV days	4
<i>Staphylococcus aureus</i>	5	Antibiotics courses	2
<i>Burkholderia cepacia</i>	8	Hospitalisations	4
Methicillin-resistant <i>S. aureus</i>	3	Non-invasive ventilation	3
		Oxygen therapy	3
		Corticosteroids	3
		Human DNase	3
<i>Comorbidities</i>		<i>Other</i>	
CFRD	7	Calendar year	1
Pancreatic insufficiency	7		
Enlarged liver	3		

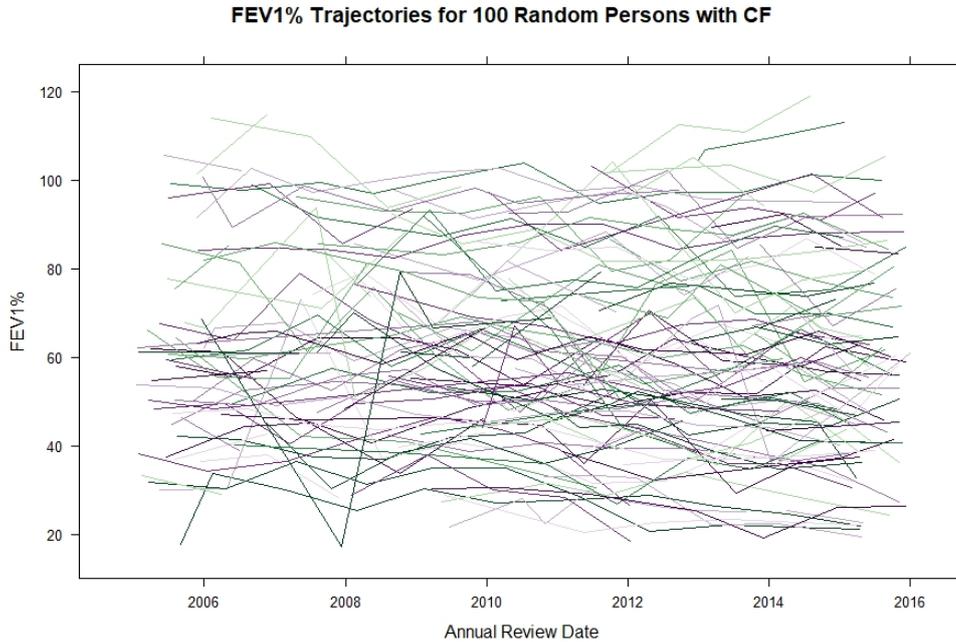


Figure 3.2: Each line plots the approximately annual measurements of FEV1% for one randomly selected person with CF. Measurements were taken between the beginning of 2005 and the end of 2015. Both right- and left-truncation can be seen. For many people, the first FEV1% measurement recorded in the UK Cystic Fibrosis Registry occurred after 2005 and/or the last measurement was recorded before 2015.

study population at landmark ages 25, 35 and 45.

In CF, poor lung function has the greatest impact on survival and quality of life [Horsley et al., 2015]. FEV1%, a measure of airway obstruction, has received the most attention as a gauge of disease severity and as a predictor of survival for CF patients [Liou et al., 2001, Szczesniak et al., 2017, Keogh et al., 2019b]. FEV1% is measured contemporaneously with FVC% using a spirometer which measures the amount of air inhaled and exhaled as well as the time it takes to exhale. Figure 3.2 plots the trajectories of FEV1% for 100 randomly selected patients. From this graphic, we see a heterogeneous mix of upward sloping, downward sloping, flat and jagged trajectories. Using a Danish cohort with monthly lung function measurements, Taylor-Robinson et al. [2012] investigated the source of the variability in FEV1% measurements. They found that approximately 10% of variability is due to measurement error and day-to-day variability; one-half of the variance was attributed to differences between patients and 40% was found to represent disease progression.

Table 3.3: Covariates from the UK Cystic Fibrosis Registry dataset used in the prediction models. The first three time-fixed variables are collected at baseline and the remaining 16 time-dependent variables are collected at each annual review.

<i>Category</i>	<i>Variable</i>	<i>Type</i>
Baseline measure	Sex (female/male)	Binary
	Genotype (F508del homozygous Y/N)	Binary
	Age at diagnosis	Numeric
Lung function	Forced expiratory volume in 1 second as percentage of predicted (FEV1%)	Numeric
	FEV1% slope as estimated from mixed effects model	Numeric
	Forced vital capacity as percentage of predicted (FVC%)	Numeric
Respiratory infection in the past year	<i>Burkholderia cepacia</i>	Binary
	Methicillin-resistant <i>Staphylococcus aureus</i>	Binary
	<i>Pseudomonas aeruginosa</i>	Binary
	<i>Staphylococcus aureus</i>	Binary
Comorbidities	Cystic fibrosis-related diabetes (CFRD)	Binary
	Pancreatic insufficiency	Binary
Other health indicators	Weight	Numeric
	Body mass index (BMI)	Numeric
	Days (past year) in hospital for IV antibiotics	Categorical
	In hospital without antibiotics (past year)	Binary
	Use of oxygen therapy (past year)	Binary
	Use of non-invasive mechanical ventilation (past year)	Binary
Calendar time	Calendar year at measurement time	Numeric

Table 3.4: Descriptive statistics for the study population at landmark ages 25, 35 and 45. For binary and categorical variables, the number and percent are shown. The median, 25th and 75th percentile (IQR) values are shown for continuous predictors.

<i>Variable</i>		<i>Landmark age 25</i>		<i>Landmark age 35</i>		<i>Landmark age 45</i>	
		No.	%	No.	%	No.	%
Sex	Female	1126	55%	515	58%	284	59%
	Male	913	45%	374	42%	195	41%
Genotype	F508del homozygous	1153	57%	437	49%	171	36%
	Other	886	43%	452	51%	308	64%
<i>B. cepacia</i>	Yes	104	5%	49	6%	23	5%
	No	1935	95%	840	94%	456	95%
MRSA	Yes	59	3%	32	4%	12	3%
	No	1980	97%	857	96%	467	97%
<i>P. aeruginosa</i>	Yes	1380	68%	607	68%	307	64%
	No	659	32%	282	32%	172	36%
<i>S. aureus</i>	Yes	859	42%	328	37%	183	38%
	No	1180	58%	561	63%	296	62%
CFRD	Yes	687	34%	348	39%	179	37%
	No	1352	66%	541	61%	300	63%
Pancreatic insufficiency	Yes	1802	88%	716	81%	343	72%
	No	237	12%	173	19%	136	28%
Hospital IV days	0 days	1191	58%	584	66%	332	69%
	1-7 days	137	7%	65	7%	31	6%
	8-14 days	241	12%	91	10%	43	9%
	15-21 days	89	4%	33	4%	18	4%
	22-28 days	110	5%	34	4%	15	3%
	>28 days	271	13%	82	9%	40	8%
Other hospital days	Yes	75	4%	26	3%	17	4%
	No	1964	96%	863	97%	462	96%
Oxygen therapy	Yes	132	6%	62	7%	39	8%
	No	1907	94%	827	93%	440	92%
Non-invasive ventilation	Yes	44	2%	15	2%	12	3%
	No	1995	98%	874	98%	467	97%
		Median	IQR	Median	IQR	Median	IQR
Age at diagnosis		0.4	(0.1,2.3)	1.3	(0.2,6.0)	4.0	(0.5,28.0)
FEV1%		64.2	(45.0,80.8)	57.2	(42.0,75.6)	53.6	(37.1,73.7)
FVC%		80.9	(64.6,93.7)	78.5	(63.3,91.3)	74.7	(59.7,90.3)
Weight (kg)		60.5	(53.3,69.1)	65.5	(57.0,75.0)	67.7	(59.4,76.9)
BMI (kg/m ²)		21.5	(19.7,23.5)	22.8	(21.1,25.0)	23.7	(21.4,26.0)

The amount of missing data in the dataset is relatively low. Less than 1% of respiratory infection and time spent in hospital data were missing and values were filled in using last observation carried forward and by setting the value to zero, respectively. For the critical lung function predictors, FEV1% and FVC%, approximately 12% of all records had missing values.

3.4 Implementation of dynamic survival prediction methods

In this section, I outline how the traditional landmarking analysis, joint modelling and the Super Learner landmark approach introduced in section 2.4 were implemented for analysis using this dataset. For all three methods, age was taken as the timescale.

3.4.1 Creation of training and test datasets

Performance of all three dynamic prediction methods was measured using both 10-fold cross-validation and a holdout sample (test data). The test dataset for validation was created from all patients at 18 randomly selected adult CF centres in the UK, representing 20% of the annual review records. Figure 3.3 depicts this division of data and the 10-fold cross-validation procedure. The test dataset contained 14% of the patients and was used to assess generalisability of the predictive model within the UK. The training dataset was formed from the remaining patient data and was randomly partitioned by patient ID into 10 folds for cross-validation. Each fold contained approximately 540 individuals and the number of events observed ranged from 94 to 119.

3.4.2 Implementation of the joint model

Because FEV1% is measured contemporaneously with FVC%, these two measures were modelled using a multivariate mixed effects model in the longitudinal sub-model. FEV1% and FVC% were modelled linearly as a function of age with random intercept and slope as in equation (2.3). The remaining time-dependent covariates, including BMI, were treated as exogenous predictors and incorporated along with the time-fixed covariates into the survival sub-model of equation (2.4) by fitting an extended Cox model with data specified in long format (i.e. one row per observed measurement time, variables identifying the beginning and end of the time interval in each row, and an event indicator per row). This approach was chosen because of the infeasibility of including 16 longitudinal predictors, many of which are binary, in the longitudinal sub-model and also because lung function is

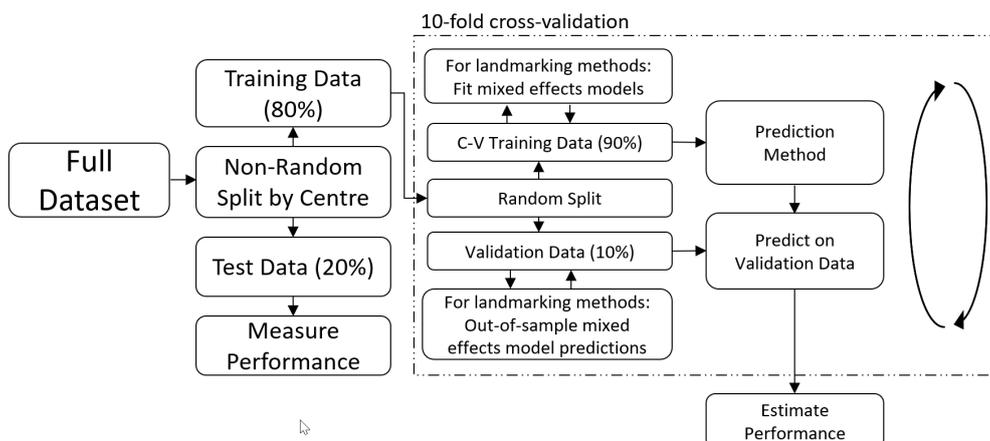


Figure 3.3: Creation of training and test datasets and use of cross-validation for estimating performance. The “Prediction Method” is either joint model, Cox landmark or Super Learner landmark. Both 10-fold cross-validation and test data are used to measure performance.

known to be the most important predictor. I review the implications of this decision in the discussion, section 3.6. The logarithm of the baseline hazard function was modelled using a B-spline with 15 knots. Three association structures were considered for the joint model: one assuming the hazard of an event at time t depends on the current value of the longitudinal outcomes at time t , one assuming that it depends on both the current values and the slope of FEV1% and one assuming that it depends on the cumulative effect of FEV1% up to time t . In the latter two, the focus was solely on FEV1% as it is a stronger predictor of survival than FVC% in CF.

3.4.3 Implementation of the traditional landmarking method

For the landmarking analysis using a Cox regression, FEV1% and FVC% were modelled as a linear function of age with random intercept and slope using a multivariate mixed effects model. The model also included the fixed effects for the other covariates in Table 3.3. Figure 3.4 shows the observed FEV1% trajectories for nine random people plotted along with each of their predicted FEV1% trajectories. Additionally, BMI was modelled using a separate random intercept and slope mixed effects model with fixed effects for sex, genotype, age at diagnosis, calendar year, CFRD, pancreatic insufficiency, *P. aeruginosa*, *B. cepacia*, and hospital IV days. Replacing the observed values of FEV1%, FVC% and BMI in the dataset with predicted values offers three advantages: (i) the predicted values account for measurement error, (ii) we can predict values at each landmark age instead of

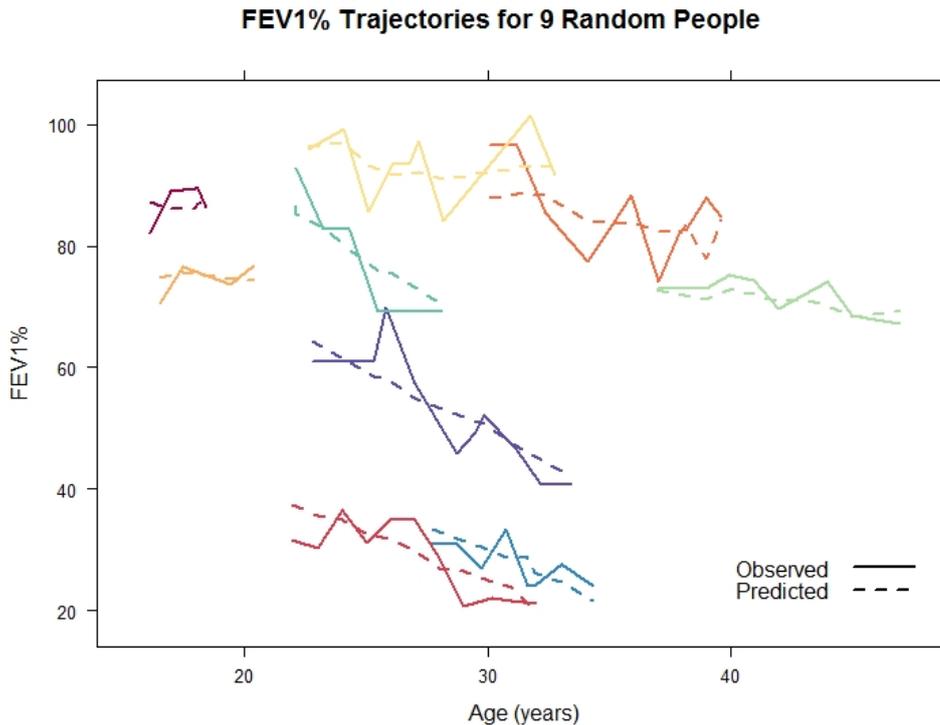


Figure 3.4: Observed (solid line) and predicted (dashed line) FEV1% trajectories for nine randomly selected individuals in the study population. Predicted values were obtained by fitting a multivariate mixed effects model to the observed FEV1% and FVC% data.

carrying forward a value from the last measurement time, and (iii) we can predict values even when observed values are missing.

To capture information about the trajectory of lung function, the individual’s modelled FEV1% slope was also added as a predictor to the landmark models [Rizopoulos et al., 2017, Keogh et al., 2019b]. In each of the 10 cross-validation steps, mixed effects models were fitted and used to predict values for the longitudinal predictors FEV1%, FVC% and BMI in the nine training folds (“C-V Training Data” in Figure 3.3). Out-of-sample predictions of FEV1%, FVC% and BMI were used in the validation fold (“Validation Data” in Figure 3.3) [Keogh, 2018].

The traditional landmarking method uses a stacked dataset. When creating a landmark stacked dataset, each individual contributes one row in the dataset for each landmark age that he or she is at risk as illustrated in Figure 2.3. In choosing a form for the Cox landmark approach, formulations with time-varying effects, with interactions between predictors and with quadratic and cubic terms added for continuous predictors were investigated.

Table 3.5: The number of rows in the super dataset for varying numbers of discrete intervals.

No. discrete intervals	No. rows in super dataset
Not discretised	28,825
5	104,348
10	206,178
20	409,677
30	612,585

3.4.4 Implementation of the Super Learner landmarking method

For the Super Learner landmark analysis, values of FEV1%, FEV1% slope, FVC% and BMI are obtained in the same way as for the traditional landmarking method. However, the landmark stacked dataset used in this approach must additionally be discretised as described in section 2.4.1. In the discretisation, follow-up time s to $s + v$ is divided into a number of adjoining time periods. At each landmark age a person is at risk, they may contribute as many rows as the number of intervals chosen to discretise the data. For this dataset, five discrete intervals were used. Therefore, each person at risk at a given landmark age will contribute up to 5 rows of data for that landmark age. They may contribute fewer than 5 rows if they have an event prior to the start of the final discrete interval. The more discretisation intervals used, the larger the stacked dataset will become. Table 3.5 shows the number of rows in the discretised super dataset for the training data as the number of discrete intervals varies. Our investigations showed that larger numbers of discrete intervals made the size of the data unmanageable and resulted in decreased predictive performance, likely due to the large variance in the estimators of the numerous parameters. Since the discrete intervals need not be of equal size, intervals were created based on quintiles of event times [Polley and van der Laan, 2011]. Table 3.6 shows a hypothetical simplified landmark super dataset before and after discretisation.

For the discrete-time survival approach used in Super Learner landmarking, a decision must be made about how to model the main effect of age (time). Maximum flexibility is allowed by adding time indicators to each row of the discretised super dataset as dummy variables but this increases the number of parameters to estimate and may sacrifice power and coefficient stability [Singer and Willett, 1993]. This study adopted a more parsimonious approach by adding an interaction between the lower bound of the discrete interval and the landmark time indicator which, for parametric models such as logistic regression, models the effect of time

Table 3.6: A landmark super dataset is shown before and after discretisation. In the discretised dataset, the survival time has been replaced by an event indicator “Ev.” that only takes the value of 1 in the discrete interval in which an event happened. The two rightmost columns in the discrete dataset represent the covariates used for modelling time, which are an interaction between the lower bound of the discrete time interval and an indicator variable for each landmark time.

<i>Landmark super dataset</i>					<i>Landmark super dataset after discretisation</i>						
ID	LM time	Surv. time	Ev.	FEV1	ID	LM time	Discrete interval	Ev.	FEV1	I(LM=20) × time	I(LM=21) × time
1	20	24.0	1	30	1	20	(20.0 , 20.8]	0	30	20.0	0
2	20	25.0	0	50	1	20	(20.8 , 21.8]	0	30	20.8	0
1	21	24.0	1	27	1	20	(21.8 , 22.9]	0	30	21.8	0
2	21	25.5	0	52	1	20	(22.9 , 23.7]	0	30	22.9	0
					1	20	(23.7 , 25.0]	1	30	23.7	0
					2	20	(20.0 , 20.8]	0	50	20.0	0
					2	20	(20.8 , 21.8]	0	50	20.8	0
					2	20	(21.8 , 22.9]	0	50	21.8	0
					2	20	(22.9 , 23.7]	0	50	22.9	0
					2	20	(23.7 , 25.0]	0	50	23.7	0
					1	21	(21.0 , 22.0]	0	27	0	21.0
					1	21	(22.0 , 23.0]	0	27	0	22.0
					1	21	(23.0 , 24.4]	1	27	0	23.0
					2	21	(21.0 , 22.0]	0	52	0	21.0
					2	21	(22.0 , 23.0]	0	52	0	22.0
					2	21	(23.0 , 24.4]	0	52	0	23.0
					2	21	(24.4 , 25.1]	0	52	0	24.4
					2	21	(25.1 , 26.0]	0	52	0	25.1

linearly over each landmark time prediction period (as shown in the sample data in Table 3.6).

3.4.5 Selection of algorithms and models for Super Learner

The Super Learner requires a list of candidate algorithms that are appropriate for the problem and that provide diversity and performance. Further, to achieve their best performance, most of these machine learning algorithms require tuning of their hyperparameters, parameters that control some aspect of how the algorithm functions. Table 2.1 lists the algorithms used in the Super Learner ensemble in this study along with references and the name of the R package used. In theory, a

large list of candidate algorithms with multiple combinations of hyperparameters could be included in the Super Learner, which then selects the best algorithms. However, such an approach will be limited by available computation power and time.

To select a list of algorithm-hyperparameter combinations that balances performance and computational requirements, some preliminary investigations were performed to look at the effect of different hyperparameter settings on predictive performance of the algorithms. The purpose of this exercise was to limit the size of the algorithm library used for our final model in consideration of the large computation time. The Super Learner was run with one learner type at a time, with multiple values of the hyperparameter(s) for that learner and then a smaller subset of learners was chosen based on the results. For example, one of the runs consisted of twelve xgBoost learners with different combinations of maximum tree depth (controls the number of levels in the decision tree), minimum child node weight (controls whether the tree will continue to be partitioned) and eta (amount of shrinkage applied to predictor weights). More conservative (lower variance) learners were performing better and eight xgBoost learners that had received a non-zero weight in the Super Learner prediction function in this preliminary investigation were ultimately selected as candidates. Table 3.7 details the hyperparameters and range of values for each that were used for our application to the UK CF Registry dataset. The final set of tuning parameters used is provided along with the algorithm weightings for the final model in the Results section, Table 3.10.

3.4.6 Performance measures

Performance was assessed first by using cross-validation on the training data and second using the held out test data. (See Figure 3.3) Using all three methods, dynamic 2-year and 5-year survival predictions were computed at each landmark age for those individuals at risk at that landmark age. The Brier score and C-index were separately calculated at each landmark age that a prediction was made, (20, 21, . . . , 50) for each analysis method using IPCW methods. The censoring distribution was estimated separately at each landmark age s using only those individuals at risk at time s . Because nearly all censoring in the UK CF Registry is administrative censoring, the assumption that C_i is independent of both T_i^* and the covariates, as required for IPCW methods, is reasonable for this dataset. We construct calibration plots by dividing the predicted survival probabilities at a given landmark age into fifths, computing the mean survival probability for each fifth (y -axis), and plotting this against the average Kaplan-Meier survival for the individuals in each fifth (x -axis).

Table 3.7: Details of preliminary investigations into algorithm-hyperparameter combinations for Super Learner.

<i>Algorithm</i>	<i>Hyperparameter Values</i>
Random forest	As recent research recommends not tuning the number of trees for classification problems, number of trees was set to 500, a common default value [Probst and Boulesteix, 2018]. Minimum terminal nodesize, the value below which a node will not be split, ranged from 1 to 5.
Gradient boosting	Model complexity was controlled by varying the maximum tree depth (values 3, 4, 5), minimum weight needed in a child node to continue partitioning (values 1, 10) and eta, the learning rate (values 0.1, 0.3).
Lasso and elastic-net GLM	The form of the penalty (lasso, ridge or a combination) was varied by tuning α across a grid of values from 0 (ridge) to 1 (lasso). The glmnet package automatically tunes the parameter λ across 100 values and biglasso used 50 λ values. λ controls the strength of the penalty.
Generalized additive models (GAMs)	Smoothing splines were used in the GAM. Values used for the degrees of freedom parameter (λ) were 2, 3, 4, 5.

3.4.7 Software

All analyses were performed using R v3.4.3 [R Core Team, 2020]. The joint model and the longitudinal submodel were fitted using the R package `JMbayes` [Rizopoulos, 2016], the Cox landmark supermodel was fitted using the R package `survival` [Therneau and Grambsch, 2000, Therneau, 2015] and the Super Learner landmark supermodel was fitted using the R package `SuperLearner` [Polley et al., 2018]. Custom code was written for the v -year IPCW squared error loss function. R packages used for the individual algorithms in the Super Learner ensemble are provided in Table 2.1. The R package `nlme` [Pinheiro et al., 2018] was used to fit the mixed effects model for FEV1% and FVC% and the Lawson-Hanson algorithm [Lawson and Hanson, 1995] which is implemented in the R package `nlsls` [Mullen and van Stokkum, 2012] for the non-negative least squares problem. The Brier score and C-index were computed using the R package `pec` [Mogensen et al., 2012].

R code for using Super Learner to fit a discrete-time landmark supermodel like the one described here is available from <https://github.com/KamTan/DynamicPrediction>. This code is illustrated using the Mayo Clinic Primary Biliary Cirrhosis dataset publicly available via the R package `survival`. This dataset was chosen as we are not permitted to make the UK CF Registry data public. However, researchers may apply for this data by completing the UK CF Registry data request form available from the CF Trust web site at <https://www.cysticfibrosis.org.uk/the-work-we-do/uk-cf-registry>.

3.5 Results

3.5.1 Cross-validated performance using the training dataset

First, the performance of a joint model, a traditional landmark super model analysis using Cox proportional hazards (“Cox landmark”) and a discrete-time landmark super model analysis using the Super Learner ensemble (“SL landmark”) is compared using 10-fold cross-validation on the training dataset.

Cox landmarking that included time-varying effects, predictor interactions, as well as quadratic and cubic terms was implemented. None of those more complex formulations offered superior Brier scores or C-index values at a majority of landmark ages as compared to the more parsimonious model including only the linear terms and so the higher order terms were discarded.

For the joint model, three different association structures were considered. Figure 3.5 presents the predictive accuracy and discriminative ability of each joint model. The model with a cumulative effects association structure performed worse than the other two models on both measures. Because no improvement in predic-

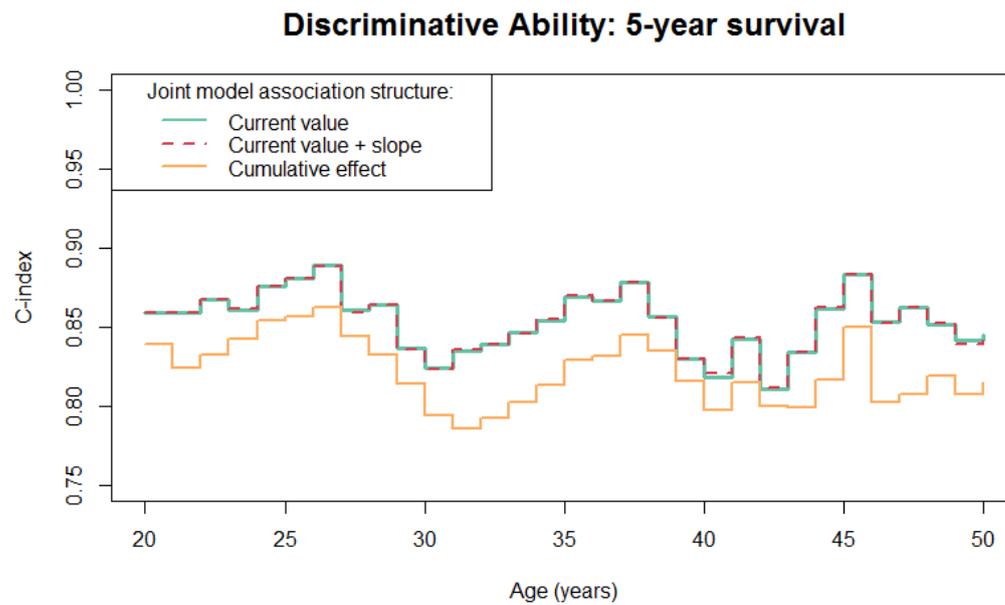
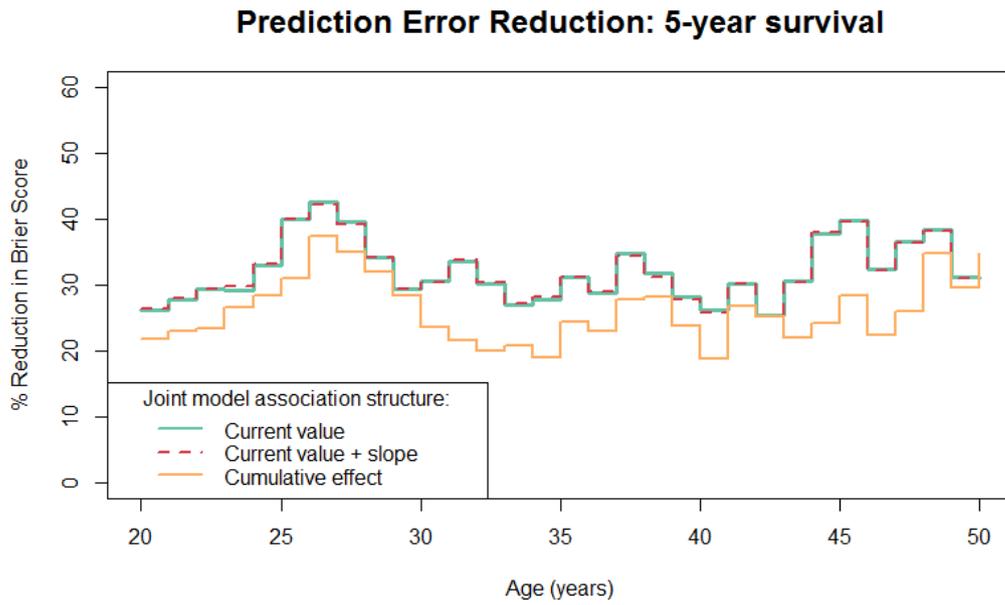


Figure 3.5: Performance evaluation of the three candidate joint models. The reduction in prediction error for 5-year dynamic survival prediction at each landmark age is shown for each of three association structures (top). Larger percent reduction in error is preferred. The C-index is plotted at each landmark age for the three joint models under consideration (bottom). Higher values of C-index are preferred.

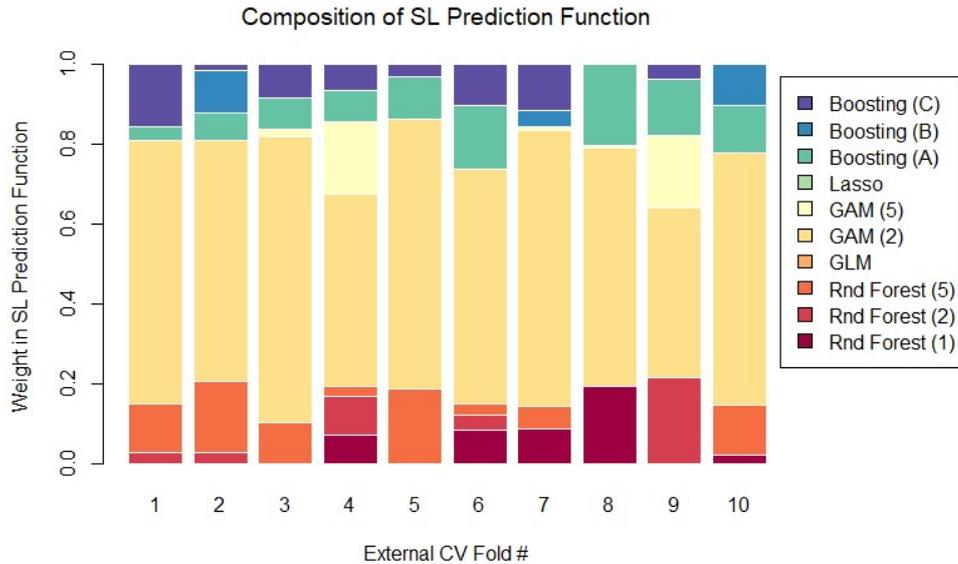


Figure 3.6: At each of the 10 external cross validation folds, a new Super Learner prediction function is estimated based on the Super Learner internal 10-fold cross validation loop. The composition of algorithms in those prediction functions is presented with each learner plotted in a different colour. The alphanumeric character in parentheses after the name of the learner serves to distinguish different configurations of the algorithm based on hyperparameter settings. For example, GAM (2) indicates a GAM with 2 degrees of freedom while GAM (5) indicates 5 degrees of freedom were specified.

tive accuracy or discrimination was achieved with the more complex association structure involving FEV1% slope, we chose to relate the hazard of an event at time t only to the current value of FEV1% at time t .

Figure 3.6 shows the composition of the Super Learner prediction function for each of the 10 cross-validation folds. The GAM (2) learner has the largest weight in the Super Learner prediction function at all 10 folds, and overall there is minor variability in which algorithms are chosen and their weights in the prediction function. For example, the Boosting (C) learner receives a weight of 0.15 (15%) in fold 1 but is given a zero weight in both fold 8 and fold 10. Neither the GLM nor the Lasso learners were given a non-zero weight in any of the 10 cross-validation folds.

In figure 3.7, the Brier score for 10 algorithms used in Super Learner landmarking is plotted for 5-year dynamic prediction of survival at each integer landmark age from 20 to 50. For comparison, the Brier score of a model with no covari-

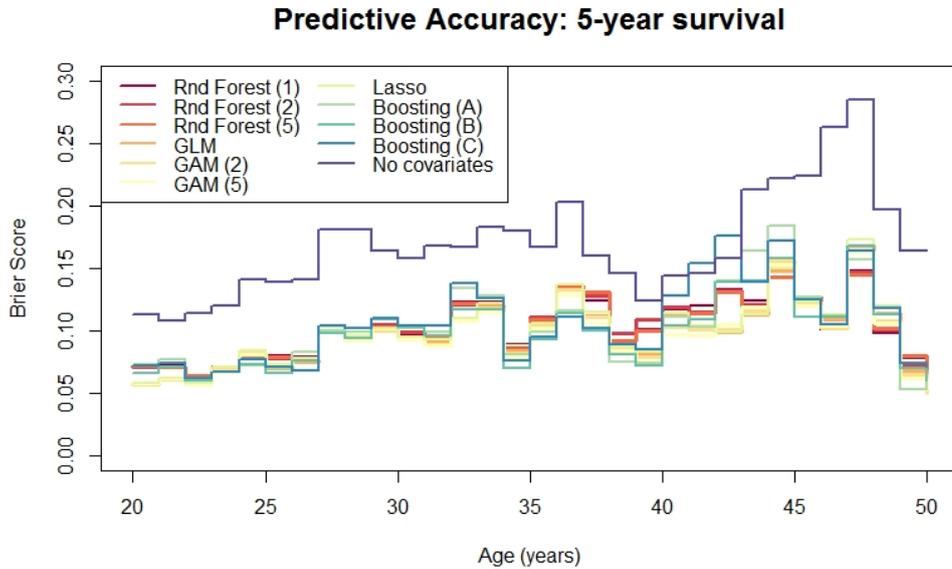


Figure 3.7: Brier score for the individual algorithms specified in the Super Learner library for 5-year dynamic survival prediction. In purple, the Brier score for a model with no covariates is plotted. The x-axis represents the landmark ages from 20 to 50 from which predictions were made.

ates is shown in purple. Different algorithms show better predictive performance at different ages with no single algorithm dominating the others. Overall, most algorithms follow a similar pattern across the landmark ages.

Table 3.8 shows the Brier scores for each method for 2-year and 5-year dynamic survival prediction. Figure 3.8 presents this information as a plot of the percentage reduction in Brier score of each method over Kaplan-Meier reference estimates. For 2-year survival prediction, the SL landmark achieved a 30% or greater reduction in prediction error at 11 of the 31 landmark ages while the Cox landmark and joint model achieved this at only three and two landmark ages, respectively. For 5-year survival prediction, the performance of the Super Learner landmark and the Cox landmark was nearly identical. The reduction in prediction error for the SL landmark ranged from 29% to 52% compared to 28% to 51% for the Cox landmark and 22% to 44% for the joint model. For this dataset, the joint model performance is inferior to that of the two landmarking methods. As seen in Table 3.8, for all three methods, predictive accuracy decreases at higher landmark ages, particularly over 40 years of age, likely due to fewer data at older ages.

As shown in Figure 3.9, the discriminative ability of all three methods for 2-year survival prediction is similar with C-index values exceeding 0.80 at all landmark

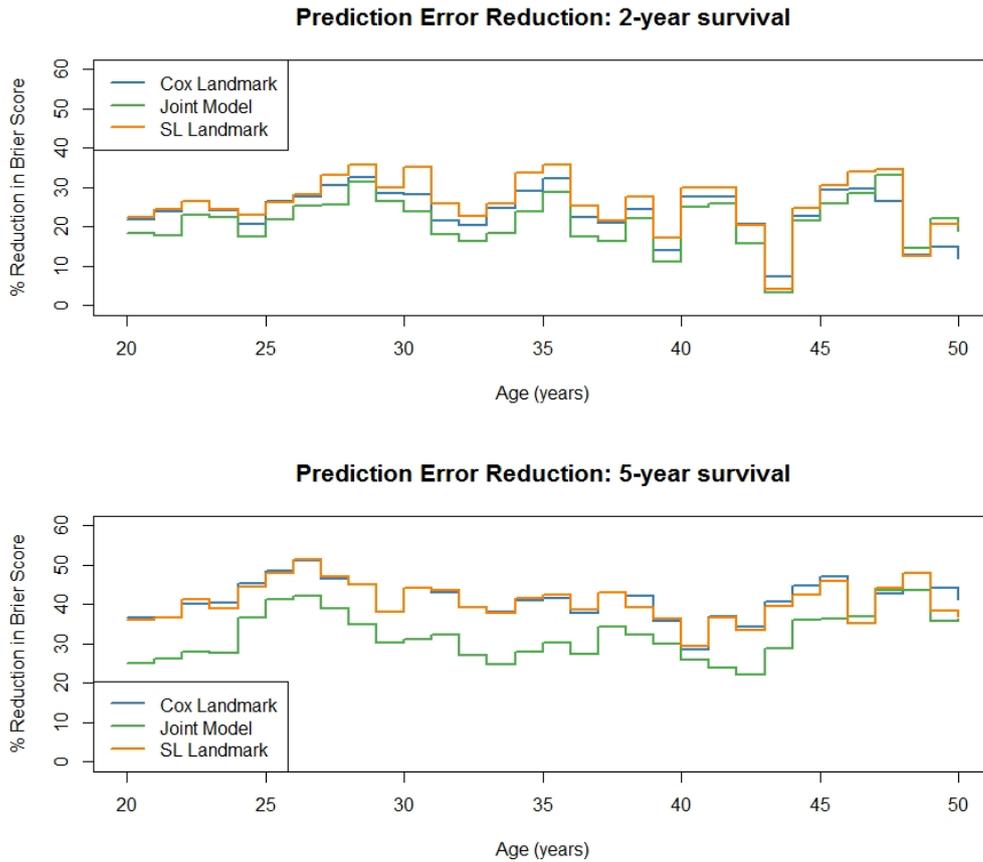


Figure 3.8: The reduction in prediction error (Brier score) over the Kaplan-Meier reference model for 2-year and 5-year dynamic survival prediction using 10-fold cross-validation on the training dataset.

Table 3.8: Values of the Brier Score for the three comparison methods for 2-year and 5-year dynamic survival prediction at selected landmark ages. Lower values indicate lower prediction error. These numbers were computed using 10-fold cross-validation on the training data.

Method		<i>Landmark age:</i>						
		20	25	30	35	40	45	50
2-yr survival	Cox Landmark	0.04	0.04	0.05	0.05	0.04	0.05	0.08
	Joint Model	0.04	0.05	0.05	0.05	0.04	0.06	0.07
	SL Landmark	0.04	0.04	0.04	0.05	0.03	0.05	0.07
5-yr survival	Cox Landmark	0.08	0.07	0.08	0.08	0.09	0.09	0.10
	Joint Model	0.09	0.08	0.10	0.10	0.09	0.11	0.11
	SL Landmark	0.08	0.07	0.08	0.08	0.08	0.09	0.11

ages except 43 and 44 years old. For 5-year survival prediction, the Cox landmark and SL landmark models have a similar ability to discriminate with C-index measures between 0.85 and 0.90 for landmark ages 20 to 40 years. The joint model has a slightly lower discriminative ability than either of the two landmark methods.

Figure 3.10 shows calibration plots for the both the 2-year and 5-year dynamic prediction models at three landmark ages: 25, 35, and 45 years of age. There is good calibration for all three methods. At landmark age 45, where there are data from fewer individuals, the calibration is less satisfactory. The joint model also appears to over-estimate 5-year survival at landmark age 35.

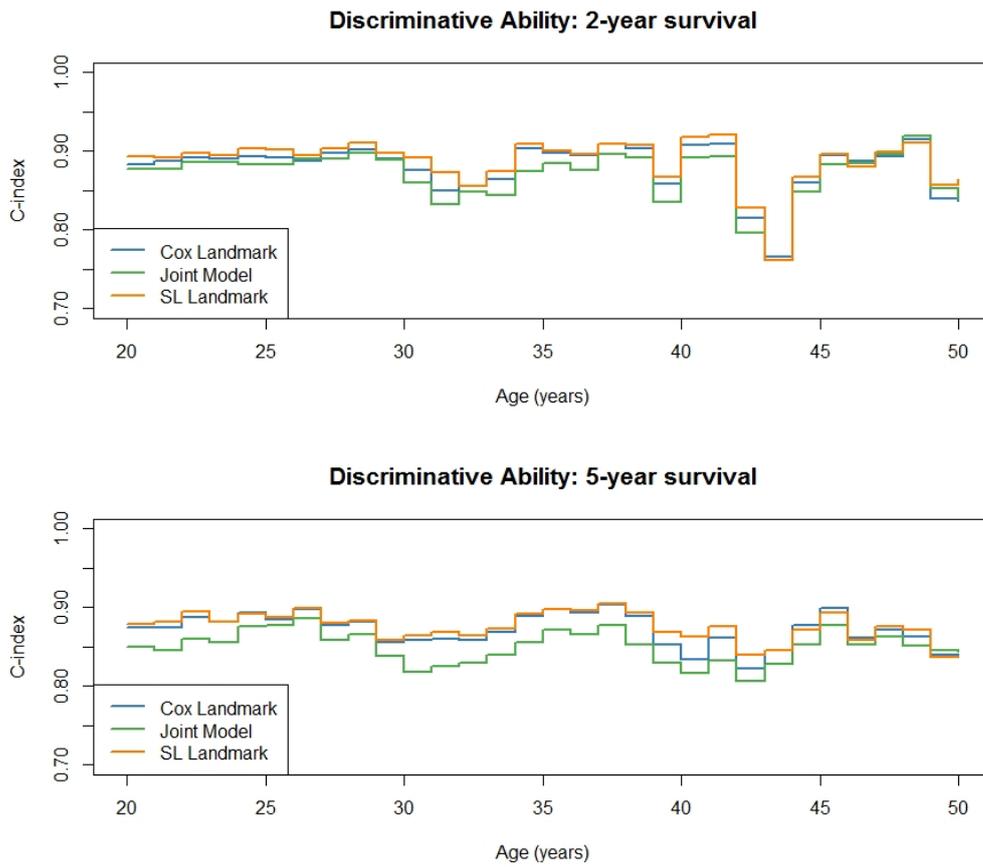


Figure 3.9: Comparison of discriminative ability for the SL landmark (orange), the Cox landmark (blue) and the joint model (green) using 10-fold cross-validation on the training data. Discrimination is measured by the C-index, with higher values preferred.

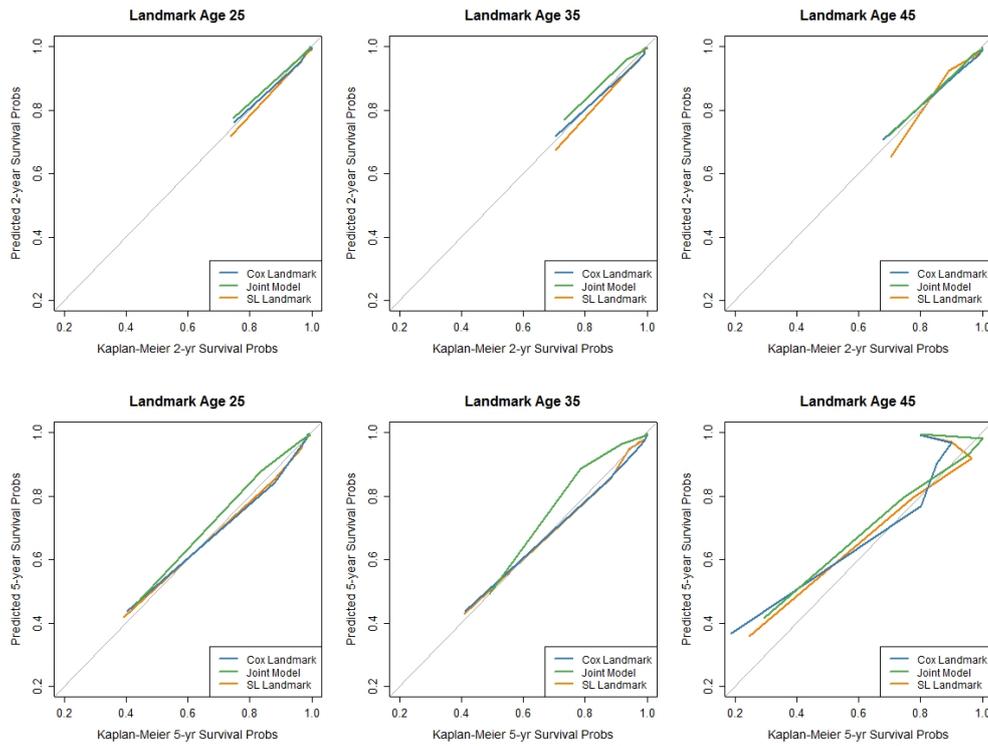


Figure 3.10: Calibration plots for 2-year (top) and 5-year (bottom) survival predictions for three selected landmark ages: 25, 35 and 45. The y-axis represents the mean survival probability by quintile group as calculated by each method. The x-axis represents the Kaplan-Meier survival probability estimated on the same individuals. The mean probabilities are averaged over the 10 external cross-validation folds for each quintile group and then joined using a blue, green or orange line depending on the method used for prediction. For reference, a 45-degree line is plotted in grey.

Table 3.9: Comparison of predictors between test and training datasets for individuals at 25, 35 and 45 years of age.

Predictor	Test: age 25 (n=181)	Train: age 25 (n=1453)	Test: age 35 (n=89)	Train: age 35 (n=648)	Test: age 45 (n=33)	Train: age 45 (n=387)
Female (%)	62	54	62	58	58	59
F508del mutation (%)	55	57	42	49	39	35
Med diag age (yrs)	0.3	0.4	1.5	1.2	8.3	4.0
FEV1% (25th ptile)	48.6	46.3	44.8	44.0	38.7	39.6
FEV1% (50th ptile)	65.6	63.2	63.7	58.1	47.7	55.1
FEV1% (75th ptile)	79.9	79.8	77.5	75.1	81.1	72.2
<i>In the past year:</i>						
Oxygen therapy (%)	7	7	9	7	9	9
Non-invasive mechanical ventilation (%)	4	2	2	1	0	3
In hospital \geq 3 weeks for IV antibiotics (%)	18	20	11	13	18	11

3.5.2 Performance using the test dataset

In addition to assessing 10-fold cross validated performance of the three methods, the performance of the three dynamic prediction methods was also compared on a test dataset comprised of patient data from 18 adult CF centres that were not part of the training data. To assess similarity between the test data and training data samples, Table 3.9 presents information on key predictors for the test dataset and the training dataset at landmark ages 25, 35 and 45. Overall, the test and training datasets have a similar composition in terms of predictors.

The predictive performance on the test dataset of the models fit to the complete training dataset was measured using both predictive accuracy and discrimination. The reduction in prediction error over a Kaplan-Meier model for each method for 2-year and 5-year dynamic prediction of survival is shown in Figure 3.11. For 2-year dynamic survival prediction, at some ages, all three methods were able to reduce prediction error by more than 30%. The percent reduction in Brier score for 5-year dynamic survival prediction across landmark ages ranged from 26% – 75% for Cox landmark, 14% – 61% for joint model and 21% – 67% for Super Learner landmark. At some landmark ages, the Cox landmark had superior predictive accuracy and at some ages the Super Learner landmark produced the smaller error.

Discrimination of all three methods for both 2-year and 5-year prediction horizons was good at younger ages, with all methods achieving a C-index in excess of 0.80. At older ages, particularly over 40 years old, all three methods showed less

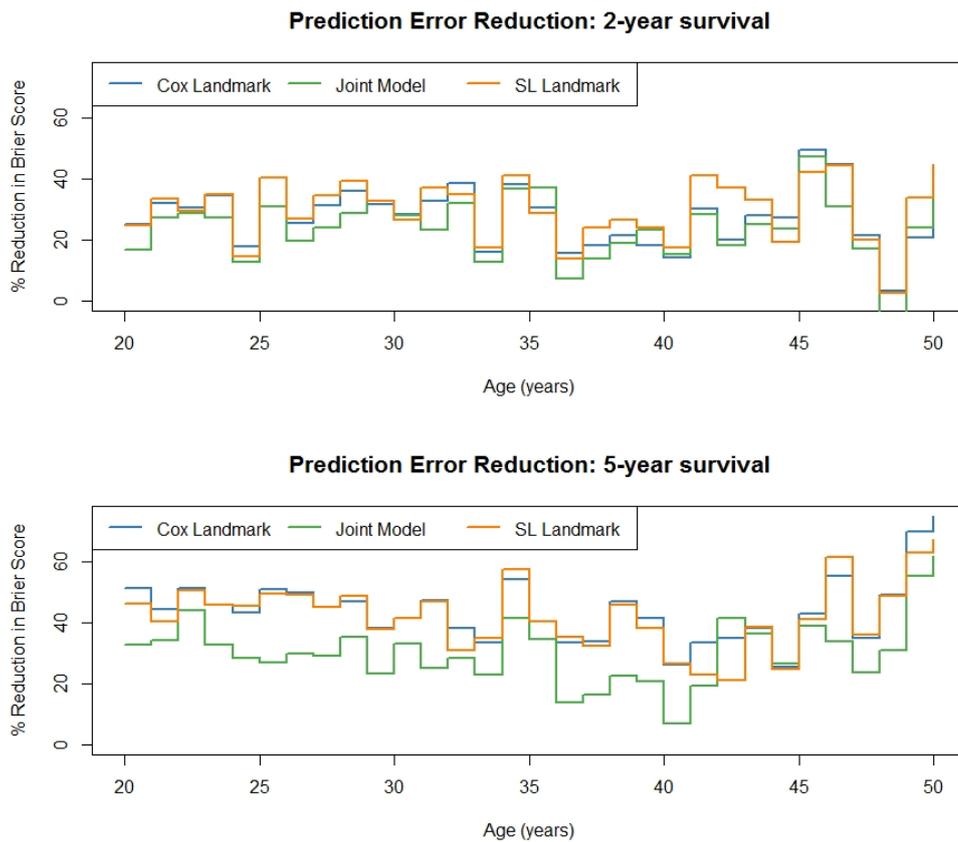


Figure 3.11: Reduction in prediction error over a Kaplan-Meier model for each method for 2-year (top) and 5-year (bottom) survival using the test dataset.

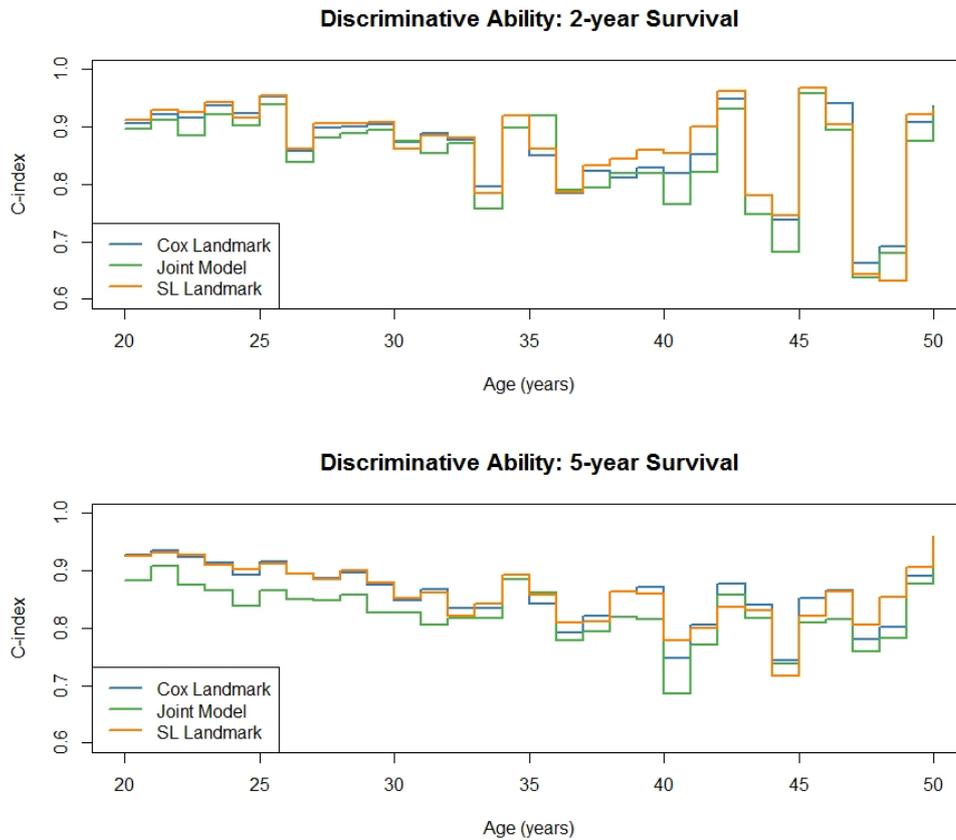


Figure 3.12: Discriminative ability of each method based on the test data as measured by the C-index for 2-year (top) and 5-year (bottom) dynamic survival prediction.

ability to discriminate. Figure 3.12 plots the C-index for each method at landmark ages 20 – 50 for 2-year and 5-year dynamic survival prediction.

Final model

Two final models, one for prediction of 2-year and one for prediction of 5-year dynamic survival probabilities were computed using the full dataset, which includes both the training data and the test data. The algorithm composition of the final model Super Learner prediction functions are presented in Table 3.10 along with the value of the v -year IPCW squared error loss for each algorithm or model. For predicting both 2-year and 5-year survival, GAM predictions make up the majority of the prediction function, 57% and 64%, respectively. Several versions of random forest and boosting algorithms contribute the remainder of the prediction function.

Table 3.10: The percent weighting of each learner in the Super Learner prediction function (SL PF) fit to the full dataset for both 2-year and 5-year dynamic prediction of survival. The IPCW squared error loss, the value minimised in the Super Learner regression to compute the prediction function weights, is also included.

<i>Category</i>	<i>Algorithm & hyperparameters</i>	2-year survival		5-year survival	
		<i>wt in SL PF</i>	<i>IPCW Loss</i>	<i>wt in SL PF</i>	<i>IPCW Loss</i>
Random forest	Ranger: no. trees = 500; min node size = 1	0.0%	0.050	8.1%	0.085
	Ranger: no. trees = 500; min node size = 2	7.4%	0.050	4.2%	0.085
	Ranger: no. trees = 500; min node size = 5	10.5%	0.050	4.9%	0.085
Gradient boosting	xgBoost: eta=0.1, max depth = 3, min obs per node = 10	0.0%	0.047	0.0%	0.083
	xgBoost: eta=0.1, max depth = 4, min obs per node = 10	0.0%	0.048	0.0%	0.083
	xgBoost: eta=0.1, max depth = 4, min obs per node = 10, sub-sample = 0.5	11.8%	0.050	0.0%	0.087
	xgBoost: eta=0.1, max depth = 4, min obs per node = 1	0.0%	0.049	0.0%	0.085
	xgBoost: eta=0.1, max depth = 6, min obs per node = 10	0.0%	0.050	5.6%	0.085
	xgBoost: eta=0.1, max depth = 6, min obs per node = 1	0.0%	0.050	0.2%	0.087
	xgBoost: eta=0.3, max depth = 6, min obs per node = 10	0.0%	0.056	14.8%	0.093
	xgBoost: eta=0.3, max depth = 6, min obs per node = 1	1.0%	0.054	4.9%	0.095
	GLM	glm	0.0%	0.046	0.0%
bayesglm		0.0%	0.046	0.0%	0.083
GAM	gam: degrees of freedom = 2	0.0%	0.046	57.3%	0.081
	gam: degrees of freedom = 3	0.0%	0.045	0.0%	0.081
	gam: degrees of freedom = 4	0.0%	0.045	0.0%	0.081
	gam: degrees of freedom = 5	64.3%	0.045	0.0%	0.081
Penalised regression	biglasso: penalty = lasso	0.0%	0.046	0.0%	0.084
	biglasso: penalty = ridge	0.0%	0.059	0.0%	0.133
	glmnet: penalty = 50% lasso / 50% ridge	0.0%	0.046	0.0%	0.083
Reference	mean model (no covariates)	5.1%	0.062	0.0%	0.141

3.5.3 Practical considerations

The computation time required for the three methods ranges from minutes to days on a PC with quad-core 3.2 GHz CPU and 16GB of RAM. The Cox landmark method produced dynamic predictions over 10 external cross-validation folds for this analysis in 5 minutes. The same results took the joint model approximately 30 hours and the Super Learner landmark with the library of candidate algorithms reported above just over 2 days. The performance changed little when models were run on a similarly powered shared cluster with 16GB of RAM and 8 CPU cores running at 2.5 GHz. The computation time for the Super Learner landmark varied dramatically based on the size and complexity of the candidate library of algorithms. As expected, more complex non-parametric algorithms such as gradient boosting and support vector machines require more computation time than parametric or semi-parametric models. Using parallelisation techniques can reduce the run-time of a Super Learner analysis significantly.

3.6 Discussion

I have described a method for adapting the landmarking technique for dynamic survival predictions to be used with a machine learning ensemble that includes statistical models and algorithms for estimation of the probability of a binary outcome. A key advantage of the ensemble is that it obviates the need to make an a priori choice as to which statistical model or class of machine learning algorithm will perform best for the problem at hand. The obvious disadvantages are greater complexity and computational cost. Another disadvantage of the ensemble is that it obscures the nature of relationships between predictors and outcomes and it is not possible to provide a full algorithm from which someone could obtain a prediction. For the UK CF Registry study population used here, the SL landmark method performed as well as the current state-of-the-art Cox landmark method for dynamic survival prediction both in external cross-validation and on a test dataset, but it did not outperform this simpler approach. Both methods had better predictive performance than the joint model.

Depending on the true data-generating distribution, different algorithms will perform best for different problems. For this dataset and this question, a traditional statistical method performed as well as our ensemble method. Given the high level of discrimination achieved by both methods, there is not much room to make gains by fitting more elaborate models or including more predictors. It is of interest to better understand when there will be performance differences between approaches. For example, certain alternative statistical models or algorithms may be particularly useful when the number of predictors greatly exceeds the number

of observations, for cases with interactions amongst the predictors and when there is non-linearity. Not surprisingly, non-parametric machine learning methods can outperform parametric models when the assumptions of the parametric models are violated. As shown by Gong et al. [2018] and Katzman et al. [2018], machine learning techniques outperformed Cox regression models in a static survival prediction setting when the effects of covariates were non-linear in the hazard. Lee et al. [2019] report a deep learning approach for dynamic survival prediction using data from the UK CF Registry that offers improved discrimination in their comparisons. Their results are difficult to compare with ours as the set of predictors used was different, they adopt a competing risks framework and the specifics of the base method implementations are not fully reported in the paper.

This work was motivated by the need for dynamic predictions of survival for patients with life-shortening conditions. Data was used from the UK Cystic Fibrosis Registry to illustrate the methods, but they could be applied to other locales and other health-related administrative databases. Because people with CF have regularly-scheduled routine care, the longitudinal data is measured at pre-planned, consistent intervals and the amount of missing data is low. Other datasets with more sporadic patient visits or without well-defined data collection may require special modelling techniques, including potentially to accommodate a model for the occurrence of patient visits. Additionally, not all datasets will be immediately generalisable to the overall population of people with the disease. The dynamic survival predictions resulting from this work are relevant for the entire UK population of people with CF as the registry data covers approximately 99% of cases in the UK. Additionally, predictive ability of the SL landmark on a non-random external test dataset suggests the model is not limited in applicability to a specific set of centres in the UK. Because the population characteristics were similar between the test and training datasets, we are unable to comment on generalisability of the prediction model beyond the UK population of people with CF.

I compared the SL landmark method to a joint model and the Cox landmark method. Instead, these methods could have been added to the Super Learner library of algorithms and each would either be selected or not selected in the final combination of algorithms. While any modelling technique can theoretically be included in the Super Learner ensemble, it may not always be practical or necessary. For example, we may not need to add a continuous-time proportional hazards learner because a discretised logistic model converges to the continuous-time Cox proportional hazards model as the time interval gets smaller [Thompson, 1977]. To add a joint model to the Super Learner framework would require additional custom code to enable the ensemble to simultaneously work with both discrete-time and continuous-time paradigms and also with models that take fundamentally different approaches to incorporating longitudinal data. However, because the joint model

produces dynamic survival predictions at given ages, there is no reason these could not be included as predictions from one of the candidate learners to determine the algorithm weightings in the Super Learner prediction function. In other words, the predicted survival probabilities from the joint model could be added as another row in the matrix Ψ and then the solution to the minimisation problem of equation (2.15) would include the joint model as a candidate.

In pursuit of the best model for dynamic survival prediction, I utilised a v -year IPCW squared error loss function. This loss function directly measures the difference in predicted and actual conditional survival probabilities. To use the Super Learner without additional custom code, a negative log likelihood loss function or a squared error loss function measured on the conditional hazard could be used. While these are still valid loss functions, they are not directly based on the conditional survival probability we are interested in and, therefore, the resulting Super Learner prediction function may be inferior [Polley and van der Laan, 2011]. To investigate how the results obtained using the squared error loss function differ from those using the IPCW squared error loss function, the Brier scores that result from running Super Learner with the same candidate library of algorithms using the two different loss functions were compared. As shown in Figure 3.13, overall, the performance was nearly identical. The difference between the two methods may be greater for a different dataset but this suggests that using the loss function defined on the conditional hazard would not give grossly different results and it has the advantages of being included in the core package in R and not requiring an estimate of the censoring function.

It is interesting to compare the IPCW squared error loss for the learners that were selected for the final prediction function with the loss for the learners that were not selected. From Table 3.10 we can see the IPCW squared error loss for 5-year survival prediction was 0.081 for all four algorithm-hyperparameter combinations of the GAMs yet a non-zero weight was only allocated to one of them. Further, the GLM learners had a loss of only 0.083 and the lasso had a loss of only 0.084 yet neither of these categories of learner were selected to contribute predictions to the final prediction function. The explanation for this behaviour lies in the attempt to balance bias, variance and covariance which is achieved through error diversity. Simply combining algorithms with the lowest IPCW squared error loss will not necessarily deliver the best performance, especially if their predictions and, therefore their errors, are strongly correlated. In this study, error diversity has been encouraged by including fundamentally different algorithms as well as multiple versions of algorithms with different hyperparameters. Other methods for bringing diversity to the ensemble include repeatedly providing different subsets of the training data to the algorithm (as in bagging), providing a different set of covariates to the algorithm, or by distorting the predictor data in some way

Comparison of Predictive Accuracy: 5-year survival

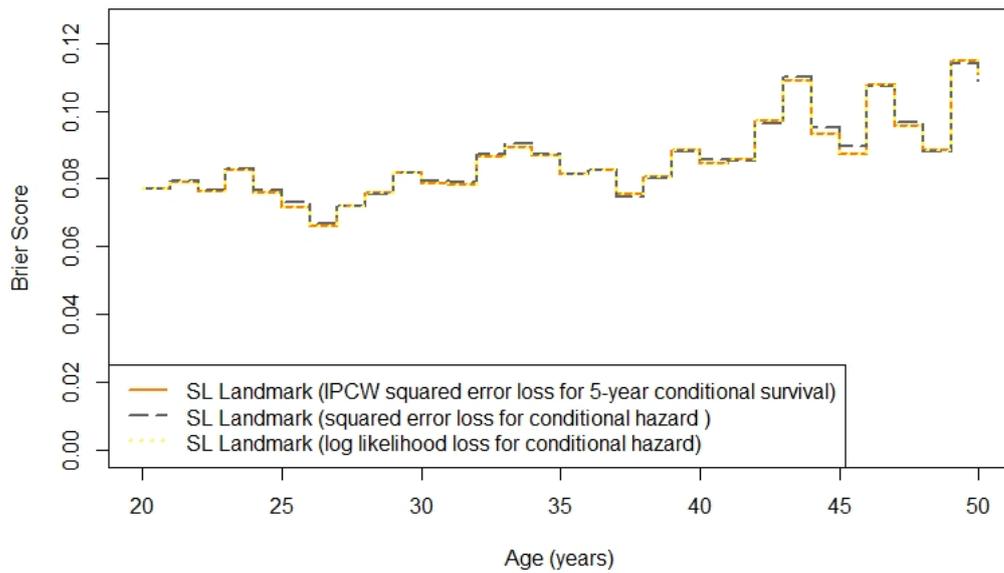


Figure 3.13: Comparison of the Brier score for 5-year dynamic survival prediction for three SL Landmark analyses with different valid loss functions. The orange line represents the performance of the method when using an IPCW squared error loss function for 5-year conditional survival. The dashed grey line represents the performance of the same method with an unweighted squared error loss function calculated with respect to the hazard. The dashed yellow line shows the Brier score when a log likelihood loss function on the hazard is used.

[Brown, 2004]. Because an ensemble balances the covariance as well as the bias and variance, diversity may be key to performance improvement.

In this study, I fit a single landmark supermodel to a stacked dataset across all landmark ages but I could have fit one separate model per landmark age in a sliding landmark analysis. A sliding landmark analysis allows for a unique Super Learner prediction function with different algorithm weightings at each landmark age. The price for this flexibility is a great deal more parameters to estimate. In this dataset, where there are significantly fewer observations at older ages, there is a benefit to borrowing data from other landmark ages in the landmark supermodel. The performance of the landmark supermodel was found to be superior to that of a sliding landmark analysis.

To take advantage of a vast range of existing algorithmic techniques, the approach of discretising continuous time-to-event data was taken. Another option for enabling the application of machine learning to survival outcomes involves augmenting the data to accommodate censoring thereby allowing standard regression-style algorithms to be used. Examples include inverse probability of censoring weighting the data [Rotnitzky and Robins, 2005, Vock et al., 2016, Goldberg and Kosorok, 2017] and pseudo-observations [Andersen and Pohar Perme, 2010, Parner and Andersen, 2010].

A common interest in machine learning is identifying important predictors. When faced with large datasets containing more (possibly collinear) predictors than observations, machine learning has had success in identifying a subset of important predictors. Feature selection may be achieved through an initial screening or filtering step using a correlation-based or information-based ranking of predictors [Guyon and Elisseeff, 2003]. Alternatively, wrapper or embedded methods for feature selection build the model and select the variables simultaneously. The lasso regression algorithm is a good example from this category [Tibshirani, 1996, 1997]. As feature selection was not the goal of this study, the list of predictors used was based on those used in current models in the literature for predicting CF survival. Because the UK CF Registry contains hundreds of variables per measurement time, this may be an interesting area for further research and was touched upon by Alaa and van der Schaar [2018] in their study of prediction of 3-year mortality for people with CF and also by Lee et al. [2019].

The relative predictive performance of the joint model in this study was inferior to the two landmarking approaches. A likely explanation involves the addition of the time-varying covariates to the survival submodel. Because these covariates can only be measured if the patient is alive (i.e. they are endogenous), they would properly be incorporated into the longitudinal submodel via mixed-effects modelling [Rizopoulos, 2012]. However, the large number of predictive time-varying covariates (many of which are binary) in this study makes this infeasible and so,

to avoid losing the information contained in these covariates, they were included as if they were exogenous in the survival submodel. When making predictions, the joint model assumes that the value of exogenous time-varying covariates remains constant in the future at the most recently measured level. This assumption is not realistic as we expect the values to change over time with the patient's health. For datasets like this one where there are a large number of time-varying covariates, the landmarking approaches may be more suitable.

The goal of this study was to provide a thorough comparison of the main methods for dynamic survival prediction, to show how a machine learning ensemble method can be used for dynamic survival prediction and to elucidate the numerous considerations and decisions that must be made during the analysis. The performance of machine learning methods and the completeness in reporting of machine learning methods is mixed. In a recent review by Christodoulou et al. [2019], many machine learning studies failed to report methods for tuning hyperparameters, did not assess calibration, and neglected to clearly report on all decisions and steps taken in the analysis. This variation in reporting standards likely contributes to the existing tension between traditional statistical methods and machine learning. Efforts to improve in this area will help to facilitate an understanding that we do not have to make a choice between machine learning algorithms and statistical models; they are both valuable tools for statisticians and analysts to use in clinical prediction models.

Chapter 4

Simulation Study: Comparison of Dynamic Prediction Methods

4.1 Introduction and aims

In the previous chapter, I compared three techniques for dynamic survival prediction using UK CF Registry data. While all methods performed well, joint modelling had inferior predictive performance to the Cox landmarking approach and landmarking with a machine learning ensemble. In this chapter, I further investigate the performance of these methods using a simulation study and a data scenario where it is hypothesised that the machine learning ensemble will outperform the Cox landmark model and the joint model for dynamic prediction of survival.

Several simulation studies have been published that compare the performance of joint models and landmarking for dynamic prediction. Ferrer et al. [2018] simulated data for three scenarios with model misspecification and one scenario where the joint model was correctly specified. The model misspecification scenarios considered the case where the model for the longitudinal biomarker was misspecified, data where the proportional hazards assumption was violated and, also, data where the slope of the longitudinal predictor affected the time-to-event but slope was not modelled as a predictor. In the simulation study of Rizopoulos et al. [2017], which also compared joint modelling to landmarking, they focused on scenarios that modelled three different association structures between the longitudinal predictor and the survival outcome. These association structures were discussed in section 2.3.1 and include the hazard depending on the current value of the biomarker, on the current value and slope of the biomarker, and on the cumulative effects of the biomarker history. In a comparison of joint models with partly conditional models, Maziarz et al. [2017] simulated data from a joint model with varying degrees of

measurement error of the longitudinal predictor.

Existing simulation studies that compare the performance of machine learning techniques to traditional approaches for survival prediction mainly focus on static survival prediction using time-fixed predictors. Gong et al. [2018] compared random survival forest and artificial neural network algorithms to a Cox regression with linear terms for static prediction of survival. They simulated data with two baseline predictors and a hazard function that was related to these predictors linearly, non-linearly, via an interaction between the two predictors and with non-linearity and an interaction and reported results based on the C-index for measuring discrimination. In a study comparing a deep neural network to a Cox regression with linear terms, Katzman et al. [2018] simulated data on 10 covariates but constructed a hazard function that depended on only two of them, both with quadratic effects.

In this chapter, several simulated datasets are created and the performance of the Super Learner landmarking approach is compared to the performance of Cox landmarking and joint modelling for dynamic survival prediction using measures of discrimination and predictive accuracy. Specifically, the aim of this chapter is to verify that a scenario exists in which the Super Learner landmarking approach will outperform the other two approaches and use this information to suggest strategies for determining when the machine learning ensemble could provide benefit to an analysis.

4.2 Data-generating mechanisms

I did not attempt to simulate data of the same form as the UK Cystic Fibrosis Registry data described in section 1.5 and used in the analysis of chapter 3 because of the large number of important predictors, the complexities of left truncation and age as the timescale, as well as the need to model two correlated longitudinal variables. It is unlikely that these complexities would lead to performance differences. Instead, a simpler scenario was simulated where observation of all individuals begins at time $t = 0$ and data on a single continuous longitudinal variable, $y_i(t)$, and two time-fixed predictors, one binary, Z_{1i} , and one continuous, Z_{2i} , are collected. This data set-up could be applied to many different dynamic prediction applications and is not limited to our cystic fibrosis setting.

The longitudinal variable and the survival time are generated jointly using a univariate mixed effects model and a Weibull hazard model. The longitudinal sub-model is defined as:

$$y_i(t) = m_i(t) + \epsilon_i(t) \quad (4.1)$$

$$m_i(t) = \beta_0 + \beta_t t + \beta_{Z_1} Z_{1i} + \beta_{Z_2} Z_{2i} + b_{0i} + b_{ti} \quad (4.2)$$

where $y_i(t)$ is the value for individual i observed at time t with measurement error $\epsilon_i(t)$ of the true longitudinal process, $m_i(t)$. The time-fixed predictors are generated using: $Z_{1i} \sim \text{Bernoulli}(p = 0.5)$ and $Z_{2i} \sim N(50, 1)$. Fixed effects parameters were set to: $\beta_0 = 60.0$, $\beta_t = -3.0$, $\beta_{Z_1} = \beta_{Z_2} = 0$ and $\epsilon_i(t) \sim N(0, \sigma_y^2)$ with $\sigma_y = 6.6$. The random effects (b_{0i}, b_{ti}) were assumed to follow a multivariate normal distribution with mean 0 and covariance matrix D . The standard deviations used for the random effects were 44.0 and 3.0 with correlation -0.9.

The Weibull survival sub-model can be written as:

$$h_i(t) = kt^{k-1} \exp(\gamma_0 + \gamma_{Z_1} Z_{1i} + \gamma_{Z_2} Z_{2i} + \alpha m_i(t)) \quad (4.3)$$

The following parameter values were used in the simulation: $\gamma_{Z_1} = 0.5$, $\gamma_{Z_2} = 0.07$, $\alpha = -0.15$ and the Weibull shape parameter, $k = 2.0$. Two different scenarios were simulated: ‘‘Scenario 1’’ with 51% of simulated individuals experiencing an event within the 16-year follow-up time ($\gamma_0 = -7.0$), and ‘‘Scenario 2’’ with 12% of the simulated population having an event ($\gamma_0 = -11.0$). All individuals still at risk at $t = 16$ were administratively censored.

1,000 training datasets of 1,000 individuals plus one test dataset of 1,000 individuals were generated. To enable prediction of 2- and 5-year survival from three different landmark times, ($t=5, 7, 10$), up to 20 longitudinal observations with a maximum follow-up time of 16 years were produced for each individual. The measurement time of each longitudinal observation was chosen randomly from a uniform distribution $[0, 16]$ to create unbalanced data.

Thus far, I have described the standard generation of data from a joint model and this served as a reference scenario. One would expect a joint model to perform very well on data generated from a joint model but, as our aim was to create a scenario where machine learning would outperform joint models and/or Cox landmarking, the longitudinal predictor was transformed. Following an idea from Kang and Schafer [2007], suppose that the data collected included survival time, T_i , covariates Z_{1i} and Z_{2i} and longitudinal data $y_i^*(t)$, a variable related to $y_i(t)$, but $y_i(t)$ itself was not collected or provided for analysis. Here, $y_i^*(t)$ was artificially created but a situation can be imagined where $y_i(t)$ itself cannot be measured and instead, data on $y_i^*(t)$ is collected as a proxy. $y_i^*(t)$ was created using one of two transformations. Transformation A, with both non-linearity and an interaction, is defined as:

$$y_i^*(t) = 30Z_{1i} \frac{(y_i(t) + 50)}{Z_{2i}} + (1 - Z_{1i}) \frac{Z_{2i}^2}{(y_i(t) + 200)/7} \quad (4.4)$$

and transformation B, with only an interaction, is defined as:

$$y_i^*(t) = 50Z_{1i}\frac{y_i(t)}{2Z_{2i}} + (1 - Z_{1i})\frac{2y_i(t)}{Z_{2i}} \quad (4.5)$$

These transformations were designed to test the ability of the models to correctly make dynamic predictions when the relationship between the predictors and the survival time involves interactions and non-linearity. Further, the transformations were structured so that the resulting $y_i^*(t)$ would appear approximately normal and, therefore, fitting a linear mixed effects model would be appropriate. Figure 4.1 shows plots of the transformed longitudinal variable, $y_i^*(t)$, versus the originally simulated $y_i(t)$ for both transformations. The presence of an interaction is evidenced by the two non-parallel clusters of points. In the figure for transformation A, the non-linearity can be seen in the downward sloping curved cluster of points.

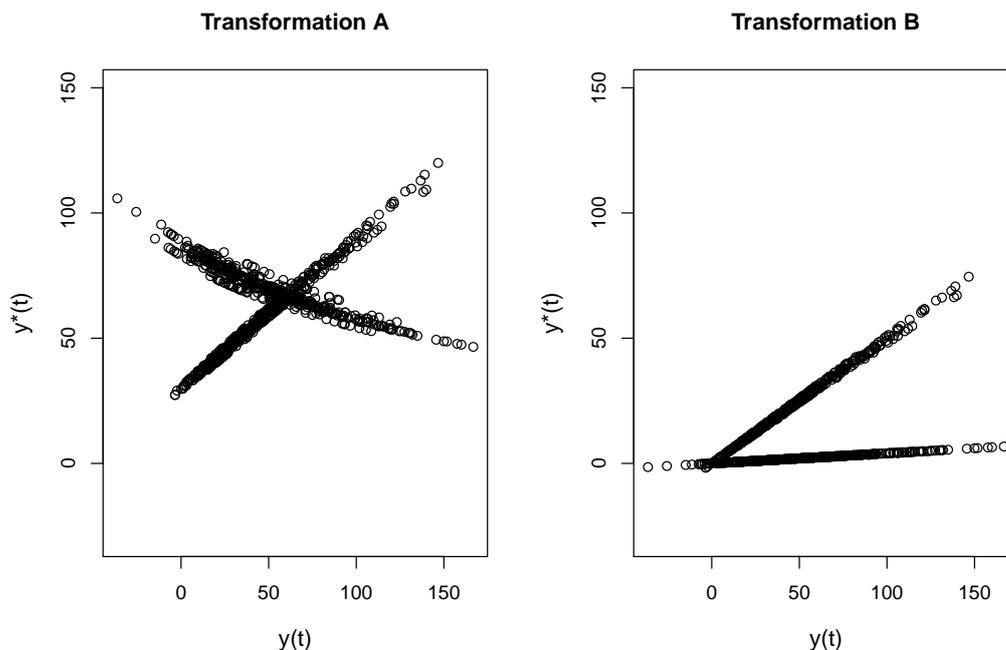


Figure 4.1: Plots of the transformed longitudinal variable, $y_i^*(t)$ versus the originally simulated longitudinal variable, $y_i(t)$ for the two transformations A and B. Data from one simulated dataset ($n=1,000$) are shown.

Five simulation cases were analysed: scenario 1 with no transformation (i.e. $y_i^*(t) = y_i(t)$) as a reference case, scenario 1 with transformation A and B and scenario 2 with transformation A and B.

4.3 Methods

The same methods used in the dynamic prediction modelling of the previous chapter were compared in this simulation study: a joint model, a Cox landmark approach, and the Super Learner landmark method. Because the data were generated by a joint model assuming a current value association structure, alternative association structures for the joint model were not considered in the simulation study. The primary Cox landmark approach assumed a linear structure for the predictors but a Cox landmark approach that allows for all possible interactions between the predictors was also considered.

Similar to the analysis of the previous chapter, in both landmarking approaches, a linear mixed effects model (LMEM) was fitted to the longitudinal outcome to generate predicted values at each landmark time that account for measurement error. The longitudinal variable was modelled as a linear function of time with random intercept and slope and fixed effects of the two other covariates, Z_{1i} and Z_{2i} . To ensure that these predicted values were not solely responsible for performance differences seen, we compared the Cox landmark and SL landmark methods both with and without predicted values of the longitudinal covariate. When not using predicted values, we used the last observation carried forward approach in which the most recent value of the longitudinal variable is used as the predictor at each landmark time. Stacked landmark datasets were created as described in sections 2.3.2 and 2.4.1. Survival predictions were made from landmark times $t = 5, 7$ and 10 using prediction horizons $v = 2$ and 5 years. The effect of time was modelled linearly over each prediction period. Because of the long computation time when running Super Learner with a large library of algorithms, a smaller number of algorithms was used in this simulation than in the analysis of the previous chapter. The list of algorithm / hyperparameter combinations selected for use in the Super Learner ensemble for this simulation study is provided in appendix table B.1

4.4 Performance measures

Each method was assessed using two performance measures: the Brier score for predictive accuracy and the C-index for discrimination. (See section 2.5 for details.) Each of the 1,000 simulated datasets was used to fit each of the three methods and then performance was evaluated based on predictions made for the test dataset. The primary outcome assessed was 5-year predicted survival. I additionally compare 2-year predicted survival for scenario 1 with transformation A. Both the Brier score and the C-index were computed at three landmark times: 5, 7, and 10. Because the maximum possible follow-up time in the simulated data is

16 years, nearly all events will have occurred for a 5-year survival prediction made at landmark time 10. Comparison of the methods is facilitated by presenting the difference between the performance of the joint model and the Super Learner landmark method and the difference between performance of Cox landmarking and Super Learner landmarking approaches.

4.4.1 Software

All analyses and generation of data were performed using R [R Core Team, 2020]. Data were simulated using the `simjm` R package [Brilleman, 2019, Crowther and Lambert, 2013]. The joint model was fitted using the R package `JM` [Rizopoulos, 2010], the Cox landmark supermodel was fitted using the R package `survival` [Therneau and Grambsch, 2000, Therneau, 2015] and the Super Learner landmark supermodel was fitted using the R package `SuperLearner` [Polley et al., 2018]. The R package `nlme` [Pinheiro et al., 2018] was used to fit the LMEM for the longitudinal variable and the R package `pec` [Mogensen et al., 2012] was used for computation of the Brier score and C-index. R code for running a simulation study that compares these three methods for dynamic survival prediction is available from <https://github.com/KamTan/DynamicPredSimulation>.

4.5 Results

Results of the simulation analysis are reported as differences in Brier score and C-index from the Super Learner landmark method; “JM - SL” (“Cox - SL”) is the Brier score or C-index of the joint model (Cox landmarking) minus the Brier score or C-index from Super Learner landmarking. For each comparison, values for the 1,000 simulated datasets are visualised using box plots. Differences in Brier score that are greater than zero imply that the method being compared to the Super Learner landmark had a poorer predictive accuracy. Conversely, differences in C-index that are less than zero suggest the comparison method had a poorer discriminative ability than the Super Learner landmark method.

In the reference scenario where the longitudinal predictor was not transformed, Super Learner landmarking and Cox landmarking performed equally well both in terms of predictive accuracy and discrimination for 5-year dynamic survival prediction. The average difference between Cox landmarking and Super Learner landmarking was 0.0 for both Brier score and C-index at all three landmark times. Joint modelling, however, had better performance on both dimensions at landmark time $t=10$ with a difference in Brier score from Super Learner landmarking of -0.03 when predicting 5-year survival. Appendix figure C.1 presents the results. The larger the landmark prediction time, the more improvement over Super Learner

landmarking the joint model showed.

Figures 4.2 and 4.3 show the results from the simulation analysis for scenario 1 (51% events) with transformation A. Compared to the joint model and Cox landmarking with linear terms only, Super Learner landmarking showed superior predictive accuracy and discriminative ability at all three landmark times for both 2-year (top row) and 5-year (bottom row) predicted survival. The difference in performance was greater for 5-year survival and increases for later landmark times. At $t = 7$, the mean Brier score for SL landmarking for 5-year dynamic survival prediction was 0.09. The mean for Cox landmarking with no interactions was 0.12 and the mean for the joint model was 0.13. Lower Brier scores indicate better predictive ability. The average C-index across simulated datasets for SL landmarking at $t = 7$ for 5-year predictions was 0.83 while the average for the joint model and Cox landmarking was 0.56 and 0.59, respectively. Higher values of C-index correspond to better discrimination. The remainder of the results presented will focus on 5-year survival.

Figure 4.4 presents the results for scenario 2 (12% events) with transformation A for 5-year dynamic survival prediction. There is little difference in predictive accuracy between the three methods as evidenced by the majority of the differences being close to zero. Also, the absolute Brier scores are very low. In contrast, Super Learner landmarking exhibits substantially greater discriminative ability than the other two methods but there is a high degree of variability between simulated datasets in the C-index differences.

The simulation analysis results for scenario 1 with transformation B are displayed in Figure 4.5. Although the Super Learner landmarking approach still outperforms the other two methods, the performance difference is smaller than in either scenario 1 or 2 with transformation A. The absolute levels of Brier score for Super Learner landmark are almost identical to those achieved in scenario 1 with transformation A. Using scenario 2 with transformation B, almost equal predictive performance was found for the three methods and this varies little across the 1,000 simulated datasets. (Results not shown) Unlike scenario 1 with transformation B where there was a large difference in C-index between the methods, there was no significant difference in C-index between methods for this case.

Because an interaction and a quadratic term were used in transformation A to transform the longitudinal variable, the performance of Super Learner landmarking to Cox landmarking formulated with interaction terms between the predictors was also compared. Figure 4.6 shows the results of this comparison where “Cox*” represents Cox landmarking with interaction terms and “Cox” represents Cox landmarking without these additional terms. The difference between Cox* and SL landmarking is much closer to zero than the difference between Cox and SL Landmarking for both Brier score and C-index, indicating superior performance

Scenario 1 / Transformation A: difference in Brier score

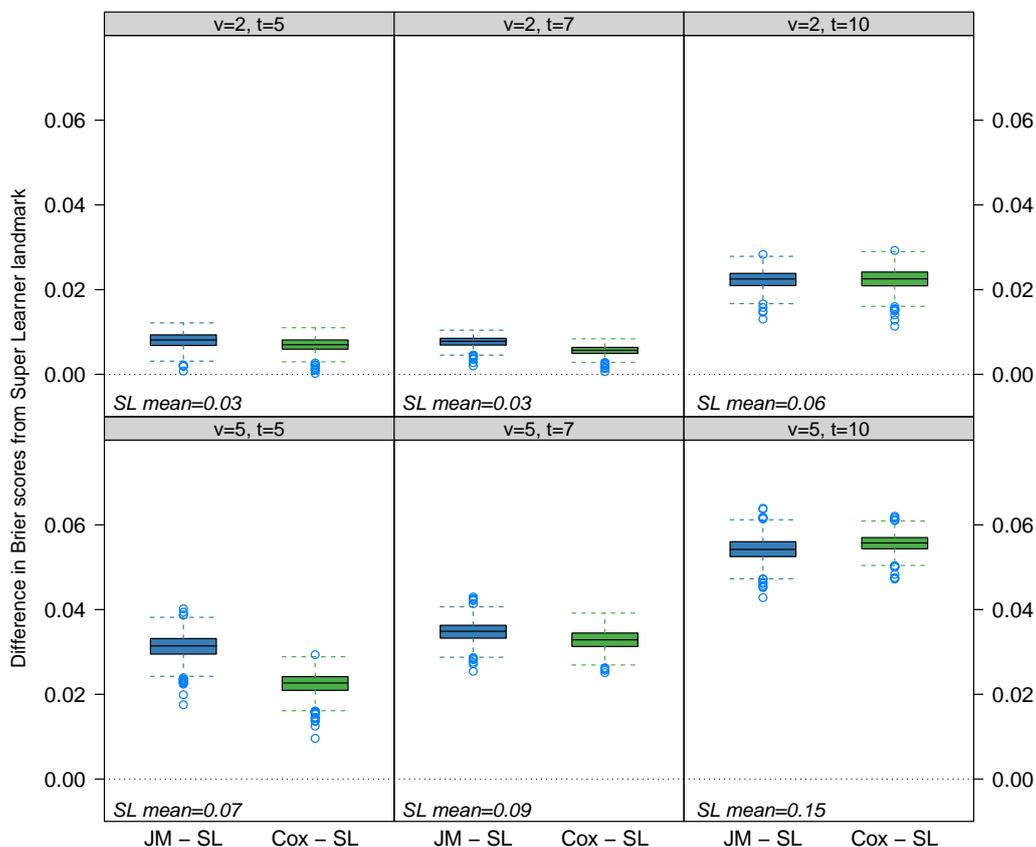


Figure 4.2: Box plots of the difference in Brier score between the Super Learner landmarking (SL) and the two comparison methods, Cox landmarking (Cox) or joint modelling (JM) for scenario 1 (51% events). Each represents the difference over 1,000 simulated datasets. The filled rectangles in the box plots represent the first through third quartiles with a line drawn at the median. Unfilled blue circles beyond the whiskers represent more extreme data points. The mean Brier score for SL is noted as “SL mean”. The top row of panels corresponds to $v=2$ -year survival predictions and the bottom row are $v=5$ -year survival predictions, at three landmark times $t=5, 7, 10$. Differences greater than zero indicate that the comparison method has poorer predictive ability.

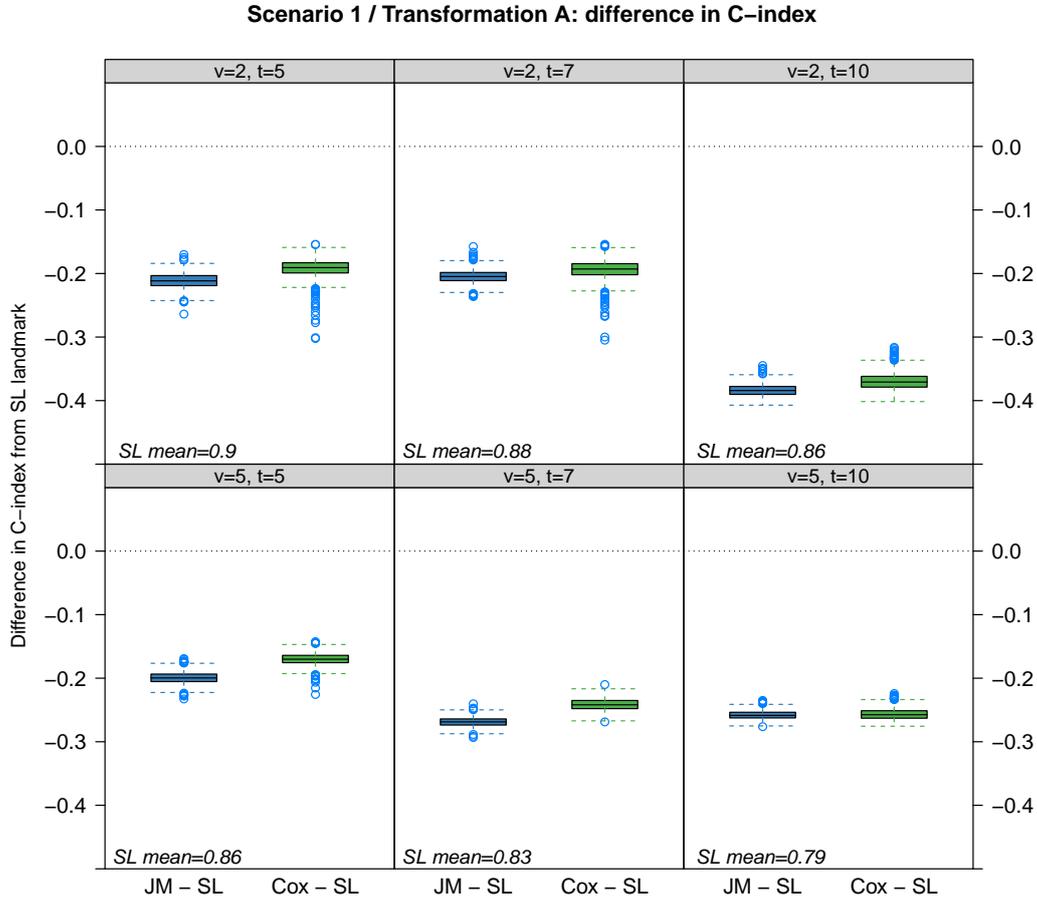


Figure 4.3: Box plots of the difference in C-index between the Super Learner landmark and the two comparison methods, Cox landmark (Cox) or joint modelling (JM), across the 1,000 simulated datasets for scenario 1, the high event scenario. “SL mean” indicates the mean C-index for the Super Learner landmark approach. The six panels represent the differences at the landmark times, $t=5, 7, 10$ for $v=2$ - and 5-year dynamic survival prediction. Because higher C-index values indicate better discriminative ability, differences less than zero in the figure indicate superior performance by SL.

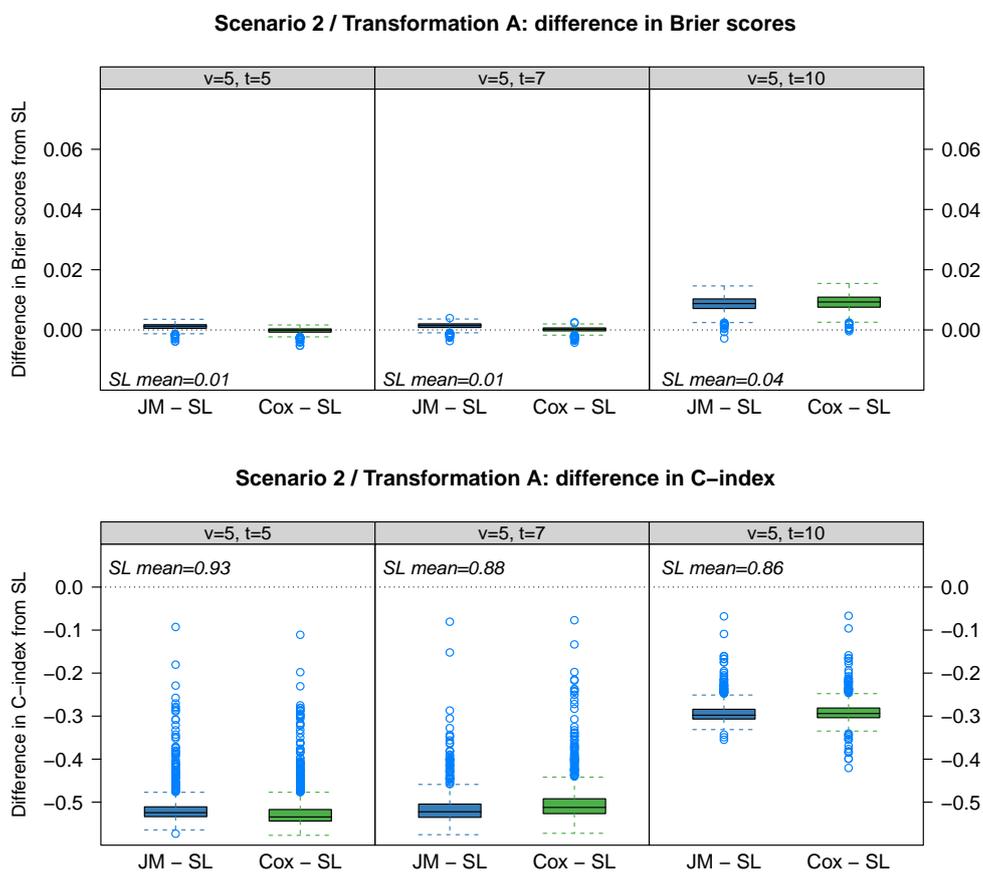


Figure 4.4: Box plots of the difference in Brier score (top) and C-index (bottom) between the Super Learner landmark (SL) and the two comparison methods, Cox landmark (Cox) or joint modelling (JM), across the 1,000 simulated datasets for scenario 2, the low event scenario. “SL mean” indicates the mean value for the Super Learner landmark approach. Each panel represents the difference at one landmark time, $t=5$, 7, or 10 for $v=5$ -year dynamic survival prediction.

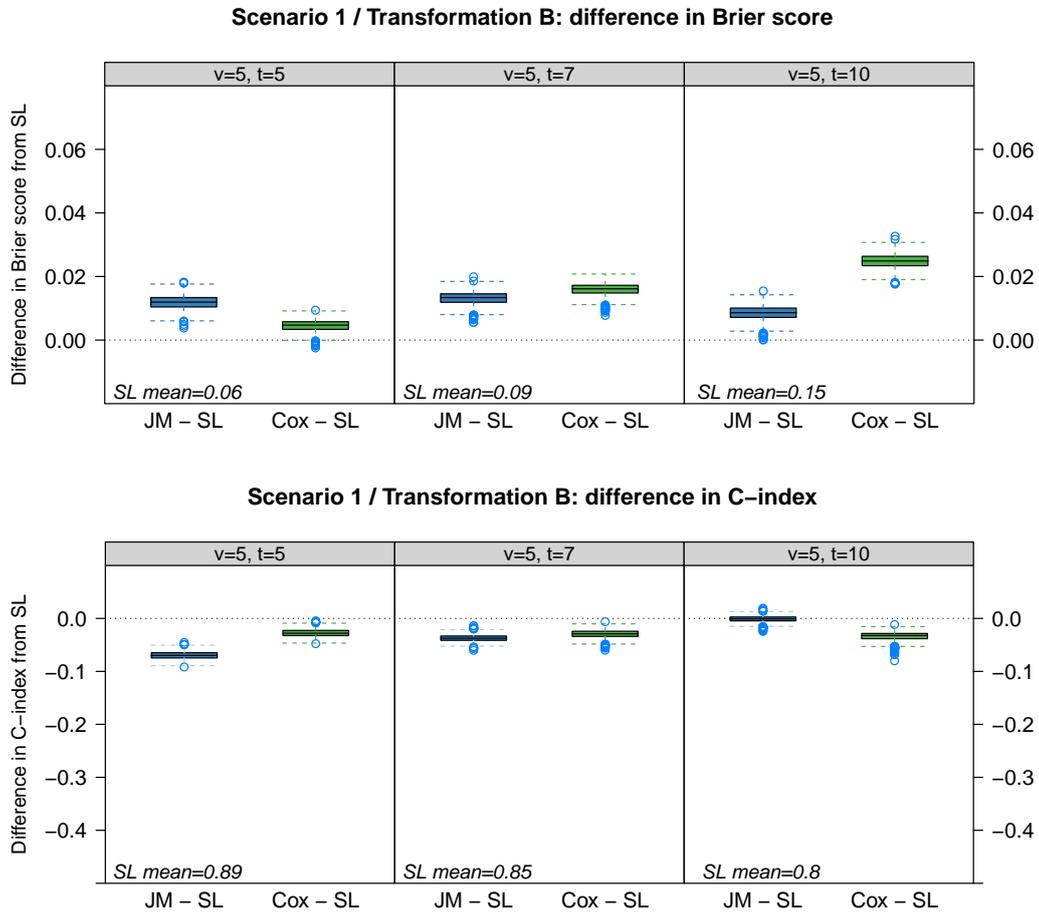


Figure 4.5: Results of the simulation analysis for scenario 1 (high event scenario) with transformation B. The difference in Brier score (top) or C-index (bottom) between the joint model (JM) or Cox landmarking (Cox) and the Super Learner landmark approach (SL) are summarised using box plots. The mean Brier score or C-index obtained by SL “SL mean” is reported in each panel. Performance measures were calculated on 1,000 simulated datasets at three landmark times: 5, 7, and 10.

of Cox* to Cox. Because the difference in Brier score (C-index) between Cox* and SL landmarking is below (above) zero for many of the simulated datasets, Cox* appears to slightly outperform Super Learner landmarking.

To investigate whether fitting a LMEM to the longitudinal variable affected the results, the comparison of the three methods was repeated on scenario 1 with transformation A using $y^*(t)$ as the predictor instead of the predicted value, $\hat{y}^*(t)$ from the fitted LMEM. These results are presented in figure 4.7. The difference in predicted accuracy between the methods is similar to what was seen when predicted values from the LMEM were used. The mean Brier score for the Super Learner landmarking indicates slightly worse performance when not using a LMEM to predict values of the longitudinal covariate. A similar pattern can be seen in the C-index.

4.6 Discussion

In this simulation study, predictive accuracy and discriminative ability of the Super Learner landmarking method was found to be superior to joint modelling and Cox landmarking without interaction terms for dynamic survival prediction in three out of the four cases where the longitudinal variable was transformed and equal in the other. The performance difference was greatest when the number of individuals experiencing an event was close to 50% and when the time-varying covariate had been transformed both non-linearly and with an interaction. When interaction terms were included in the Cox landmarking approach, its performance matched and sometimes exceeded that of Super Learner landmarking. Interestingly, in our analysis of the UK CF Registry data, it was found that including interactions and non-linear terms in the linear predictor for Cox landmarking did not improve performance (see section 3.5.1). In the reference case where the longitudinal variable was not transformed, all methods performed equally at the early landmark prediction time ($t=5$) but joint modelling exhibited superior performance at the last landmark prediction time ($t=10$). The joint model was expected to predict well on this untransformed data as it was simulated from a joint model.

In the presence of non-linearity and interactions, we intuitively expect an ensemble that includes machine learning algorithms such as random forest and boosting to outperform a parametric model specified without interaction terms. One of the known strengths of random forest is the ability to detect interactions. The further away from a linear transformation we get, the more advantage the Super Learner landmarking method will have unless the analyst has prior knowledge of the non-linearities and can explicitly add them to the Cox or joint model. In the simplistic example used here with only three covariates, it was straightforward to add the three 2-way and one 3-way interaction terms into the Cox landmarking

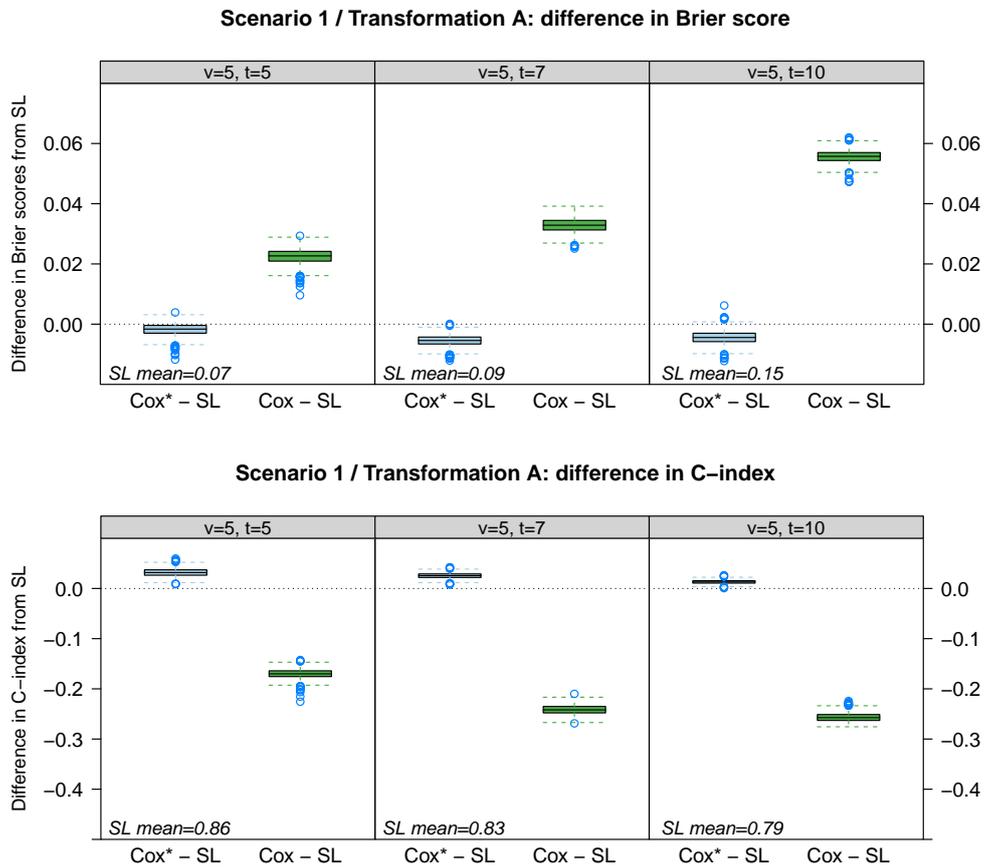


Figure 4.6: Box plots of the difference in Brier score (top) and C-index (bottom) between the Super Learner landmarking approach and two Cox landmarking approaches: “Cox” with only linear terms and “Cox*” with the addition of 2-way and 3-way interaction terms between the predictors. “SL mean” indicates the mean Brier score or C-index for the Super Learner landmark approach.

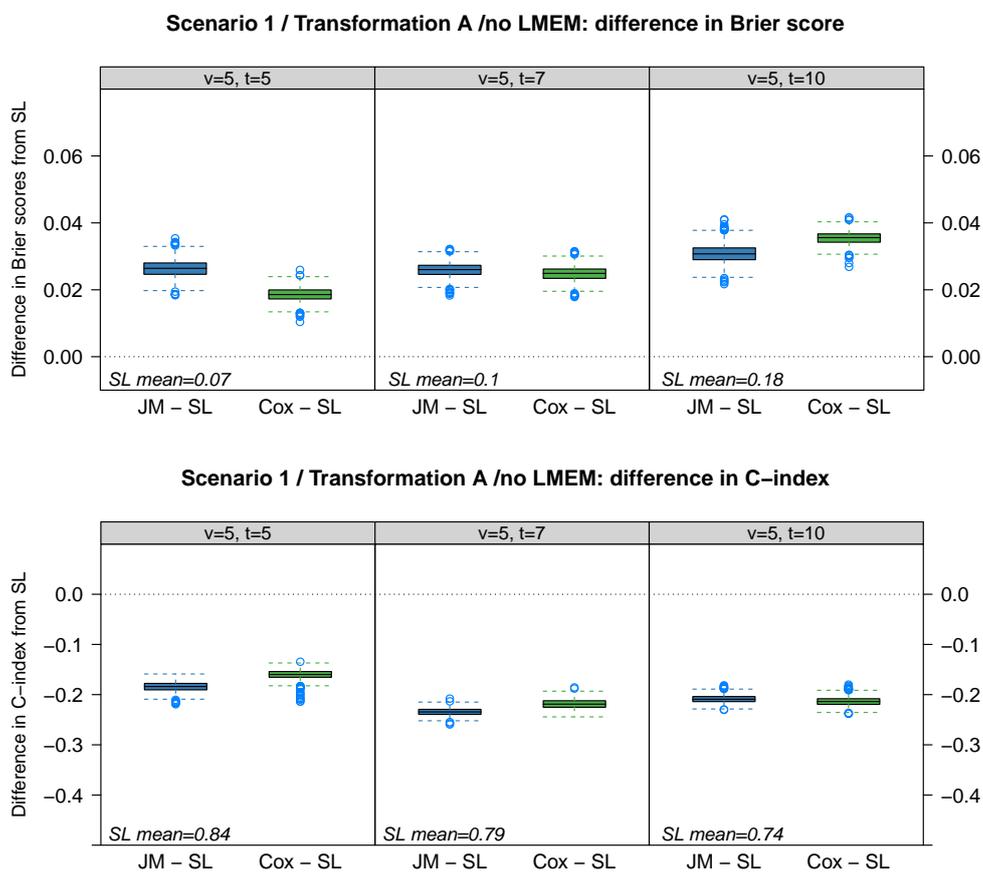


Figure 4.7: Difference in Brier score (top) and C-index (bottom) between Super Learner landmarking (SL), Cox landmarking (Cox) and joint modelling (JM). In this comparison, the raw longitudinal variable $y^*(t)$ was used instead of the predicted value from fitting a LMEM. Simulation scenario 1 had 51% events and transformation A included both non-linear and interaction terms. “SL mean” indicates the mean Brier score or C-index for the Super Learner landmark approach.

approach (referred to as Cox* in section 4.5). For analyses with a small number of covariates or clear subject matter knowledge about existing interactions, explicitly accounting for interactions may be possible, but for many analyses, there are too many predictors to add in all suspected interactions. Having algorithms that can accommodate these without explicit specification in a model serves as an insurance policy against failing to include such terms and, therefore, failing to include these effects when predicting. The transformations used here were extreme and it is possible that when working with actual data the presence of such extreme interactions would be known by subject matter experts.

That the Cox landmarking method with interaction terms equalled or even surpassed the performance of the Super Learner landmarking method is not unexpected. Because the data was generated using interactions that were subsequently included in the Cox* model, we should expect excellent performance. In Polley et al. [2011], the authors describe the “frustrating” scenario of a data-adaptive non-parametric estimator “losing” to a parametric model (even a misspecified parametric model) because the non-parametric method is more variable. This leads to the key idea behind Super Learner – we don’t know which algorithm will perform best up front, so we give ourselves the best chance of good performance by including a broad range of algorithms in our library. Had a Cox regression specified with interaction terms been included in the Super Learner library of algorithms, Super Learner would have had this method to choose from when calculating the prediction function and likely would have given it a large weighting. If so, a performance difference would not have been seen between the two methods.

In the results from scenario 2 (12% events) and transformation A, all three methods achieved a similar Brier score (SL mean = 0.01 at $t=7$) but the Super Learner landmarking showed much stronger discriminative ability. In this case, the Brier scores were all very similar because there were so few events to predict. Even a method that predicted 100% survival would achieve a respectable Brier score simply because 88% of the simulated population survived. It is only when we try to rank two individuals in terms of their risk of an event that we are able to see the performance difference of the Super Learner in such a low event scenario.

To highlight a situation where Super Learner landmark would outperform traditional methods, the longitudinal predictor was transformed. Under both transformations, interaction plus non-linearity and interaction only, the machine learning ensemble exhibited better discriminative and predictive ability than joint modelling or Cox landmarking restricted to linear terms. Although this procedure is artificial in the sense that one would never intentionally transform data in this way, it is common to need to work with data for which the data generating process is unknown and, therefore, the functional form of the relationship between it and the outcome is unknown. This transformed variable may represent data

we can observe that is acting as a surrogate for an underlying biological process that cannot be measured. The transformations applied here represented extreme cases, designed specifically to violate modelling assumptions of the two traditional methods. Before any dynamic survival prediction analysis begins, I recommend not only statistical data exploration but also discussions with subject matter experts to uncover known relationships in the data. Knowledge about the data can inform many decisions in the model building process.

Part II

Communication of Survival Predictions

Chapter 5

Communication of Survival Predictions

“I think life expectancy is... the elephant in the room, the topic everyone is thinking about but no one is talking about.”

-Clinician in CF specialist centre (2019)

5.1 Introduction

The estimated median survival age for a baby born today in the UK with cystic fibrosis is 49 years [UK Cystic Fibrosis Registry, 2020]. New treatments and improved standards of care have extended the lifetimes of people with CF dramatically and enabled them to pursue higher education and full-time employment as well as start families and take mortgages to buy homes. People with CF face a number of multi-faceted life decisions and the need to provide information about predicted prognosis to help plan for the future is increasingly important. In the clinic, these predictions can assist with provision of care and answering patient questions.

Life expectancy estimates have long been in the public domain. The Office for National Statistics in the UK publishes life expectancies for males and females on its web site, in addition to providing a life expectancy calculator [Office for National Statistics, 2021]. This calculator provides estimates of life expectancy based on current age and sex. Beyond overall population data such as this, numerous clinical prediction models exist that aim to provide survival or risk predictions in specific clinical settings such heart disease or cancer, to name two, many of which are available as publicly available web-based tools. For example, QRISK, originally developed in 2007, is an algorithm designed to predict cardiovascular disease risk in the UK [Hippisley-Cox et al., 2007]. Now in its third version, this

algorithm is publicly available on the web; based on up to 22 clinical and demographic characteristics, the calculator provides a 10-year estimated risk of heart attack or stroke. More recent work by some members of this same group produced QCOVID, an algorithm to estimate the risk of catching COVID-19 and dying; QCOVID is also available as online tool at <https://qcovid.org/> [Clift et al., 2020]. Many clinical prediction models have been developed in oncology to predict either the risk of recurrence or death. The Memorial Sloan Kettering Cancer Center web site contains a publicly accessible section with a menu of 15 different prediction tools for various types of cancer. Following surgical resection of colorectal cancer, for example, one tool uses seven pieces of input data about the patient, tumour(s) and lymph node(s) to provide a predicted probability of 5-year post-surgery survival [Weiser et al., 2011, Memorial Sloan Kettering Cancer Center, 2021]. Two web tools built by the Winton Centre for Evidence and Risk Communication, Predict and Predict Prostate, offer the option to compare survival probabilities under different treatment regimes in the breast cancer and prostate cancer settings [University of Cambridge, 2021a,b]. Clinical prediction models also exist in the chronic disease setting. For example, Celli et al. [2004] developed the BMI, Airflow Obstruction, Dyspnea, and Exercise Capacity (BODE) Index for predicting risk of death in chronic obstructive pulmonary disease with the goal of improving on the sole use of FEV1% to predict risk of death. Using four criteria, the BODE index provides a prediction of 4-year survival. An online calculator for the BODE index and hundreds of other prognostic calculators covering over 200 patient conditions is available at www.mdcalc.com, a web site created by doctors for doctors with the intent of making it easier to include such algorithms in clinicians' workflow [MDCalc, 2021].

All of these online prognostic calculators have the same high-level workflow: data is entered about the person and then a prediction or projection is produced. However, there is no standard design associated with these calculators; the style of the inputs and methods used to present the results vary widely across tools. See Figure 5.1 for a collage of four of the online calculators previously mentioned. As illustrated in these examples, sometimes the results are presented as a life expectancy, the age to which someone is expected to live, and sometimes the information is provided as a v -year survival probability, the probability of surviving an additional v years. Different graphical formats and labelling are used to convey the information: a line graph, a pictogram, a column chart, or text only. The inputs used in the calculations may be vastly different between tools as well, with some requiring detailed medical inputs such tumour size or FEV1%, and some relying only on basic information such as age and sex.

In CF, life expectancy and the median age of death at the population level and by gender is available from organisations such as the Cystic Fibrosis Trust in the



Overall Survival Probability Following Surgery

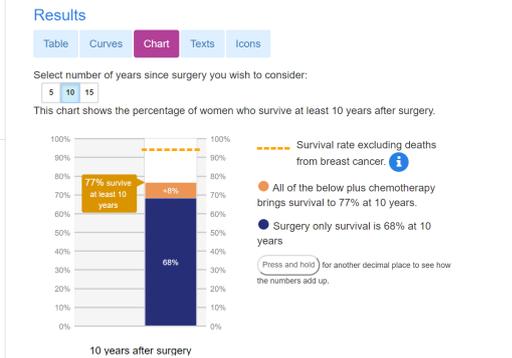
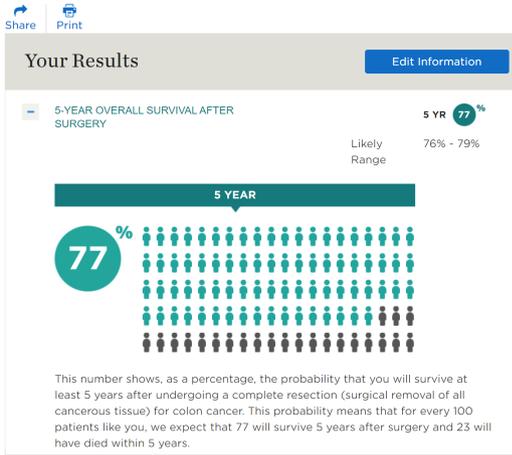


Figure 5.1: Screen captures from four publicly available online prognostic calculators. Clockwise from top-left: Life expectancy calculator for the UK population by age and gender [Office for National Statistics, 2021]. Results from post-surgery predicted 5-year survival calculator [Memorial Sloan Kettering Cancer Center, 2021]. Results from Predict model showing predicted 10-year survival for different breast cancer treatment regimes [University of Cambridge, 2021a]. 4-year predicted survival probability produced by the BODE index for people with chronic obstructive pulmonary disease [MDCalc, 2021].

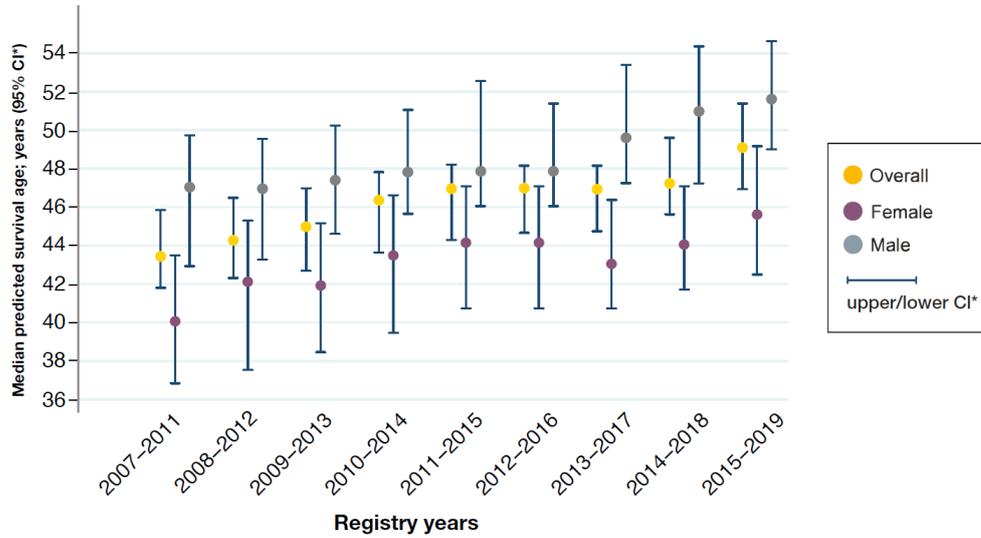


Figure 5.2: Graphic showing median predicted survival age, reproduced from the UK Cystic Fibrosis Registry Annual Data Report 2019 [UK Cystic Fibrosis Registry, 2020].

UK and the Cystic Fibrosis Foundation in the USA [UK Cystic Fibrosis Registry, 2020, Cystic Fibrosis Foundation, 2020]. Figure 5.2 shows a graphic available in the UK Cystic Fibrosis Registry Annual Data Report 2019 presenting predicted survival for rolling 5-year samples of registry data. Despite there being a number of survival prediction models in the literature (see Chapter 1 for a review), I am not aware of any of these having been translated into a publicly available tool for people with CF to obtain information about their life expectancy. (The AutoPrognosis model developed by Alaa and van der Schaar [2018] has been translated into an RShiny web app as an accompaniment to their manuscript but its availability has not been generally publicised nor can it be found by a simple internet search for cystic fibrosis life expectancy.) Szczesniak et al. [2019] developed a clinical prediction model for CF which predicts rapid decline of lung function. They have translated their algorithm into a point-of-care tool designed for use in clinic.

A possible reason for the absence of online tools for life expectancy in CF is the sensitive nature of the topic and the desire not to upset people. Chapman et al. [2005] studied end of life care for people with CF and found that discussions around death were common in early adulthood but that clinicians were sensitive to the fact that not all patients were ready for such a discussion. Conversations around survival may also happen when the clinical team is concerned about adherence or to introduce the topic of transplantation. They further note the appropriateness of such discussions with adults who are making treatment decisions. Addressing the

question directly, Keogh et al. [2019a] conducted an online survey to understand if adults with CF are accessing life expectancy information and whether they want more personalised life expectancy information. They found that the majority of respondents had received life expectancy information from one or more sources. For those who said they had not looked beyond their care team for this information, the most common reasons were because they didn't feel the information would be applicable to them or because they didn't want to know. In this survey, over 70% of the respondents said they would be interested in more personalised life expectancy. 40.3% of this group who would like to access personalised life expectancy felt this information would be best delivered by their health care team, 32.3% preferred to access it on their own and the remaining 27.4% were happy to access it either way. The authors conclude that although many adults with CF would find more personalised information useful, some find this information negative and do not want it and, therefore, the format and delivery of this information must be sensitive to both points of view.

Building on the findings of Keogh et al. [2019a], the aim of this section is to further explore the desire for and methods to present tailored life expectancy predictions to adults with CF. I began by conducting a review of methods for presenting life expectancy or risk information to patients. This informed the development of different prototype ways of presenting life expectancy information in CF. We solicited the views of health care professionals on the risks and benefits of providing this information directly to people with CF and in clinic. Using semi-structured interviews, we trialled a prototype life expectancy presentation for ease of comprehension, graphical format preferences, emotional response and level of personalisation. Themes that emerged from these interviews are summarised here and presented with representative quotes. Finally, based on the feedback to the prototype, a web app was developed that could be used to provide life expectancy information directly to people with CF. Developing a web app for use in clinic was not in scope and evaluation of the web app is reserved for future work.

5.2 Presentation of life expectancy information

5.2.1 Graphical displays

There is a rich literature comparing various graphical displays for communication of risk and/or survival information and excellent overviews can be found in Ancker et al. [2006], Spiegelhalter et al. [2011], Spiegelhalter [2017]. In looking across studies, there is no single graphical format that appeals to most people most of the time but the use of graphics has been consistently found to be useful. Also, the graphic type most preferred may not be the same as the one best comprehended.

Clarke et al. [2008] reported that a survival curve was preferred to a bar chart, pie chart or pictogram by surgical patients while Edwards et al. [2006] found bar charts were more helpful than pictograms or thermometer-style graphics amongst people with diabetes. In contrast, Downen et al. [2017] studied survival in the context of chronic kidney disease and found that a pie chart showing the percent expected to be alive/dead was preferred over a pictogram, histogram or Kaplan-Meier curve; the pictogram was the best interpreted. In a study about comprehension of different formats by cancer surgery patients, Davis et al. [2010] found that over 95% of participants correctly interpreted a simplified Kaplan-Meier curve and that 70% correctly interpreted all of the formats (pictogram, Kaplan-Meier curve, bar chart and narrative only). Good comprehension of survival curves (>90%) was also found by Rakow et al. [2012]. They also noted that long-term outcomes received more attention than short- or intermediate-term outcomes. This is consistent with the work of Fortin et al. [2001] who report that most subjects preferred lifetime estimates over 10- or 20-year time horizons, though the time horizons of most interest may be expected to differ substantially depending on the context.

Pictograms, sometimes called icon arrays, have been well studied. These typically show a grid of icons representing individuals, with colours or shading used to indicate the number that is expected to survive to a given time horizon. An example is shown in Figure 5.1 (top right). Price et al. [2007] aimed to identify characteristics of pictograms that led to the greatest speed and accuracy in interpreting risk predictions. They learned that when the number of rows and columns differs, comprehension is better in a horizontal (more columns than rows) format. In a study of the icons used in a pictogram, Schapira et al. [2001] found that icons representing humans were easier to identify with than geometric shapes and Burkell [2004] believes the only drawback to a pictogram is the space they use. Both of these studies discourage use of partial shading of icons to represent a more detailed estimate as it led to inaccurate interpretation. Both Clarke et al. [2008] and Edwards et al. [2006] received feedback that smiley-face/sad-face icons in pictograms were seen as childish and unhelpful.

All of these graphical formats also require effective labels or text. For communication of probabilities, much research has focused on the use of frequency statements. Burkell [2004] found that frequency statements of the form ‘20 out of 100’ were preferred to percentage format and that raw probability numbers such as 0.2 were to be avoided. Other studies revealed that the denominator used in the frequency statements can affect the interpretation and perceived reliability. Small denominators like ‘1 in 10’ seemed less reliable as people questioned whether the sample size was only 10 people [Schapira et al., 2001]. Frequency statements with larger denominators, however, were perceived as more risky as the larger number of events had the greatest impact on perception of risk [Burkell, 2004]. For example,

Yamagishi [1997] found that people perceived 1,286 deaths out of 10,000 people to be riskier than 24.14 deaths out of 100 people despite the latter being nearly twice as likely. In another study about the denominator of a frequency statement, Naik et al. [2012] found that a denominator of 100 or 1,000 was best understood. Both Gigerenzer [2003] and Trevena et al. [2013] emphasise the importance of making the reference class clear in all formats so that the probability is not misinterpreted. ‘20 out of 100 people will develop condition X’ begs the question, to which people is this estimate relevant? All people? People who have never had condition X before? Frequency statements make it easier to clearly define the reference class [Spiegelhalter, 2017].

5.2.2 Uncertainty

For statisticians, uncertainty is a fundamental part of our work and much of our time is spent quantifying it, for example through confidence intervals on estimates. However, communicating uncertainty to non-statisticians is often a challenge [Spiegelhalter et al., 2011]. First, there is evidence that patients have difficulty interpreting uncertainty information. Engelhardt et al. [2017] noted that the idea of a probability being correct (unbiased) yet having uncertainty was particularly difficult to understand. Muscatello et al. [2006] and Trevena et al. [2013] found that confidence intervals were generally poorly understood, even amongst a more educated audience. Also, Bowman [2019] notes that the boundaries of confidence intervals create a notion of precision that conflicts with the uncertainty one is trying to portray. Second, there is the concern that communicating uncertainty could lead to a loss of trust and confidence in the information. Han [2013] explains that ‘ambiguity aversion’ when presented with uncertainty on probabilities can reduce people’s confidence in the estimates beyond the level of the uncertainty alone. van der Bles et al. [2020] conducted an empirical study to test the validity of this concern. They found that while communication of uncertainty did lead people to find the information less certain, it did not lead to significantly lower trust in the numbers or the source when the uncertainty was presented numerically. Verbal statements about uncertainty, however, did lead to a reduction in trust. There are also ethical questions about presentation of uncertainty. Respecting the patient’s right to make choices suggests that providing uncertainty could aid their decision-making ability. Innes and Payne [2009] found that while patients wanted their clinicians to be honest, they also wanted some level of ambiguity. This ambiguity-seeking behaviour may be due to uncertainty providing a source of hope [Han, 2013]. On the other hand, if the uncertainty overwhelms the patient or leaves them overly-pessimistic about their future, perhaps this is not in the patient’s best interests.

5.2.3 Preservation of hope

Hope is recognised as being important to patient well-being [Elliott and Olver, 2009] and is inversely related to anxiety and depression [Olver, 2012]. In presentations of life expectancy, preservation of hope is an important consideration. For example, the labels appearing on the graphics are as important as the graphical style and can be directly linked to hope. Downen et al. [2017] received feedback that use of the word ‘dead’ was perceived negatively. Mozumder et al. [2018] and Kiely et al. [2010] concur and suggest presenting survival probabilities rather than death (risk) probabilities. If we are concerned about framing bias (positive framing makes the risk seem less risky), presentation of survival and mortality simultaneously is possible and is effectively what a pictogram achieves [Gigerenzer, 2003, Spiegelhalter, 2017]. Kiely et al. [2010] propose a method for presenting survival while preserving hope by providing not only the median life expectancy, but also a best case and worst case scenario. In a follow-up study, Kiely et al. [2013] found that a majority of people who were surveyed believed the three scenario presentation conveyed hope and is what they would want to hear for their own prognosis rather than simply a median. Much of this work, however, has been conducted in cancer research and reflects survival after diagnosis in contrast to our setting of a chronic disease.

5.3 Methods

5.3.1 Design and development of prototype life expectancy presentation

Based on our review of graphical displays for presenting survival information, we chose to test three main graphical formats for life expectancy information: a pictogram, a bar chart and a line graph (survival curve). In the pictogram, we used an icon similar to a bust because it clearly represents a person and is less abstract than a geometric shape. In addition to the median life expectancy, the 25th and 75th percentiles of life expectancy were also presented. Above the 10x10 pictogram graphic is a frequency statement label formatted as ‘ X out of 100 people with CF are expected to live beyond age A ’ where X is either 75, 50 or 25 and A is the corresponding survival age. Bar charts contained either 4 or 5 bars and were labelled either in age format ‘Number out of 100 expected to survive beyond age A ’, where A is the age in years, or in years format ‘Number out of 100 expected to survive X or more years (beyond Y)’, where Y is the calendar year (e.g. 2030) and X is an integer. The survival curve format was presented with several different labelling options, including best and worst case scenarios, defined as the survival ages for

the top 10% and bottom 10%. Also, because confidence bands have been found to be confusing [Trevena et al., 2013], we tested a format where other hypothetical survival curves were shown to indicate that the predicted survival curve had uncertainty. These graphics can be found in the prototype life expectancy presentation in Figure 5.3. All life expectancy numbers used in the prototype were fictitious and not based on any statistical model.

In addition to the different graphical formats, we also illustrated different levels of personalisation in the prototype. Personalisation levels included:

- no personalisation
- personalisation based on gender, genotype and age
- personalisation based on gender, genotype, age, age at diagnosis, infections in the past year, diagnosis of cystic fibrosis-related diabetes, lung function and number of days in hospital receiving IV antibiotics in the past year

For the most personalised presentation, the characteristics were selected due to their predictive ability as found in our work in the previous section. The goal was threefold: to get feedback on how much personalisation was desired, to determine whether people would be able to fill in these characteristics accurately, and to get clinician feedback on the usefulness of such a personalised prediction. The graphical displays with varying levels of personalisation were put into a presentation to illustrate how this might be implemented in a web application.

The overall design of the prototype life expectancy presentation was created using a persona approach. Originally described by Cooper [2004], personas are a set of fictional people, each representing the different goals and characteristics of a potential group of people in the target audience. Personas help focus the design on the needs and motivations of the users, a complex and heterogeneous group.

A persona consists of a name, a narrative that gives a sense of the people represented by the persona, and a set of characteristics such as age, gender, socio-economic status, social support structure and health. This allows us to imagine the situation of each persona and to project their interaction in other scenarios [Pruitt and Grundin, 2003]. While a set of personas is not intended to be exhaustive in the sense of describing every potential user, examination of the characteristics of a set of personas may help identify obvious omissions, e.g. no male personas (Ibid). Vosbergen et al. [2015] used a persona approach to help create educational content for heart disease patients in the Netherlands. Their approach used k-means cluster analysis to define the personas. More traditional approaches combine quantitative data analysis with subject matter expertise. As the main purpose of personas for our application was to elucidate a range of emotional responses, we relied primarily on experiences shared with us by researchers, clinicians, staff at the CF Trust and a person with CF.

Table 5.1 introduces the six personas created to aid the design of the prototype life expectancy presentation. They were designed to span a range of responses to seeing personalised life expectancy information. The Terry and Olivia personas both want access to this information and they want something tailored to their situation. They differ in that Olivia has a specific need for the information – she is considering starting a family. Terry’s mindset is that the more one knows, the more one can plan and shape one’s future. Terry has more severe disease than Olivia and is not as well supported. In contrast to the Terry persona is Eleanor. Eleanor is living her life in the way she believes is best and feels that knowing her estimated life expectancy will not change anything about how she lives. The Jamie persona is somewhere in between and has not decided whether this information would be reassuring or depressing. Jamie may feel comfortable looking at overall survival information for people with CF but not at personalised predictions. The Liam persona is in early adulthood and is struggling with his CF. He misses treatments and sometimes engages in risky behaviours as he tries to live a “normal” life. He is vulnerable and his clinicians worry that based on his severe disease, his predicted life expectancy may be lower than he expects, and seeing this information could lead him to give up and stop taking his medication altogether. The health care professional is represented by the Aryan persona. Aryan wants what is best for his patients. He wants to motivate them to maintain their treatment regimen, give them hope, and honestly answer their questions.

These personas suggest that there is a differing level of comfort with the idea of accessing life expectancy information, especially personalised information. Some people will not want this information at all, some people will only want very general information, some want as much detail as possible and some people could be hurt by the information. Because we would not know anything about the person accessing the life expectancy information were it to be made public on a web site, the overall design must be sensitive to all of these situations. To achieve this, we designed a layered presentation where layers represent increasingly personalised life expectancy information. People can choose whether to proceed to the next layer to receive more tailored information. The top layer contains the information everyone would see upon accessing the app and is limited to population-level life expectancy information of the type already in the public domain, such as in annual Cystic Fibrosis Registry reports. The subjects can then choose whether to continue to the next level of personalisation or to stop at the population-level data. The second layer contains survival predictions conditional on age, gender and genotype. Again, subjects choose whether to continue to the next layer or to stop there. The third and final level of personalisation contains survival predictions conditional on age, gender, genotype, age at diagnosis, lung function, cystic fibrosis-related diabetes diagnosis, infections in the past year and hospitalisations. The prototype

Table 5.1: Personas used in development of prototype life expectancy presentation.

Persona Name	Narrative	Age	Disease severity	Social Support	Education /Numeracy
<i>Persons with CF</i>					
Olivia	She’s making a big decision and her expected survival is an important input into that decision	20s / 30s	Low	Good	Medium
Liam	He’s struggling with adherence and has a negative outlook. There are concerns that this information could cause him to give up.	early 20s	High	Low	Low
Jamie	Doesn’t know if he or she wants this information	40s	Low-Med	Medium	Low-High
Eleanor	Doesn’t want to know because there’s nothing she can do about it	30s / 40s	Low-High	Medium	Low-High
Terry	Believes knowledge is empowering	late 20s	Medium	Medium	High
<i>Health care professional</i>					
Aryan	His main concern is the health of his patients. He wants to motivate them to make positive lifestyle choices.	NA	NA	NA	High

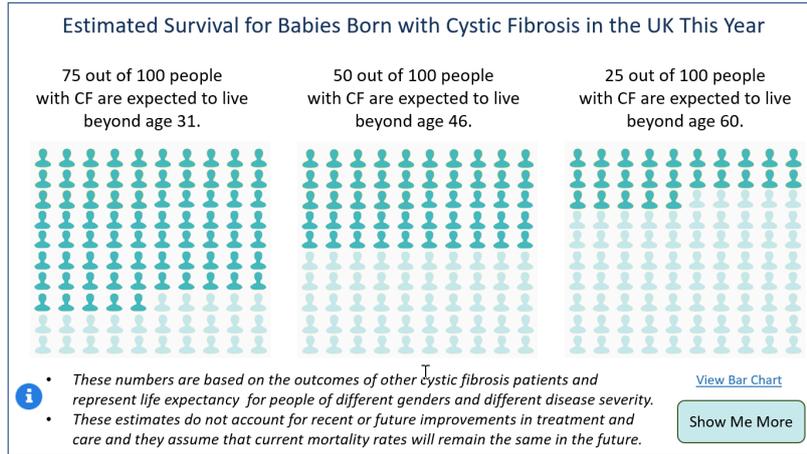
life expectancy presentation shown in the interviews can be found in Figure 5.3 and was built using Microsoft PowerPoint. Slides 1A and 1B are from the first layer; 2A, 2B and 2C are from the second layer; 3B and 3C correspond to the third layer. Because this was a paper-based mock-up of a web application, in the interviews, subjects were asked if they felt comfortable proceeding to the next level of personalisation. Note that the prototype presentation used fictitious survival information so that interviewees would not be seeing their own predicted survival.

5.3.2 Design of semi-structured interview, recruitment and analysis

The prototype presentation information described above was tested using interviews with people with CF and health care professionals working with CF patients. A combination of convenience sampling and respondent driven sampling was used to recruit participants for this study. People with CF were recruited initially through word of mouth, with a further advert posted on various social media profiles. Health care professionals were recruited by contacting specialist CF centres in the UK. Prior to commencement of the interview, all participants were asked to read a participant information sheet. In addition to background

Figure 5.3: Prototype life expectancy presentation

Slide 1A



Slide 1B

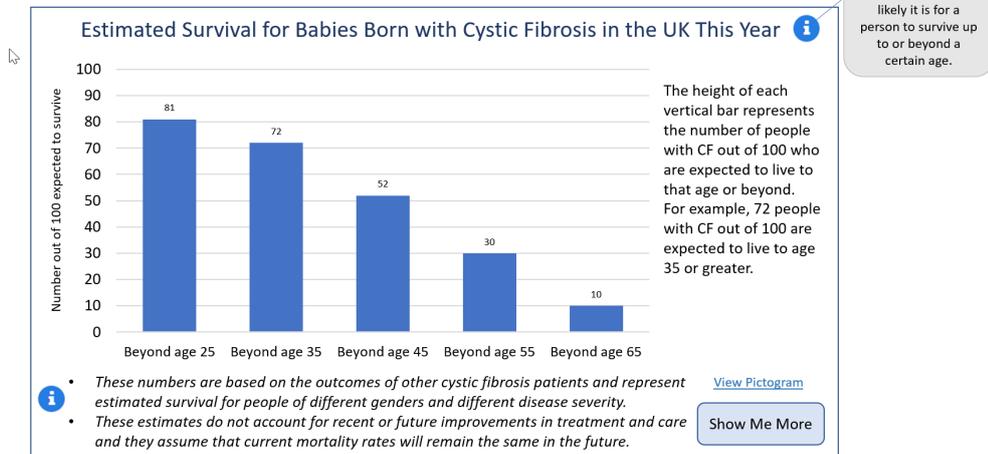
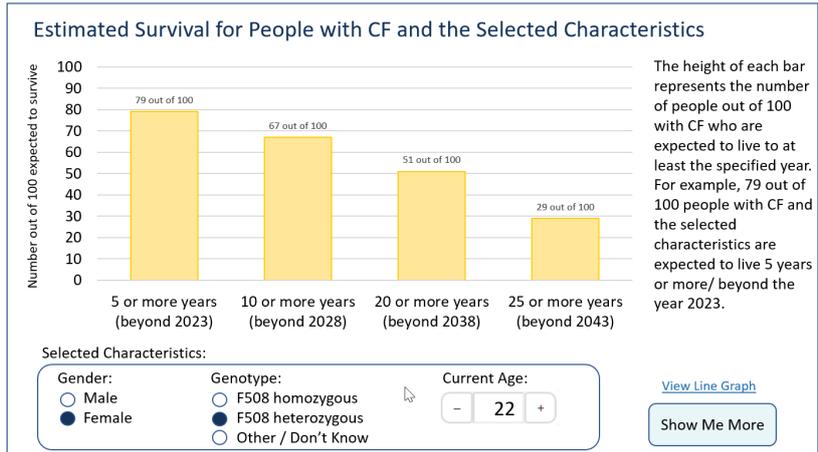


Figure 5.3: Prototype life expectancy presentation (cont'd)

Slide 2A



Slide 2B

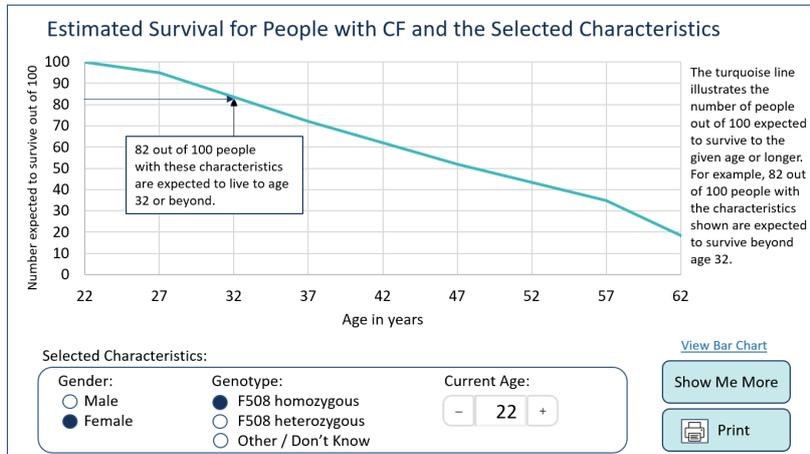
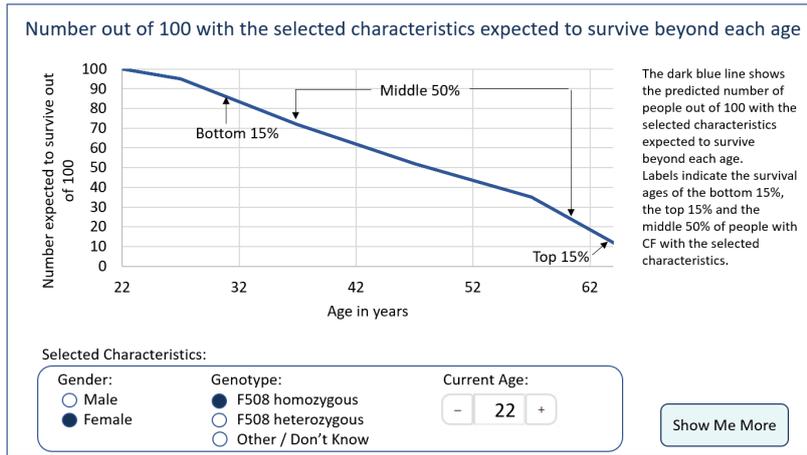


Figure 5.3: Prototype life expectancy presentation (cont'd)

Slide 2C v2



Slide 2C v3

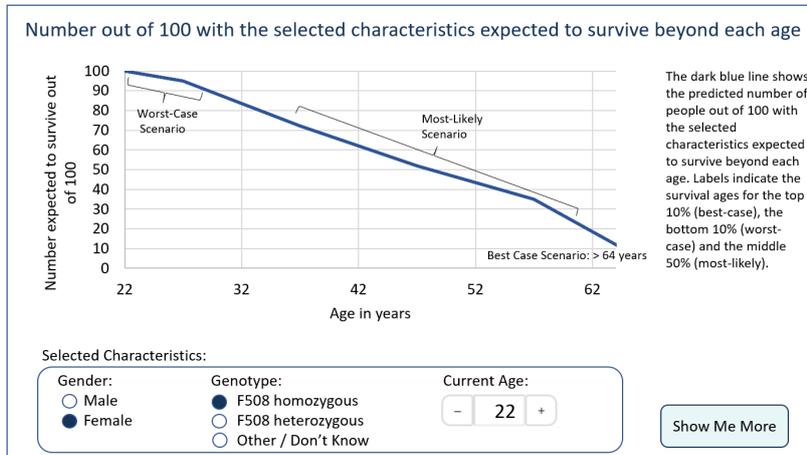
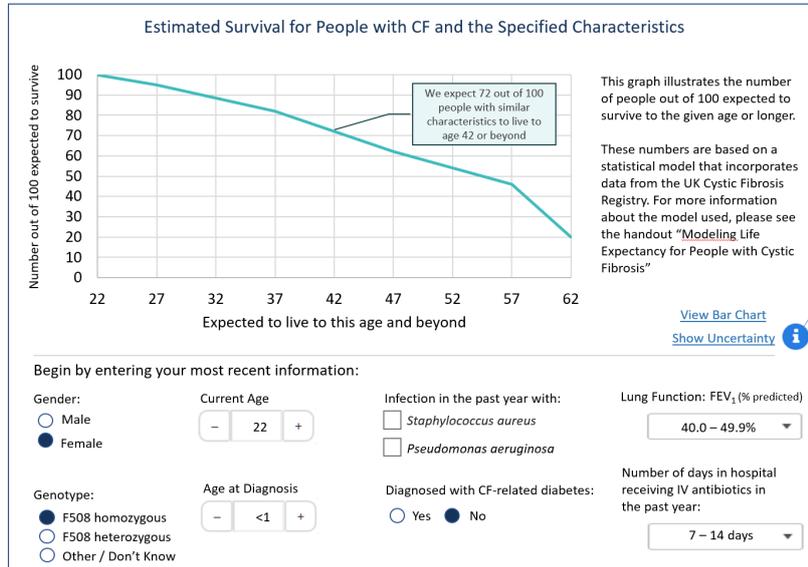


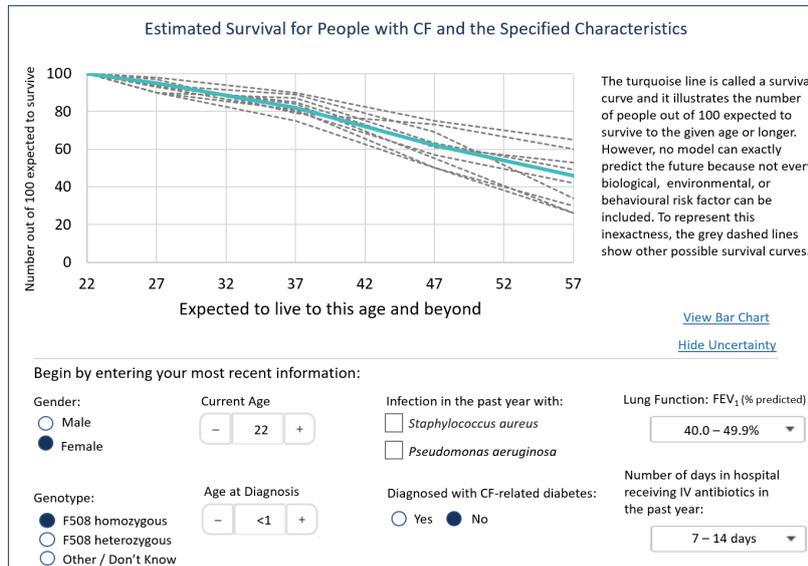
Figure 5.3: Prototype life expectancy presentation (cont'd)

Slide 3B



No model can exactly predict the future because not every risk factor can be included. We call this inexactness "uncertainty".

Slide 3C



information about the study, the information sheet explained that their participation was voluntary, they would not be compensated for participating and they could withdraw at any time. It was also explained that their responses would be anonymous and that no identifying information would be included in transcripts of the interview. Participants were then asked to sign a consent form to participate in the research study.

Topic guides were developed for the semi-structured interview with content focused on three main themes: current provision and accessibility of life expectancy data in CF, feedback on a prototype life expectancy presentation described above, and the impact of having personalised life expectancy information available. The full topic guide for an interview with a person with CF can be found in appendix D. Interviews were designed to last between 20 and 60 minutes and were conducted either in person or via video call between February and June 2019. Questions were open-ended and designed to elicit the respondent’s opinions and feelings. Topic guides differed somewhat depending on whether the subject was a person with CF (PWCF) or a health care professional (HCP). This study and the materials used were approved by the London School of Hygiene and Tropical Medicine (LSHTM) Observational / Interventional Research Ethics Committee (Reference 16138). A pilot version of the topic guide was created and tested by interviewing several colleagues in the clinical trials unit at LSHTM.

Interviews were recorded (voice-only) and transcribed. Thematic analysis of the data was guided by the approach of Braun and Clarke [2006]. Data was collated by respondent type (either PWCF or HCP) and themes were identified and refined. Representative quotes for each theme were extracted from the transcripts. For questions assessing comprehension of the prototype only, responses were coded as correct or incorrect for the analysis.

5.3.3 Development of web app

Based on the findings from our qualitative research described above, an R Shiny app [Chang et al., 2021] for presenting a subset of the survival information considered in the prototype materials was developed. Survival statistics in the web app were calculated based on Kaplan-Meier survival curves generated using annual review data from the UK CF Registry for calendar years 2013-2017, inclusive. This time frame was chosen so that the statistics would match those published in the annual UK CF Registry report. Survival curves were estimated for the total population, separately for males and females, and conditional on survival to a specified age from 20-50 years old. Survival estimates, including 95% confidence intervals, were pre-calculated at each integer year of age and loaded into the app from an R data file.

Table 5.2: ID and time of diagnosis for the seven persons with CF interviewed.

<i>ID</i>	<i>Age at diagnosis</i>	<i>ID</i>	<i>Age at diagnosis</i>
PWCF-1	child	PWCF-5	infant
PWCF-2	adult	PWCF-6	adult
PWCF-3	child	PWCF-7	infant
PWCF-4	infant		

5.4 Results

5.4.1 Respondents

Seven people with CF were interviewed. Table 5.2 shows the IDs I will use to refer to each respondent along with their time of diagnosis: as an infant (<1 year old), as a child (<12 years old) and as an adult (>18 years old). No respondents were diagnosed as teenagers in our sample. Further demographic information was not gathered to ensure anonymity.

Thirteen health care professionals (HCP-1 - HCP-13) including specialist nurses, physiotherapists, pharmacists, psychologists and CF clinicians from eight CF specialist centres in the UK were also interviewed. Identifying characteristics were not collected for the HCP interviewees to preserve their anonymity.

5.4.2 Feedback from people with CF

Desire for life expectancy information

To assess the desire for life expectancy information, respondents were asked multiple questions at different points in the interview on this topic. They were asked “What kinds of things would life expectancy information help you plan for?” as well as several questions about whether they would use an online tool containing the information they were presented with in the prototype presentation. While all of the PWCFs interviewed expressed an interest in life expectancy information in general, they had different opinions about how and when they would want to receive this information, how detailed it should be, how it would impact them personally, and how important it was. Representative quotes from the PWCF interviews, organised by theme and sub-theme, are presented in Table 5.3.

Some respondents felt they had a right to know about their life expectancy. PWCF-5 said, “I think if these numbers are out there then I really don’t think it’s moral or ethical to keep that away from people using the excuse it might make them upset.” There was a feeling that other people were making the decision about offer-

ing this information but it should be up to the people with CF to decide. PWCF-6 expressed frustration at trying to access life expectancy information which they believe exists but isn't being provided. "I am constantly asking my team and they kept going we don't know, we are not sure.... And I can't find out because I don't have access to the data."

Not all respondents have sought out this information. One person explained that, "It's not that I'm not interested [in my life expectancy], I think I just prefer not to know sometimes." (PWCF-3) Another explained that because they have a low FEV1%, their life expectancy is likely to be low and they don't want to know that. Others said that not having any information was worrying. PWCF-5 told us, "The very difficult thing is not knowing anything and anytime you cough it turns into a gloom and doom, because you don't know anything." Many respondents had an idea of life expectancy for people with CF but for several the information was outdated and for others they weren't sure where they had obtained the information.

Opinions were mixed about whether access to life expectancy information is preferred on their own via a web site or together with a member(s) of their clinical team. Those wanting independent access voiced feelings that it was personal and that they preferred to have time to reflect on the information before speaking with their doctor. Further, one respondent was unhappy with the idea that clinicians are gatekeepers of this information. On the other hand, some felt that it would be important to have the support of their clinical team to understand and process this information. "I think anything around survival is emotional and you need a bit of support and understanding, what it means and what it means for different therapies you're on or are trying." (PWCF-3) There was a concern in particular that if the predicted life expectancy was much lower than what the person had in their mind, this could have a negative impact and availability of a counsellor would be important. All agreed that receiving this information should be the decision of the PWCF. Several respondents expressed concern that if access to life expectancy information was provided in clinic, this could pressure people into having a discussion they preferred to avoid, especially in the potentially stressful situation of a clinic visit. PWCF-4 said, "I think if it were offered and the information was there in a folder waiting to be opened, someone might feel pressure to have that information."

Impact of life expectancy information

The usefulness of life expectancy information to PWCFs depends on whether they believe it applies to them. Many respondents explained that they didn't feel currently-available information was relevant to them because it did not account for their particular situation. "Because I'm already a certain age, I'm expected to live longer than I was when younger, so like all of those more complex stats and

concepts [they] get ignored or oversimplified and it's not mentioned that that's the case." (PWCF-5) CFTR mutation, age at diagnosis, current age and colonised bacteria were all cited as characteristics differentiating the respondents from the population average they believe the data is based on. Information tailored to their circumstances was perceived to be more useful. PWCF-1 said, "If I could get a figure more relevant to me, I personally would definitely try to access it and also make life decisions based around that."

Four of the seven PWCFs told us they would use personalised life expectancy information to help them plan for the future. The types of things they needed to plan for included: their pension (2 people), a mortgage (1 person), work / career (2 people), travel plans (1 person) and starting a family (1 person). PWCF-2 explained, "Yeah, it [personalised life expectancy information] would help me plan, . . . I ask do I get a mortgage, how long do I get the loan for? Its constantly there in my mind, you know thinking about my pension provision. It factors into every major decision really." Two other PWCFs had the opinion that life expectancy information could be used to inspire positive lifestyle changes or adherence to daily therapy. They believed it would be motivational and challenge them to follow best personal care practices. In contrast, one respondent reflected on their difficult time in early adulthood and felt that news about a limited life expectancy would have had the opposite effect; it would have led them to feel more isolated and less engaged with their care because it was a time when they were so desperate to be normal.

Table 5.3: Themes, sub-themes and representative quotes from semi-structured interviews with people with CF.

Theme / Sub-theme	Quote
<i>I. Desire for information on life expectancy</i>	
	"I think if these numbers are out there then I really don't think its moral or ethical to keep that away from people using the excuse it might make them upset. CF is upset[ting] and there's this idea of let's keep this from people so they don't think about it." (PWCF-5)
PWCFs have the right to know	"If the data exists it shouldn't be up to people like doctors and people not living with CF to make and keep this from us." (PWCF-5)
	"I am constantly asking my team and they kept going we don't know, we are not sure. They don't want to say or they don't know. And I can't find out because I don't have access to the data." (PWCF-6)
Not knowing is difficult	"..the very difficult thing is not knowing anything and anytime you cough it turns into a gloom and doom, because you don't know anything." (PWCF-5)
Not knowing is preferred	"It's not that I'm not interested, I think I just prefer not to know sometimes." (PWCF-3)

II. Receiving life expectancy information

Prefer to receive from my clinical team	“I think anything around survival is emotional and you need a bit of support and understanding.... And if I was just looking at myself, I’d struggle to interpret, and I’d have many questions” (PWCF-3)
	“I think it would have to be importantly when a counsellor was immediately available for any fallout that might occur” (PWCF-4)
Prefer to access on my own	“The people who suffer should be able to access it without having to go through a third party like the clinician really.” (PWCF-1)
	“...the appointments are stressful, you’re worried and scared, and you’re in hospital. It’s not a relaxing environment. While it would be good to have them explain things that aren’t clear, I would definitely want to look at it alone, and a few times to process and then I’d go to ask questions.” (PWCF-5)
Access must be optional	“Some people might want to ignore it. If it’s part of the clinical setting, it seems prescribed like you have to do it, and if you have had tests done you don’t want to hear your life expectancy is looking different now.” (PWCF-5)
	“I think if it were offered and the information was there in a folder waiting to be opened, someone might feel pressure to have that information.” (PWCF-4)

III. Impact and use of life expectancy information

Use to plan for the future	“It would help me plan, . . . I ask do I get a mortgage, how long do I get the loan for? It’s constantly there in my mind, you know, thinking about my pension provision. It factors into every major decision really.” (PWCF-2)
	“I did think, I’m past the [median survival] age so I’m pretty close to the limit. Do I carry on living? go travelling? tell my family? plan my finances?” (PWCF-6)
Use for motivation or to modify lifestyle	“But if you knew there were other factors coming into play like people did specific things to keep them well, a specific drug, then it would make me more motivated with that behaviour. . . . If there was anything modifiable, that’s what I would do.” (PWCF-3)

Prototype life expectancy presentation

All of the PWCF participants correctly answered questions that required them to interpret the provided life expectancy graphics. When asked which graphical format they preferred, pictogram or bar chart, all chose the pictogram. When offered a choice between pictogram, bar chart or line graph, the line graph and pictogram were both liked. Some noticed that the line graph format contains more information; PWCF-4 said, “I really like this line graph in that hypothetically you could find any age and any number.” In response to questions about the maximum age they wanted to see life expectancy for, responses varied. Retirement age was cited as important by multiple respondents. Fewer labels/text were preferred on

the graph as were labels that referred to survival by age rather than by year.

In the graphics, survival was presented as ‘the number of people out of 100 expected to survive’. At greater ages, this number is quite low and we asked participants how they felt about seeing information that at a given age, only 10 out of 100 people were expected to survive. Two respondents felt that information should not be shown when the percent survival was this low. Two felt that the information was encouraging because it showed that some people with CF do live to an advanced age. The other three had mixed feelings. PWCF-1 said, “It is quite hard hitting obviously, but again I think it’s down to personal preference. I don’t mind seeing that but I can imagine [some] people will.” It was also suggested that this information may be particularly stressful for younger people.

When asked if they would use an app that gave life expectancy information for people with CF based on their age, sex and genotype, three respondents said yes. One said no because they felt their situation was too different from the average and one said no because it would not change how they were living their life. The other respondent said they wanted to see the information but did not believe it would apply to them because of their unusual genotype.

We also asked if they would want more tailored life expectancy information that was based on age, sex, genotype, lung function, age at diagnosis, respiratory infections and number of days in hospital receiving IV antibiotics. Four participants said yes, one said no and one was sceptical that it still would not be personalised enough to be relevant. (The seventh participant was not asked this question.) When asked if they would be able to fill in all of the information requested for the more personalised prediction, six said yes and one said they knew most of the information. Five respondents said they did not believe this information was too personalised. PWCF-5 said, “No I think it’s still quite general, I mean, I assume there will be quite a few people in the same position.” They also offered feedback on the characteristics used with some suggesting we add BMI, amount of physical exercise, and more different respiratory infection types.

The participants expressed varying degrees of familiarity with the concept of uncertainty. After a brief explanation of uncertainty, four people said they preferred the survival line graph with the addition of uncertainty although they preferred uncertainty shown as a shaded region rather than individual possible trajectories. PWCF-3 explained how the uncertainty was cause for optimism: “I think it makes it a bit more positive really. If you can see there’s uncertainty you know you can be a bit different and change things and be a bit lucky.”

5.4.3 Feedback from health care professionals

Impact of life expectancy information

Most of the 13 HCPs we interviewed saw both positive and negative aspects of the presentation of life expectancy information to the people with CF they care for. Representative quotes from the HCP interviews, organised by theme and sub-theme, are presented in Table 5.4. Several respondents voiced concerns about how this information might impact people and their motivation to keep fighting and to stick to their treatment plan, particularly if the estimated survival they saw was worse than what they had expected. HCP-2 explained, “They might go away disheartened. And think there’s no point in continuing treatment.” There was particular concern for younger people as they may simply accept the number they see and not question it or realise that it is not a certainty. It was acknowledged that each person’s reaction to this information will be unique. HCP-12 said, “For some people this will be very helpful and reassuring but for other people it can be very distressing so the range of different reactions can happen.” This was at the heart of the HCPs concerns – the reaction of the more vulnerable people who might access this information.

HCPs also told us they thought this information could empower people with CF and help answer some of their questions. In clinic, it could also help to give them a “good starting point for a conversation” (HCP-3). When asked if more tailored models of life expectancy would be helpful in their ability to provide care, some HCPs mentioned the lack of information available for people who have lived well into adulthood. Many PWCFs are still recalling a life expectancy number they learned as children and it has not been updated for their current age. HCP-1 explained, “We do have a [PWCF] who said ‘When I was a child, I was told I would be dead by 30. Now I’m 30 and I don’t know what to do. I thought I was going to be dead. What do I do now?’ And it feels tragic that that information has come to light only when they are 30.”

Some HCPs also told us that life expectancy information could be used to motivate people to adhere to treatment regimens and give them hope. HCP-6 explained how they would interpret the pictogram information to provide hope. “I would say there’s 100 people and this number will live to 47 but it doesn’t say if you are or aren’t one of those people. There’s nothing to say you are this or that so let’s assume you’re that person, [someone who has survived], and that’s how you motivate them to keep going. You have got to give them hope, you’ve got to be positive because nobody knows.” Several other HCPs also voiced their idea of using this information, particularly the graphic showing uncertainty in the estimates, to encourage positive behaviours. HCP-4 thought, “If you’re presented with something like this [life expectancy with uncertainty] and you say to the

patient, ‘potentially if you do your treatment, if you do your physio, and your medicine, we could be looking at the higher end of the spectrum [rather] than the lower end.’”

Access, personalisation and format of life expectancy information

HCPs were unanimous that information consistent with the full prototype life expectancy presentation should not be made publicly available without any safeguards for access. Some were in favour of sharing information based on a limited set of characteristics (e.g. age, sex, genotype) publicly and some were not in favour of sharing this information at any level of detail publicly. HCPs were primarily concerned with the impact that this information would have on their patients’ mental well-being but some also expressed concerns that the predictions were inaccurate and, therefore, should not be shared. Referring to the graphic with estimated survival based on age, sex and genotype, HCP-11 said, “It feels like a lot to handle especially if they are looking at it without the right support and understanding. Maybe better to discuss this with the team.... It depends on their outlook and their health beliefs and their resilience and their anxieties and mental health state.” HCP-13 felt more strongly and told us, “I would not be happy with my patients being able to see this.” There was a general feeling that patients who wished to know more about their life expectancy should be given that information but with support available. Those concerned about the validity of the estimates cited the use of median versus mean, the assumption that treatments would not change over time and the concern that population data included a large percentage of non-adherent people with CF which the HCP perceived could underestimate the probability of survival.

When asked about use of current information on expected survival of people with CF (which refers to survival from birth), a number of HCPs told us that the figures were too general and did not apply to their group of patients or to any specific patient due to their failure to account for current age and the large variability of disease progression in CF. HCP-10 voiced the concern, “The problem is, the figure bounded around is not relevant to adults in the clinic”. HCP-2 told us, “I’m always a bit cagey about any population data to inform patients with CF because it depends on genotype to a point, it depends on how compliant they are with treatment, it depends on drop in FEV1. Because two identical patients with same genes and treatment could look very different in 10 years.” In the context of an application that could be used in clinic, many HCPs said this would be helpful, especially in terms of providing more personalised life expectancy information to use in conversations with their patients. HCP-1 said, “I’m looking for personalized info, something more specific than the overall median life expectancy... that doesn’t tell you anything about the individual sitting in front you who has got features

that can push them one way or another.... As a clinician, you are always dealing with an individual patient not a group of people.” HCPs also expressed a need to be cautious about the idea of ‘personalised’ predictions. Some thought it made the predictions seem too certain and others felt that it was not possible to provide a personalised prediction based on a handful of characteristics. “So it’s the risk of introducing a new fancy interactive tool on a website that it’s then perceived as more of a certainty which obviously it isn’t although it is providing more accurate data.” (HCP-12)

Regarding the design of the life expectancy information as shown to the respondents in the prototype, nearly all of the HCPs said they preferred the pictogram to the bar chart if they would be showing it to their patients. For their own use, more preferred the line graph and bar chart presentations because they contain more information. Additionally, several pointed out that the declining survival probabilities at older ages were more obvious in the line chart and might be negative. HCP-1 explained, “...you can extrapolate the line down to zero and that’s something that might be distressing to people with CF.” Most HCPs believed that PWCFs would be able to fill out the majority of the characteristics asked for in slides 3B and 3C but they expressed concerns about some of the inputs. Some felt that PWCFs would not know their genotype or would be unfamiliar with the scientific terms homozygous/heterozygous. Also, if the PWCF did not remember their last lung function measurement correctly, the results could be wrong. The input for IV days was also a concern in terms of people correctly entering the information. Finally, as newborn screening was introduced in the UK in 2007, the age at diagnosis input may become less valuable over time and could bias predictions if pre-2007 data were used to train the algorithm to make post-2007 predictions. These concerns were cited as an additional reason that HCPs recommended against providing this level of detail on a public web site.

HCPs were also generally in favour of showing a measure of uncertainty in the predictions. Some thought it could assist them in motivating people in their treatments and lifestyle choices and others thought it was important to illustrate that models are not perfect. HCP-1 said, “I think its important to include a measure of uncertainty by doing that it’s a reminder for people that this is only a prediction and not set in stone.” HCPs preferred a more traditional presentation of uncertainty as a confidence interval than the possible trajectories presented in slide 3C. There was also some apprehension about the statistical nature of uncertainty and whether it was adding too much complexity. HCP-10 told us, “I think it’s the right thing to do ... but I think it therefore starts to become more complicated and how these things are calculated it can raise questions that may not be answered.”

Table 5.4: Themes, sub-themes and representative quotes from semi-structured interviews with health care professionals.

Theme / Sub-theme	Quote
<i>I. Life expectancy information could be damaging to PWCF</i>	
It could cause them to	“They might go away disheartened. And think there’s no point in continuing treatment.” (HCP-2)
disengage from their treatment regimen	“This can make our lives as clinicians very difficult. I’ve seen patients die of depression and I’ve seen patients die of discouragement because they just disengage and stop.” (HCP-13)
Some people may be more negatively affected	“I think in the majority of the cases if this was publicly available it wouldn’t be a problem but it’s just the small percentage of vulnerable patients accessing the information at the wrong time and it just creating a lot of anguish for them.” (HCP-9)
	“For some people this will be very helpful and reassuring but for other people it can be very distressing so the range of different reactions can happen.” (HCP-12)
<i>II. Life expectancy information could be empowering and motivating</i>	
Knowledge empowers people	“As a clinician, I’m in favour of these things because they empower people. They give you a good starting point for a conversation.” (HCP-3)
	“Access to it would significantly change conversations we have with people with CF and, potentially, their life decisions.” (HCP-1)
Information may help HCPs provide hope	“I would say there’s 100 people and this number will live to 47 but it doesn’t say if you are or aren’t one of those people. There’s nothing to say you are this or that so lets assume you’re that person [someone who has survived] and that’s how you motivate them to keep going. You have got to give them hope, you’ve got to be positive because nobody knows.” (HCP-6)
Information may motivate positive behaviour	“If you’re presented with something like this [life expectancy with uncertainty] and you say, ‘potentially if you do your treatment, if you do your physio, and your medicine, we could be looking at the higher end of the spectrum [rather] than the lower end.’” (HCP-4)
	“Really helpful to tell them what their median survival is dependent on age and allowing them to plan but also giving them a wake up call to stick to treatment.” (HCP-9)
	“I think it’s interesting for some of our patients who have a negative view or impression that they may not live that long and are non-compliant and take risky behaviours, so this would be good to challenge their thoughts and support them with this.” (HCP-11)
Some people didn’t expect to still be alive	“We do have a [PWCF] who said ‘When I was a child, I was told I would be dead by 30. Now I’m 30 and I don’t know what to do. I thought I was going to be dead. What do I do now?’ And it feels tragic that that info has come to light only when they are 30.” (HCP-1)

III. Access to personalised life expectancy information

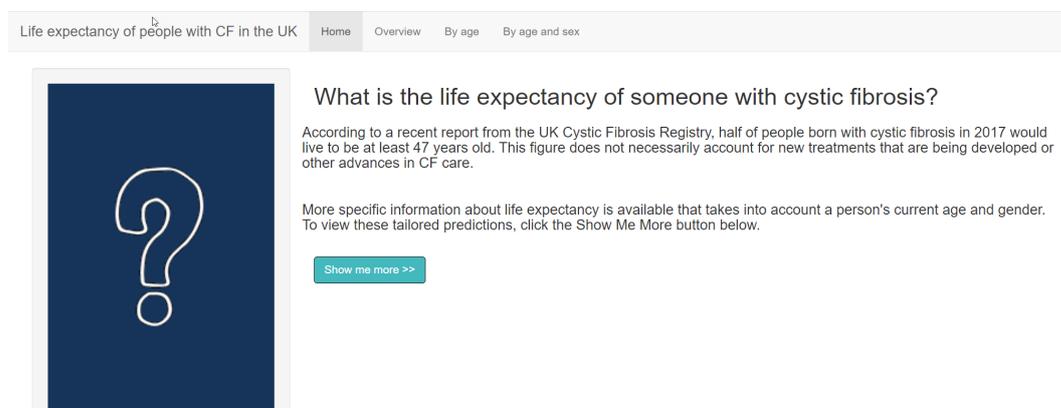
Needs to be controlled	“If you can access this via [a public web site] then this is even more dangerous because there is no one there to say, “Remember...” ... I would not be happy with my patients being able to see this.” (HCP-13)
	“I think that some patients may feel a bad result or short life expectancy has taken hope away but that’s why a careful conversation with a clinician to discuss what their options are is crucial and why they shouldn’t be accessing the super personalized information at home because there’s the danger that they may just give up on the basis of a result whereas if they were with a clinician they could be guided on what to do next, and what they could do about it.” (HCP-1)
	“I think I have an idea of who would be robust enough to take in the info and who might not.” (HCP-2)
	“I think people have a right to see the info that’s out there and the evidence that’s out there.... The way it was available and the safeguards and messages around that would need to be carefully thought about.” (HCP-12)

IV. Personalisation

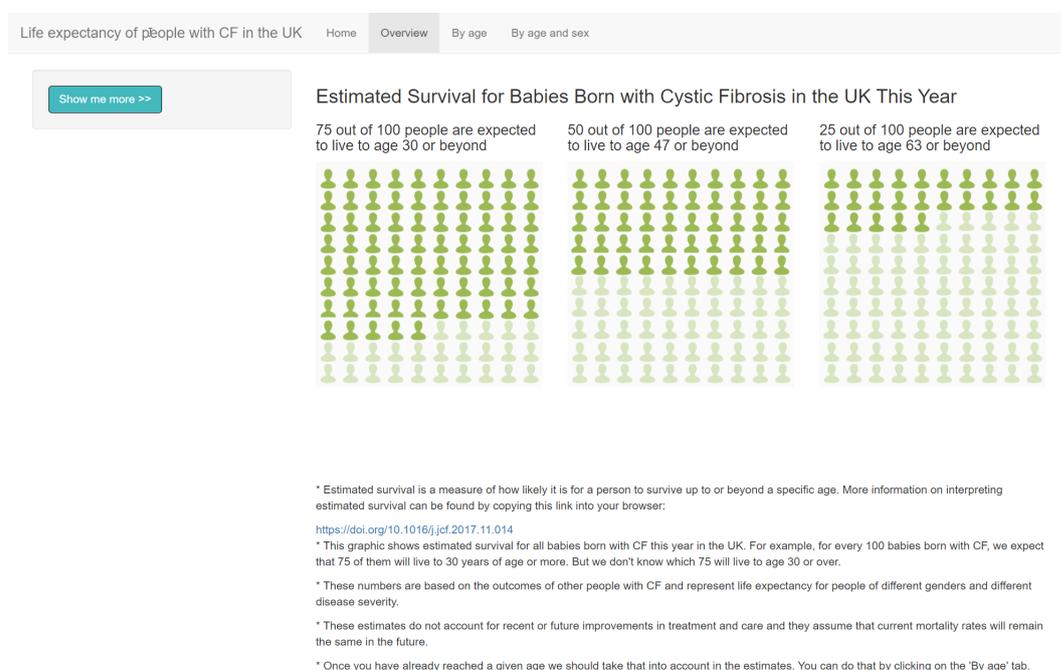
Current information is too general	“The problem is, the figure bounded around is not relevant to adults in the clinic.” (HCP-10)
	“I’m always a bit cagey about any population data to inform patients with CF because it depends on genotype to a point, it depends on how compliant they are with treatment, it depends in drop in FEV1, because two identical patients with same genes and treatment could look very different in 10 years.” (HCP-2)
Personalised information would be helpful	“I’m looking for personalised info, something more specific than the overall median life expectancy... that doesn’t tell you anything about the individual sitting in front you who has got features that can push them one way or another.... As a clinician, you are always dealing with an individual patient not a group of people.” (HCP-1)
	“The clinical team would love it, I think especially if you’re worried someone is declining and you are worried about them maybe needing a transplant or something. You can say to someone who’s 52 and in 5 years time they could be down here, you can start having those conversations about do you want a transplant.” (HCP-8)

5.5 Translation of results into life expectancy web app

Based on HCP and PWCF feedback on the prototype life expectancy presentation as well as input from the Cystic Fibrosis Trust, a corresponding prototype web app was developed. There are four screens available for viewing: Home, Overview, By



(a) Home page



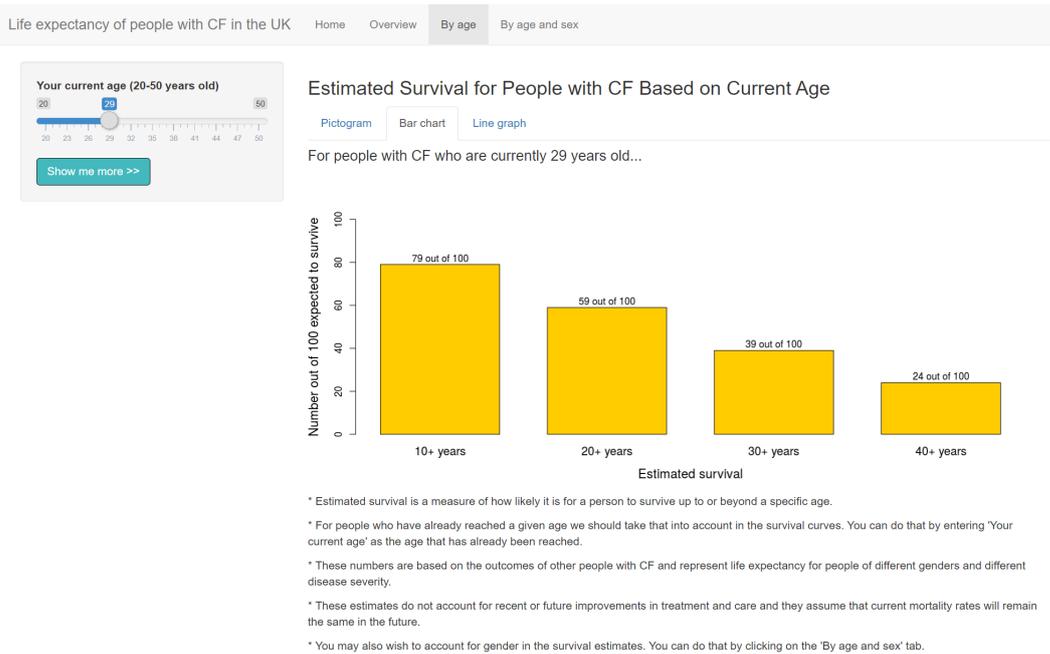
(b) Overview page.

Figure 5.4: Landing page of the web app (top) designed to reconfirm the user's desire for more information on life expectancy. A turquoise 'Show me more' button takes the user to the next screen. Alternatively, top tabs can be used for navigation. Overview page (bottom) provides median life expectancy for babies born today in the UK as well as the 25th and 75th percentile estimates using a pictogram.

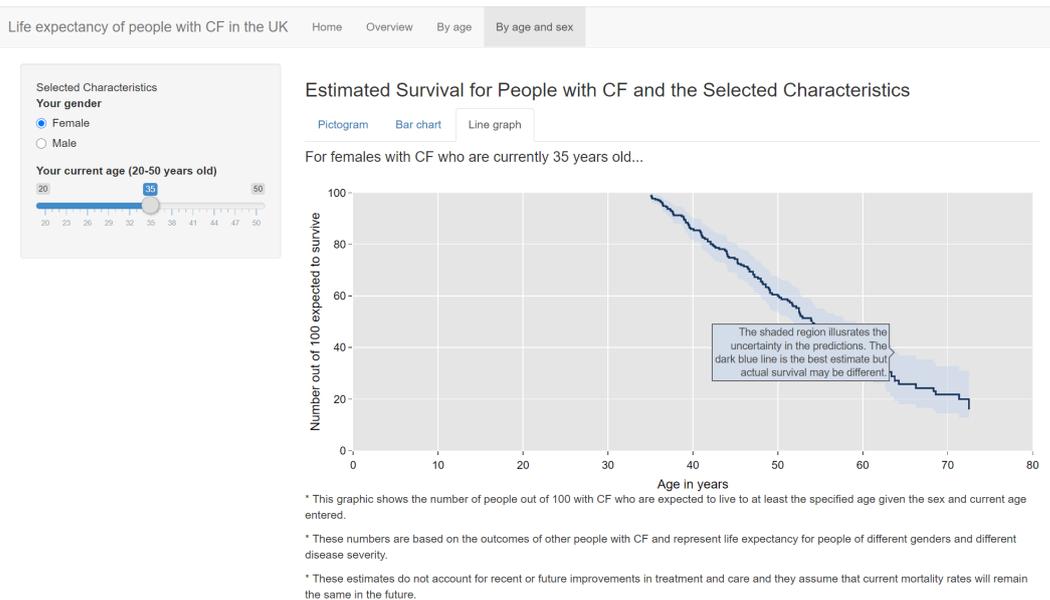
age, and By age and sex, representing increasingly more personalised information. These screens can be reached successively via the ‘Show me more’ button. The web app differed from the prototype life expectancy presentation in several ways. First, to prevent people from inadvertently seeing life expectancy information, a home page was added that requires the user to reconfirm that they want to access survival predictions. Also, because many users felt the pictogram was the simplest presentation format, we designed each page to begin with a pictogram display of the information. Additional disclosure text was added to each page as well as a link to information about how to interpret estimated survival. Colours were chosen based on the branding specifications of the CF Trust web site.

In keeping with the layered approach of revealing increasingly tailored predictions (described in section 5.3.1), after the Home screen, users are first presented with population-level information (‘Overview’ tab) and then have the option of seeing estimated survival for people of their current age (‘By age’ tab), and then estimated survival for people of their current age and sex (‘By age and sex’ tab). Because different respondents preferred different graphical formats, all were incorporated into the web app. Estimates conditional on age or age and sex can be viewed as pictograms, bar charts or line graphs by selecting the tab corresponding to the desired display type. For the line graph presentation of a survival curve, uncertainty in the model estimates is shown with a shaded band around the survival curve and is always displayed. Text appears explaining uncertainty when the user hovers over the area with their mouse. If the mouse is on the survival curve itself, text appears providing the number out of 100 that are expected to live beyond the age where the mouse is.

We limited the characteristics people can personalise the predictions with to current age and sex. Survival estimates based on genotype were not provided for two reasons. First, some PWCFs do not know their genotype. Also, while F508del mutations are the most common, 48.6% of the UK population have two copies of this mutation and 89.7% have one copy [UK Cystic Fibrosis Registry, 2020], many other mutations exist. People who are F508del heterozygous or have two non-F508del mutations form a heterogeneous group of people with both severe and mild forms of CF. Survival predictions for these heterogeneous groups will be less tailored and so, to avoid confusion, an option for specifying genotype was not included in the web app. We also omitted options to specify more detailed characteristics such as lung function and bacterial infections as shown on slides 3B and 3C in the prototype presentation (Figure 5.3). This was in response to concerns by HCPs that predictions with this level of detail should not be available outside of the clinic and because several PWCFs said they would prefer to receive this information in the company of someone from their clinical team.



(a) Life expectancy conditional on age



(b) Life expectancy conditional on age and sex

Figure 5.5: The 'By age' page (top) provides predicted survival information conditional on the user's current age. The 'By age and sex' tab (bottom) provides life expectancy information conditional on the user's current age and gender. Pictogram, bar chart or line graph display formats may be chosen. The line chart offers additional information when users hover their mouse over the shaded region or the survival curve itself.

5.6 Discussion

We undertook this study to learn how to present life expectancy information to people with CF in a sensitive and accessible way. Information about life expectancy is emotive and we saw varying levels of desire for this information by PWCFs. One of the key messages was that receipt of this information must be at the person's request. Even including life expectancy as an optional discussion topic at the annual review could exert pressure to have a conversation on the topic. On the other hand, some people feel frustrated that information is being kept from them and that others are making the decision about whether it would upset them. Our interviews also revealed a broad range of opinions about how this information might be provided and what its impact would be. There is no 'typical person with CF' and there is no single presentation of life expectancy information that will suit all. Amongst health care professionals interviewed, there was agreement that while tailored life expectancy information could be valuable, it would be best delivered within the clinical setting where support and clarifications could be offered.

An interesting outcome of the interviews with persons with CF was the feeling many had that current population median life expectancy information did not apply to them because their personal characteristics were different. This finding is consistent with the online survey of Keogh et al. [2019a]. For example, two of the respondents were diagnosed with CF in adulthood and we can surmise they have a milder form of CF; their life expectancy is not likely to be the same as someone diagnosed at birth who has suffered from lung disease and respiratory infections since childhood. The idea of personalised or tailored survival predictions was appealing to the participants as the information may be more relevant to their individual situation. However, the term 'personalised' may imply a level of certainty or a level of customisation that cannot be offered. Either care must be taken to explain what personalisation means or this term should be substituted for something more generic such as 'tailored' or 'group-specific'.

We also must contend with data limitations when attempting to provide personalised life expectancies. Considering genotype, for example, many genetic mutations in CF are rare and there is not sufficient data to create a predictive model specific to them all. Using the most common mutation, F508del, we could provide predictions for those who are F508del homozygous and those who are not, but the latter group would contain a mixture of people with severe and mild genotypes who will not have a homogeneous disease progression. One possible solution is to include more characteristics. By modelling recent lung function and other health indicators that genotype affects, we may overcome the need to model genotype itself.

The notion of a 'personalised' prediction is controversial. Taylor-Robinson and Kee [2019] describe an "inconvenient truth in the risk prediction industry" that the

predictions produced by clinical prediction models are not for any specific individual. Rather, they are predictions for a group of people with similar characteristics to the individual, suggesting that the label ‘personalised’ is misleading. Further, what models are able to achieve in terms of prediction precision for a group of people may not translate to predictions for individuals [Altman and Royston, 2000]. An example given in Smith [2011] of a group-level prediction not always translating to the individual is of a 100-year old healthy woman who has smoked for 93 years. We know that smoking causes lung cancer but that is not the same thing as saying that every smoker will get lung cancer. Even if we could put a probability on getting lung cancer, any given individual either will or will not get it and we cannot say which people will fall into which group with certainty. Smith [2011] explains that what is almost random at the individual level may be predictable at the population level. Several of the PWCFs and HCPs interviewed seemed to have an intuitive sense of this. HCP-11 commented “Even within genotype and gender and bugs grown, people’s paths are different and you can’t predict that different things may or may not happen.”

If personalised predictions actually pertain to a group (not an individual) and they lack accuracy, are they useful? The first issue is solvable by a label change and appropriate explanation. For the second problem, we must focus on the goal. Our goal in this work is to provide life expectancy information to people with CF (who want this information) to help them consider the future. One of the most powerful scenarios in favour of providing survival predictions that are conditional on individual characteristics involves conditioning on current age. We saw in the interviews that several PWCFs were not aware that their life expectancy is different based on their survival to age 20, 30, 40, etc. One person wondered what they should do now that they are past the quoted median survival age (which refers to survival from birth). Life expectancy information for the group of people with the same sex and current age may help answer their questions and dispel the notion that they are past some limit. While this does not represent a truly personalised prediction as it is only specific to two characteristics, it represents a large step forward in terms of information they currently have access to.

When discussing survival, studies have found that patients want their doctors to be honest and to treat their situation individually [Hagerty et al., 2005, Parker et al., 2007] but clinicians may not feel armed with sufficient information. Health care professionals have seen prognostic uncertainty first hand, both through patients who exceeded expectations and through those who declined more rapidly than expected. This uncertainty and how to communicate it to the patient is a key reason these discussions are challenging [Holloway et al., 2013, Parvez et al., 2015, Brighton and Bristowe, 2016]. Clinicians draw on their experiences and on the information available to them but currently there are no tools for providing

life expectancy information based on the characteristics of a particular patient. Several HCPs interviewed thought a tool such as the one proposed here could be a valuable resource for them in preparing for a conversation about survival and in helping to motivate patients and provide hope.

The biggest limitation of this work is the low number of participants with CF. We were only able to interview seven participants and the respondent-driven sampling approach means that we were more likely to interact with people involved with and aware of scientific research in CF. Our sample likely excludes more vulnerable patients and those with a lower level of education and socio-economic status. The aim of this work was not to attempt to achieve consensus amongst the UK population of people with CF on whether tailored life expectancy information should be made available. There will always be those who do not want to know and those who do. This study was a beginning in understanding how such information could be provided in a sensitive and accessible way and towards that end, the differing circumstances and opinions of our sample provided many insights. Note also that paediatric patients were explicitly excluded in this study as the life expectancy presentation is not intended for children. We do acknowledge, however, that if the app is to be released to the public, we would not be able to prevent young people from looking at it without some type of verification process.

It is the hope that the work presented here on communication of life expectancy will be the beginning of a larger undertaking. (See Figure 5.6) Although not in scope for this thesis, future work could include publishing the web app and surveying adults with CF on their feelings about having such an app, whether they would use it and if they find the information presented sensitively and clearly. Based on the results, the web app may be able to be released to the public, possibly with permission rules or passcode access. One idea is to let the CF clinicians provide a PWCF with the passcode and offer to review the information together. The PWCF could accept this offer or choose to view the information privately at home. The clinician would be able to counsel vulnerable patients against accessing the information if they could be harmed by viewing it. This would avoid inadvertent access to the information while browsing and could discourage use by minors. Development of a tool for health care professionals to use in clinic, either by themselves or with patients, is a second area for future exploration.

Based on this study and the previous work done to assess desire for life expectancy information, we believe that providing predicted survival information, tailored by certain characteristics, would benefit many people wanting to know more about what to expect. The fact that some people want access is sufficient that we should endeavour to find a way to provide it. The remaining question is how to provide access while balancing the needs of PWCFs who do not want this information and the risks inherent in vulnerable people accessing this information.

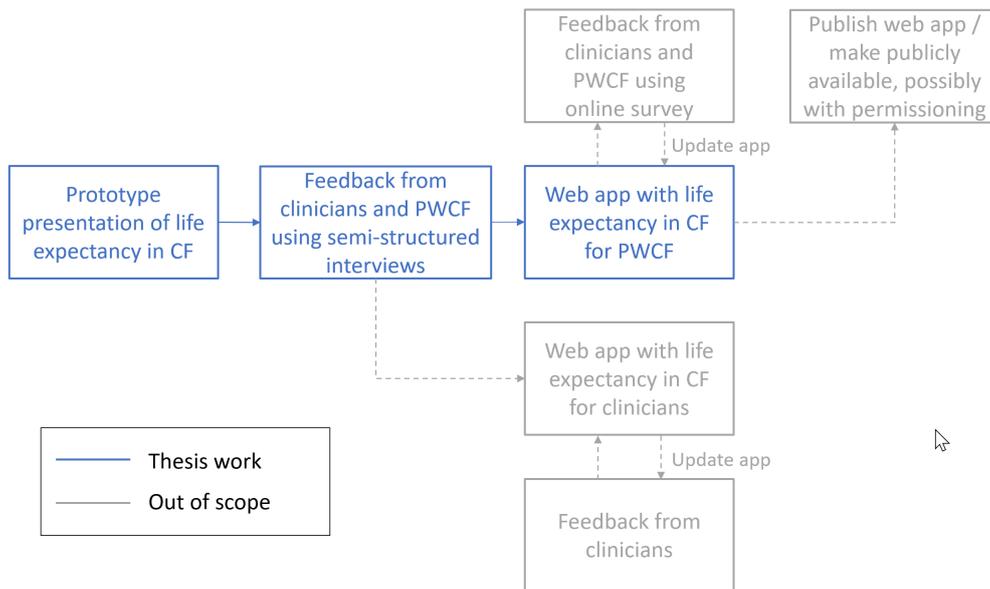


Figure 5.6: Illustration of this work (blue) and ideas for a larger project that could ultimately lead to the creation of an online tool for people with CF and/or health care professionals. All items in grey are out of scope for this thesis but represent possible future directions for research.

5.7 Contributions

This section includes a description of some work performed jointly with Fahad Malik, a qualitative researcher and Ruth Keogh, my PhD supervisor. Specifically, Dr Malik prepared the ethics submission to obtain permission to conduct interviews with clinicians and people with CF. This included creation of a consent form and participant information sheet. I performed the literature review of preferred graphical formats and labels, ability to understand graphics by a broad audience and best practices in survival communication. Based on that review, I designed a paper-based sample presentation of life expectancy information and accompanying script to be used in the semi-structured interviews with clinicians and people with cystic fibrosis (CF). Dr Malik prepared the overall interview scripts including information about our study, current access to information, understanding how the interviewee felt about accessing life expectancy information and incorporation of the paper-based model. Of the 7 people with CF interviewed, Dr Malik interviewed 6 of them; Ruth Keogh and I interviewed one. Of the health care professionals interviewed, Dr Malik interviewed 7, Ruth Keogh interviewed 3 and I interviewed 3. Dr Malik transcribed all of the recorded interviews. I used thematic analysis to summarise key themes expressed in the interviews and selected repre-

sentative quotes. Using the results of the interviews, I updated the paper-based life expectancy presentation and built an R Shiny online tool for presentation of life expectancy conditional on age and sex. Throughout this process, Ruth Keogh provided supervision and helpful ideas and comments.

5.8 Funding

The work in this section of the thesis was funded by a Vertex Pharmaceuticals, Inc. Circle of Care award, awarded to the Cystic Fibrosis Trust and Ruth Keogh.

Part III

Mediation Analysis for Survival Outcomes

Chapter 6

Cystic Fibrosis-Related Diabetes

6.1 Introduction

Improvements in our understanding of the biological impact of CFTR mutations and in management of the pulmonary infections and nutritional needs of people with CF have allowed many to enjoy longer lives. However, along with a greater life expectancy has come an increased risk of many co-morbidities. CF complications in adults can affect the pancreas, gall bladder, GI tract, kidneys, liver, bones and joints, skin and also mental well-being [Regard et al., 2019]. The most common of these complications is cystic fibrosis-related diabetes (CFRD). CFRD is associated with worse lung function [Lanng et al., 1992, Koch et al., 2001, Taylor-Robinson et al., 2012, Kerem et al., 2014], increased respiratory infections [Marshall et al., 2005, Brennan et al., 2007, Lehoux Dubois et al., 2017] poor nutritional status [Lanng et al., 1992, Koch et al., 2001, Marshall et al., 2005], and ultimately greater mortality [Milla et al., 2005, Moran et al., 2009, Chamnan et al., 2010, Lewis et al., 2015]. The mechanisms for these effects of CFRD are not yet fully understood. In Part III of this thesis, using CFRD as the motivating example, I explore causal mediation techniques for identifying mechanisms through which an exposure affects survival. First, in chapter 6, the aim is to review the current state of knowledge about the effect of CFRD on survival and provide a descriptive analysis based on multi-state models and other techniques. In chapter 7, I review two recently proposed methods for mediation analysis in the setting of longitudinal data and a survival outcome and apply them to the question of the extent to which lung function, respiratory infections and/or nutritional status mediate the effects of CFRD on survival. Finally, in chapter 8, a simulation study is used to compare the performance of the methods under different scenarios, to investigate possible sources of bias in these analysis methods and to assess the impact of some observational data characteristics.

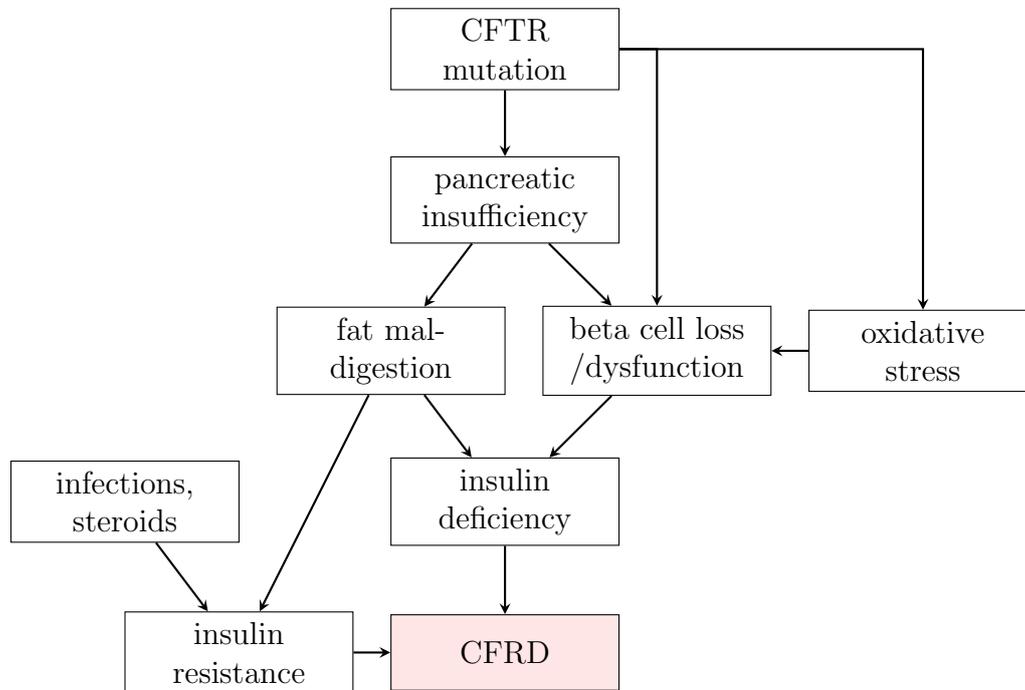
6.2 Cystic fibrosis-related diabetes

CFRD is one of the most common co-morbidities of CF affecting 32% of adults in the UK [Cystic Fibrosis Trust, 2017]. Risk factors for developing CFRD include female gender, age, more severe genotypes, reduced lung function, liver disease, pancreatic insufficiency and use of corticosteroids [Adler et al., 2008]. Prevalence increases with age and is estimated at 45-50% for those over age 40 [Moran et al., 2009]. This section provides an introduction to CFRD aimed at a non-clinical audience. More thorough treatment of the pathophysiology, care, prevalence and impacts can be found in excellent overviews of CFRD by Moran et al. [2009], Kayani et al. [2018], Moran et al. [2018], Granados et al. [2019], Frost et al. [2020].

Although CFRD has similarities to both type 1 and type 2 diabetes mellitus, it is a distinct form of diabetes requiring unique management. Like type 1 diabetes, CFRD is fundamentally characterised by a lack of insulin. People with CFRD may also have a degree of insulin resistance, which may become more severe during exacerbations [O’Riordan and Moran, 2015]. The onset of CFRD is typically gradual [Kayani et al., 2018, Granados et al., 2019]. Because most people with CFRD experience few, if any, symptoms, screening is considered essential [Moheet and Moran, 2017, Moran et al., 2018] and annual screening from 10 years of age is recommended in the UK by the National Institute for Health and Care Excellence [National Institute for Health and Care Excellence (NICE), 2017]. Even with screening, CFRD can be difficult to diagnose because of the temporary hyperglycaemia common in people with CF, especially during periods of acute illness [Bridges et al., 2018, Moran et al., 2018]. People with CF often exhibit a daily rhythm of abnormal glucose levels with increases after meals and few are considered to have normal glycaemic control [Bridges et al., 2018, Moran et al., 2018, Frost et al., 2020]. Although CFRD is rarely diagnosed in pre-adolescents, abnormal glucose tolerance may be present in infants [Granados et al., 2019]. In contrast to other forms of diabetes, pulmonary failure, not macrovascular complications, is the main cause of death for people with CFRD [Moran et al., 2018, Granados et al., 2019].

CFRD is associated both with more severe CF-causing genotypes and with pancreatic insufficiency but the pathophysiology is complex and may include a direct effect of the CFTR mutation on insulin secretion [Marshall et al., 2005, Adler et al., 2008, Lewis et al., 2015]. Figure 6.1 illustrates some of the pathways believed to be involved in development of CFRD. Pancreatic insufficiency is present in 85% of adults and CFRD occurs most frequently in those with pancreatic insufficiency [Kayani et al., 2018, Bridges et al., 2018, Regard et al., 2019]. Damage to the pancreas causes a loss of beta islet cells, the cells responsible for insulin production and release [Kayani et al., 2018]. The CFTR mutation itself leads to oxidative stress (too many free radicals compared to antioxidants) that also likely affects

Figure 6.1: Some of the physiological processes associated with development of CFRD. CFRD is primarily characterised by a deficiency in insulin, the hormone that regulates blood sugar levels. Additionally, people with CFRD may be insulin resistant meaning the body does not respond properly to insulin.



beta cells and, over time, may lead to characteristic decline in glucose tolerance (Ibid.). Decreased absorption (malabsorption) of fat also contributes to CFRD via a hormonal pathway associated with poor glycaemic control after meals [Granados et al., 2019]. Because insulin resistance increases during periods of acute infection and/or when taking corticosteroids, these also play a role in CFRD (Ibid.).

6.3 Cystic fibrosis-related diabetes and survival

A number of studies have found that CFRD negatively impacts survival. Using a cohort of 1,081 CF patients between 1987 and 2002, Milla et al. [2005] observed an increased burden of CFRD in females. They reported that the median survival age for females without diabetes was 47.0 years compared to only 30.7 years for females with diabetes. The median survival ages for males with and without diabetes were not significantly different from each other at 47.4 and 49.5 years, respectively. However, this Kaplan-Meier analysis did not account for the time-dependent nature of CFRD. Using data from three consecutive 5-year intervals through 2008,

Moran et al. [2009] found that female mortality with CFRD decreased by 50% from the period 1992-1997 to 2003-2008 and that there was no difference in mortality between females and males diagnosed with CFRD in the period 2003-2008. Overall CFRD mortality had decreased from the mid-1990s to the mid-2000s but mortality rates in those with CFRD still exceeded mortality in non-diabetic people with CF. Using UK CF Registry data from 1996-2005, Chamnan et al. [2010] used multivariate Cox regression on baseline data to show that increased hazard of death was associated with lower FEV1%, female gender, lower BMI, infection with *S. aureus* or *B. cepacia*, use of corticosteroids and diagnosis of CFRD. They reported an adjusted hazard ratio for CFRD of 1.31 [95% CI:1.03,1.67]. They also computed age-adjusted mortality rates of 4.2 [95% CI:3.4,5.1] per 100 person-years for those with CFRD compared to 1.5 [95% CI:1.3,1.7] per 100 person-years for those without diabetes. These mortality rates were directly standardised using the 2005 population in England and Wales. In a more recent study, Lewis et al. [2015] hypothesised that the decline in CFRD mortality they had seen from 1992-1997 to 1998-2002 and again in 2003-2008 would continue as a result of increased screening leading to earlier diagnosis and treatment of CFRD. For the period 2008-2012, they found age-adjusted mortality for CFRD patients to be 1.8 [95% CI:1,3] per 100 person-years versus 0.5 [95% CI:0.3,0.9] for those without diabetes, rates similar to those found in 2003-2008. The continued decline in mortality they predicted was not seen. The difference in mortality rates between those with and without CFRD was greatest in adults over 30. Lewis et al. [2015] also reported that the significant increase in hazard of death associated with CFRD was still seen when separately performing the analysis by genotype categories. This finding allowed them to exclude the possibility that the increased mortality in the CFRD group was due solely to the increased proportion of people with CFRD having a deleterious severe genotype. Although there is evidence that CFRD affects mortality, the mechanisms for this effect are not clear. One theory is that CFRD negatively affects lung function which then affects survival. Other theories involve increased respiratory infections and compromised nutritional status in those with CFRD. Each is considered in turn.

Although CF affects multiple systems in the body, its effect on pulmonary function is the key driver of mortality. The relationship between CFRD and poor lung function is well documented. In a study of over 7,500 people from the European Epidemiologic Registry of Cystic Fibrosis, Koch et al. [2001] reported significantly lower lung function using both FEV1% and FVC% measurements for all age groups when comparing those with CFRD to those without CFRD. The same pattern was seen in analyses restricted only to people with F508del homozygous genotype and when patients were separated by presence of certain pathogens. Using logistic regression analysis and a population from the European Cystic Fi-

brosis Society Patient Registry, Kerem et al. [2014] found that the odds ratio of severe lung disease (defined as FEV1% < 40%) for those with CFRD compared to those without CFRD was 1.8 [95% CI:1.6,2.2] after adjustment for age, BMI, pancreatic status and infection with *P. aeruginosa*. Additionally, FEV1% was estimated to be 8.0 [95% CI:6.6,9.5] percentage points lower for those with CFRD than those without CFRD after adjustment for age, country, sex, BMI, pancreatic status, genotype and infection with *P. aeruginosa*. Taylor-Robinson et al. [2012] noted that while CFRD was associated with a lower absolute value of FEV1%, (point estimate -2.5% [95% CI: -3.6% , -1.4%]), it was not associated with a steeper rate of decline in FEV1% over time in a model adjusted for age, cohort, pancreatic insufficiency and infection with *P. aeruginosa* using the Danish Cystic Fibrosis register. Similarly, Edwards et al. [2019] found that those with CFRD had an average FEV1% 5.5 percentage points lower than those without CFRD using UK CF Registry data. There is also some evidence that this decline in lung function begins prior to diagnosis of diabetes. Lannig et al. [1992] found significant differences in FEV1% between pre-diabetic and control patients four years prior to diagnosis after adjusting for genotype and infections. These results could be explained by declining lung function causing CFRD or by pre-diabetic conditions causing pulmonary decline. Pitocco et al. [2012] suggest the lung is a “target organ” in diabetes; a decline in respiratory function is also seen in non-CF people with type 1 and type 2 diabetes. Hyperglycaemia may affect the lung by reducing its elasticity and thickening the lining which reduces diffusion of oxygen (Ibid.). Additionally, reduced lean body mass and increased respiratory infections can impact lung function [Frost et al., 2019, Granados et al., 2019].

Because good nutritional status is associated with better survival and lung function, management of this is a key component in CF care. CF often leads to fat and protein malabsorption and this, combined with the increased energy requirements of the disease itself may necessitate enzyme and/or calorie supplementation [Taylor and Connolly, 2015]. For people with CF who also have CFRD, nutritional status can be further compromised. Comparing people with CF with CFRD to those without CFRD using the European Epidemiologic Registry of Cystic Fibrosis, Koch et al. [2001] found that BMI was lower for all age groups over 15 years of age in the CFRD group. Similarly, Marshall et al. [2005] found people with CFRD attained lower height-for-age and weight-for-age percentiles compared to those without CFRD in a study population of individuals over the age of 13 in the Epidemiologic Study of Cystic Fibrosis register. Lannig et al. [1992] found that differences in weight manifest themselves up to 4 years prior to diagnosis with CFRD. This compromised nutritional status can impair lung function by reducing lean body mass necessary for strong respiratory musculature [Pitocco et al., 2012, Granados et al., 2019]. Malnutrition could also affect the body’s ability to repair

tissue in the lungs after infection [Van Sambeek et al., 2015]. The need to maintain weight leads to a recommended diet high in calories and higher in fat and salt than would be seen in the recommended diets for type 1 or 2 diabetics [Frost et al., 2019].

Another hypothesis concerning the impact of CFRD on mortality is that it leads to more respiratory infections. CF is characterised by a “vicious cycle of infection and inflammation” [Horsley et al., 2015]. CF leads to mucus in the lungs which is a favourable environment for bacterial growth and colonisation; bacterial infection causes inflammation which damages the airway and leads to a further build-up of mucus (Ibid.). In CFRD, lack of insulin causes high blood glucose but there is mixed evidence about the impact of this on respiratory infection. Brennan et al. [2007] found that when blood glucose was elevated, presence of airway glucose was more frequent and airway glucose was associated with increased growth of common respiratory bacteria *in vitro*. Van Sambeek et al. [2015], however, did not find a correlation between CFRD and levels of sputum glucose although they and Marshall et al. [2005] both found that CFRD subjects had more pulmonary exacerbations than non-CFRD people. Lehoux Dubois et al. [2017] found that bacterial colonisation of *S. maltophilia*, *P. aeruginosa* and a group of other pathogens increased in people with poor glycaemic control compared to those with normal glucose tolerance but the difference was only significant for *S. maltophilia*. Unlike with FEV1% and BMI, Lanng et al. [1992] found no evidence for a difference in infections in the pre-diabetic state.

People with CFRD have an insulin deficiency and insulin therapy is the recommended treatment [Moran et al., 2018]. Insulin therapy helps regulate blood sugar levels and improve the absorption of nutrients, but in CFRD, insulin’s anabolic effects are critical. Insulin deficiency is associated with protein catabolism and possible loss of muscle mass; insulin, an anabolic hormone, may reverse these effects and improve nutritional status (Ibid.). Initiation of insulin has been found to reverse the declines in BMI seen in people with abnormal glucose tolerance [Koch et al., 2001, Mohan et al., 2009, Moran et al., 2009, Frost et al., 2018]. Also, insulin has been found to initially improve lung function as measured by FEV1% but the improvement was not maintained [Mohan et al., 2009, Frost et al., 2018]. Although Mohan et al. [2009] reports that the rate of decline in FEV1% is the same after initiating insulin treatment, Edwards et al. [2019] found that insulin decreased the rate of decline by a small but significant amount. There is also hope that CFTR modulators may help reduce the impact of CFRD and the risk of developing CFRD. CFTR modulators are mutation-specific therapies designed to correct the function of the protein encoded by the CFTR gene. Ivacaftor, also known as Kalydeco, is one such therapy. In a study of 1,256 (411) people treated with ivacaftor in the US (UK) and 6,200 (2,069) comparator patients, Bessonova

et al. [2018] found the relative risk of diagnosis of CFRD in a one-year period was lower in the group taking ivacaftor than in the non-ivacaftor group with relative risk of 0.77 [95% CI:0.70,0.84] (0.71 [95% CI:0.58,0.87]).

Although there is strong evidence that CFRD is associated with increased mortality, reduced lung function and nutritional status, and increased number of pulmonary exacerbations, the specific mechanisms through which these associations arise are not fully understood. Screening and treatment for CFRD have led to reduced mortality over time but they have not closed the gap. Lewis et al. [2015] hypothesises that even with insulin treatment, there continues to be an inflammatory burden affecting mortality. In the next section, I analyse the UK CF Registry using a variety of techniques to describe the relationship between CFRD and survival.

6.4 Descriptive analysis of CFRD and survival in the UK CF Registry

6.4.1 Study population

Data were extracted from the UK CF Registry on 11,427 people with an annual review between 1/1/2008 and 31/12/2017. Limiting the study population to those aged 18 and over reduced the study sample to 6,592 individuals of whom 6,487 had information about whether or not they have two copies of the F508del CFTR mutation. Particular attention is paid to this CFTR mutation as it is found on at least one copy of the CFTR gene in 90% of people with CF [Boyle and De Boeck, 2013] and it is associated with more severe disease. This period prevalent cohort contains both people who have already been diagnosed with CFRD prior to the start of follow-up and people who will be diagnosed with CFRD during the study period. 1,975 people had been diagnosed with CFRD prior to the beginning of our study period. Although these prevalent cases make up a substantial proportion of the study population, their data will not be used for most of the analyses. Because there is no diagnosis date for some of those already diagnosed with CFRD and to avoid including people with very different onset times, the focus is on incident cases only in the majority of the analyses [Brookmeyer and Gail, 1987, Vonesh et al., 2000]. After excluding the prevalent cases, the incident cohort consisted of 4,512 adults with CF and 29,940 annual review records. During the 10-year study period, 1,032 individuals (23%) were diagnosed with CFRD and there were 587 deaths or lung transplants (13%).

The data have a low level of missingness overall. Less than 1% of respiratory infection and hospital IV days data were missing and values were filled in using the last observation carried forward and by setting the value to zero, respectively.

For BMI and FEV1%, 3% and 17% of values were missing, respectively, and we used the last observation carried forward to fill these in.

6.4.2 Methods and implementation

The aim is to perform a descriptive analysis of the association between CFRD and survival using standard methods, as a pre-cursor to the mediation analyses in the next chapter. These descriptive analyses will inform interpretation of the more complex analyses, and also allow for verification of associations that have been found in earlier studies. Multiple methods were used to describe the association between CFRD and mortality.

First, Poisson regression was used to estimate age-specific mortality rates by sex and CFRD status per 100 person-years. Using the 29,940 annual review records from the incident cohort, a person-years table was constructed by gender and CFRD status per integer year of age. For example, consider a female who was first observed at age 24.5, diagnosed with CFRD at age 26.6 and died at age 28.8. She would contribute person-years at ages 24, 25 and 26 without CFRD and at ages 26, 27 and 28 with CFRD as shown in table 6.1. Individuals who were censored contribute person-years up to their censoring time. Our mortality rate model is a Poisson regression with the addition of an offset equal to the log-person years in the linear predictor. The mortality rate equals μ/t where μ is the number of events and t is the person-years of follow up. We can write:

$$\log(\mu_i) = \log(t_i) + \beta_0 + \beta_{CFRD} * CFRD_i + \beta_{sex} * sex_i + \beta_{age} * age_i + \beta_{age^2} * age_i^2 \quad (6.1)$$

where i indexes the row in the person-years table. Age was centred at 30 years (the average age was 30.26). Sex was included in the model because life expectancy for a person with CF differs by sex [UK Cystic Fibrosis Registry, 2020] and because some studies have found that females have worse outcomes with CFRD [Milla et al., 2005, Chamnan et al., 2010]. Including interaction terms for CFRD with age or CFRD with sex did not significantly improve the fit of the model so these terms were not included in the analysis. Overdispersion was checked by examination of the ratio of the residual deviance to the degrees of freedom and by comparing the fit of a negative binomial regression by checking whether the dispersion parameter was significantly different from 0.

Cox regression analysis was used to investigate whether CFRD is associated with the hazard of death. Three analyses were performed: one with baseline-only values of the predictors and of CFRD from the period prevalent cohort and two with time-varying values of the predictors and of CFRD from the incident cohort. The analyses were adjusted for sex, BMI, FEV1%, FVC%, F508del homozygous (Y/N), diagnosed by neonatal screening (Y/N), *B. cepacia* infection, *S. aureus*

Table 6.1: Sample person-years calculation for a female first seen in the data at age 24.5 who was subsequently diagnosed with CFRD at age 26.6 and died at age 28.8.

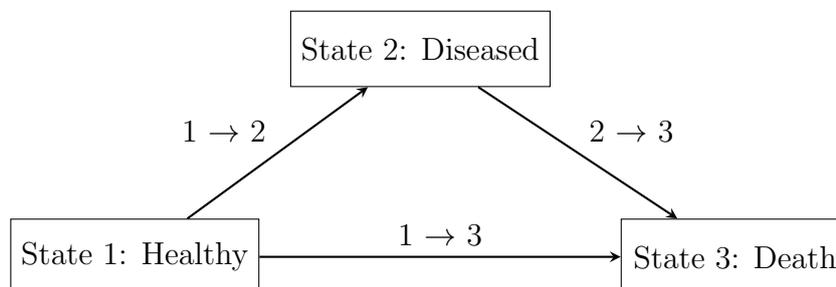
Start	Stop		CFRD = No		CFRD = Yes	
Age	Age	Sex	Person-years	Events	Person-years	Events
24.5	25.0	F	0.5	0	-	-
25.0	26.0	F	1.0	0	-	-
26.0	27.0	F	0.6	0	0.4	0
27.0	28.0	F	-	-	1.0	0
28.0	28.8	F	-	-	0.8	1

infection, number of days in hospital receiving IV antibiotics, corticosteroid use (Y/N), and liver disease (Y/N). The timescale was age. The two Cox analyses with time-updated covariates differed in that one used CFRD (Yes/No) as a time-dependent covariate and the other included the duration of CFRD as a covariate where duration equals 0 until the time of CFRD diagnosis. In these extended Cox analyses, each time-varying measurement was accompanied by a start and stop time equivalent to the annual review date where the measurement was taken and the next known annual review date or date of event / censoring, respectively. Proportional hazards was assessed by examining plots of scaled Schoenfeld residuals, testing each covariate for a parameter equal to zero in a term that interacts with time, and a global chi-square test [Grambsch and Therneau, 1994].

I also use multi-state modelling to investigate the association of CFRD with mortality or transplant. Multi-state models allow for the analysis of a history of events and are particularly useful for estimating the probability of a future event when the occurrence of an intermediate event may affect prognosis. Excellent overviews of multi-state modelling methods can be found in Putter et al. [2007] and Geskus [2016]. A three-state illness-death model is specified where individuals may transition from a healthy state to a diseased state or to death; diseased individuals may only transition to death. (See Figure 6.2) In the CFRD setting, the “Healthy” state contains the CF study population that has not yet been diagnosed with CFRD. The “Illness” state represents a diagnosis of CFRD and the “Death” state represents the composite outcome of death or transplant.

Three different model specifications are investigated, two using a clock-forward approach (“Clock-forward model” and “Clock-forward + duration model”) and one using a clock-reset approach (“Clock-reset model”). The Clock-forward model uses age as the time scale and it is hypothesised that advancement to CFRD and death depends most strongly on age. This model assumes that progression from CFRD to death does not depend on the sojourn time spent in the CFRD state and all

Figure 6.2: The classic illness-death model is defined by three states and three transitions. Death is an absorbing state. Individuals may transition from the healthy state directly to death or transition to a diseased state and then death.



times are with respect to the age of the individual at their first observation in the study population. The transition intensity from state g to state h can be modelled using standard models such as the Cox proportional hazards model. Assuming a Markov model with a clock-forward time scale, the hazard of transitioning from state g to state h at time t can be written:

$$\lambda_{gh}(t | Z) = \lambda_{gh,0}(t) \exp\{\beta_{gh}^T Z\} \quad (6.2)$$

where $\lambda_{gh,0}(t)$ represents the baseline hazard at time t for the transition from g to h and may be different for each transition. Z represents a vector of covariates and β_{gh} represents transition-specific regression coefficients for each covariate. This formulation allows each covariate to have a different effect on each transition hazard. This transition hazard model is analogous to a cause-specific hazards model in the competing risks setting.

In this descriptive analysis, the focus is on presenting average results as opposed to adjustment for many confounders. Therefore, I adjust for only four key covariates: gender (M/F), F508del homozygous (Y/N), FEV1% (continuous: centred at 50) and BMI (categorical: underweight (<18.5), normal weight (18.5-24.9), overweight/obese (≥ 25.0)). BMI categories are based on World Health Organisation category definitions [WHO Regional Office for Europe, 2021]. Additionally, an interaction between genotype and gender was included for the transition to CFRD. To test whether the effect of each covariate could be assumed equal across transitions, likelihood ratio tests were used comparing a model with the selected covariate effect assumed equal across transitions to a model where it was allowed to vary by transition. This was done for each covariate, one-by-one, with the result that a better fit was obtained by allowing transition-specific covariate effects. The model in equation 6.2 is fit to a dataset in stacked, long format where there is one row per individual per possible transition [Geskus, 2016]. Separate Cox regressions can be fitted to each transition or a single model may be fitted with the baseline

hazard stratified by transition (Ibid.).

The Clock-forward model is compared to a second clock-forward model where follow-up time in state 2 (CFRD) was split into one-year increments and the duration of CFRD was added as a time-varying covariate. In this model, the Clock-forward + duration model, age is again used as the primary timescale but a second timescale was added, CFRD duration, using a time-varying covariate. This is a relaxation of the Markov assumption. Unfortunately, the use of a time-varying covariate means that prediction of transition probabilities is difficult. The Clock-forward model was also compared to a Clock-reset model, which differs fundamentally from the clock-forward models in that time since entering each state is the timescale not age. Age at first observation divided by 10 is adjusted for using both linear and quadratic terms. This semi-Markov model allows the transition from the CFRD state to death or transplant to depend on time spent in the CFRD state. As above, the proportional hazards assumption was checked by examining plots of scaled Schoenfeld residuals, a test that the parameter for a term for each coefficient times times is zero, and a global test [Grambsch and Therneau, 1994].

In addition to exploring the transition hazards, transition probabilities are estimated. The transition probability is the probability of transitioning from state g to state h by time t given that the person was in state g at time u and given their event history \mathcal{H}_u up to time u and covariate values Z [Geskus, 2016]. In general, for some future event at time t , \mathcal{E}_t , we wish to estimate $\Pr(\mathcal{E}_t \mid \mathcal{H}_u, Z)$ [Putter et al., 2007]. Following the notation of Putter et al. [2007], in our setting with age as the timescale, the probability of an individual who is first seen in the registry without CFRD (state 1) at age 20 remaining in state 1 at age 30 conditional on their history at age 20 and covariate values is:

$$\Pr_{11}(u = 20, t = 30) = \Pr(R > 30, T > 30 \mid \mathcal{H}_{20}, Z) \quad (6.3)$$

where R is the time of the intermediate event (diagnosis of CFRD) and T is the time of death or transplant. Similarly, the individual's transition probability from state 1 at age 20 to state 3 at age 30 without having been diagnosed with CFRD is written:

$$\Pr_{13}^1(u = 20, t = 30) = \Pr(T \leq 30, T < R \mid \mathcal{H}_{20}, Z) \quad (6.4)$$

Here, the superscript 1 in \Pr_{13}^1 refers to the fact that the individual went directly to state 3 from state 1 without going via state 2. This transition may have been made at any time after time u up to and including time t .

$$\Pr_{13}^2(u = 20, t = 30) = \Pr(R \leq T < 30 \mid \mathcal{H}_{20}, Z) \quad (6.5)$$

is the corresponding transition probability from state 1 at age 20 to state 3 at age 30 where the individual also passed through state 2, diagnosis of CFRD. For an

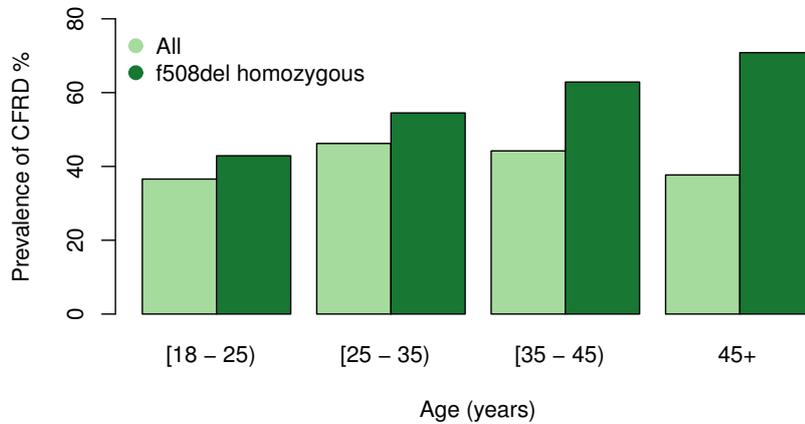


Figure 6.3: Prevalence of CFRD amongst a cross-section of the study population in 2017. Light-green represents the prevalence amongst all people in the study population and dark-green shows the prevalence for the 46% who have two copies of the F508del CFTR mutation.

irreversible illness-death model and an individual beginning in state 1, $\Pr_{11}(u, t) + \Pr_{12}(u, t) + \Pr_{13}^1(u, t) + \Pr_{13}^2(u, t) = 1$. There are 4 possible transition paths so the sum of their transition probabilities must equal 1. To estimate these transition probabilities, first the cumulative transition hazards specific to a set of covariate values are estimated for each transition and then an Aalen-Johansen-type estimator is used [de Wreede et al., 2010, 2011].

All analyses were performed using R v4.0.2 [R Core Team, 2020]. Poisson regression was fit using the `glm` function in the `stats` package and the negative binomial model using the `glm.nb` function in the `MASS` package [Venables and Ripley, 2002]. The Cox proportional hazards models were fitted with the R package `survival` [Therneau and Grambsch, 2000, Therneau, 2015]. The multi-state model predictions and data preparation in stacked, long format were performed using the `mstate` package [Putter et al., 2007, de Wreede et al., 2010, 2011].

6.4.3 Results

Prevalence of CFRD

Using a cross-sectional sample of 5,005 adults with CF from the UK CF Registry in 2017, the prevalence of CFRD by age group for all adults and for those who have 2 copies of the F508del mutation is shown in Figure 6.3. The prevalence of

CFRD in this group was 36% for 18-25 year olds and was highest (46%) for 25-35 year olds. Looking only at those individuals who are F508del homozygous, the prevalence of CFRD is larger in all adult age groups, reaching 71% for those over age 45.

Characteristics of the incident cohort

Table 6.2 provides descriptive statistics at the start of follow-up for each of the 4,512 individuals in the incident cohort. The median age at the start of follow-up was 22 years and 2,009 (45%) of the individuals were female. 3,224 (71%) individuals were pancreatic insufficient and 1,949 (43%) were homozygous for the F508del mutation. Because this is an incident cohort, no individuals had been diagnosed with CFRD at the beginning of their follow-up. By the end of follow-up, 1,032 (23%) were diagnosed with CFRD.

Mortality rate analysis

The crude mortality/transplant rate in our 10-year cohort, estimated as the total number of events divided by the total person-years of follow up, was 2.0 per 100 person-years. The final Poisson mortality rate model included adjustment for CFRD, sex and age (with linear and quadratic terms). A negative binomial fitted with the same covariates did not improve the fit. The dispersion parameter 0.04 [95% CI:-0.04,0.11] was not significantly different from 0 so the Poisson regression was chosen. The estimated mortality/transplant rate at age 30 for a male without CFRD was 1.2% [95% CI:1.0%, 1.4%] compared to 4.6% [95% CI:4.0%, 5.4%] for males with CFRD. Thirty-year old females with CFRD have a higher estimated mortality/transplant rate than males of the same age with CFRD at 5.7% [95% CI:4.9%, 6.7%] but the confidence intervals are overlapping indicating that this gender difference may not be significant. Estimated mortality for females at age 30 without CFRD was 1.5% [95% CI:1.3%, 1.7%]. Figure 6.4 shows the estimated mortality rates by sex, age and CFRD status. 95% confidence intervals are shown as dashed lines. The higher estimated mortality associated with CFRD is evident at all adult ages and for both sexes.

Cox regression analysis

Multivariable Cox regression analysis was used to explore the association between CFRD and the composite outcome of death or transplant. Results from all 3 analyses are presented in table 6.3. The first regression used baseline data (data from the first annual review record for each individual in the study period) from the period prevalent cohort. Evidence was found that the following covariates were associated with the hazard for mortality: CFRD, sex, BMI, FEV1%, FVC%, F508del

Table 6.2: Descriptive characteristics for the incident cohort at the start of follow-up for each individual. For binary and categorical variables, the number and percent are shown. The median, 25th and 75th percentile (IQR) values are shown for continuous predictors.

<i>Variable</i>		<i>No.</i>	<i>%</i>
Pancreatic insufficiency	Yes	3224	71
	No	1288	29
Sex	Female	2009	45
	Male	2503	55
Genotype	F508del homozygous	1949	43
	Other	2563	57
Neonatal screening	Yes	817	18
	No	3695	82
<i>B. cepacia</i>	Yes	198	4
	No	4314	96
<i>S. aureus</i>	Yes	1977	44
	No	2535	56
Hospital IV days	0 days	2979	66
	1-7 days	418	9
	8-14 days	488	11
	15-21 days	138	3
	22-28 days	156	3
	>28 days	333	7
Corticosteroid use	Yes	2147	48
	No	2365	52
Liver Disease	Yes	513	11
	No	3999	89
<i>Variable</i>		<i>Median</i>	<i>IQR</i>
Age at start of follow-up (years)		21.7	(18.7,30.0)
BMI (kg/m ²)		21.8	(19.9,24.3)
FEV1%		71.2	(52.7,87.5)
FVC%		85.6	(71.3,97.1)

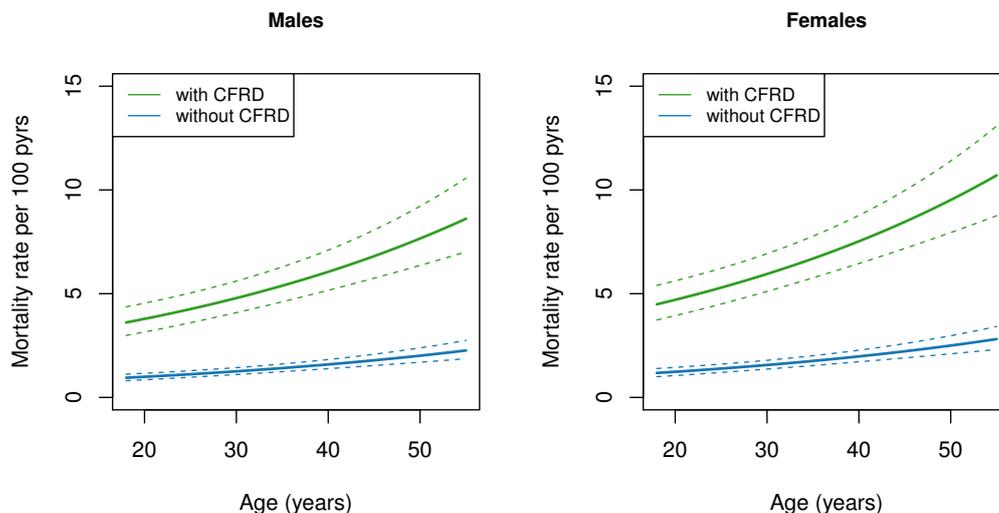


Figure 6.4: Estimated mortality/transplant rates for adults with CF by age, sex and CFRD status. 95% confidence intervals are shown as dashed lines. (pyrs = person-years)

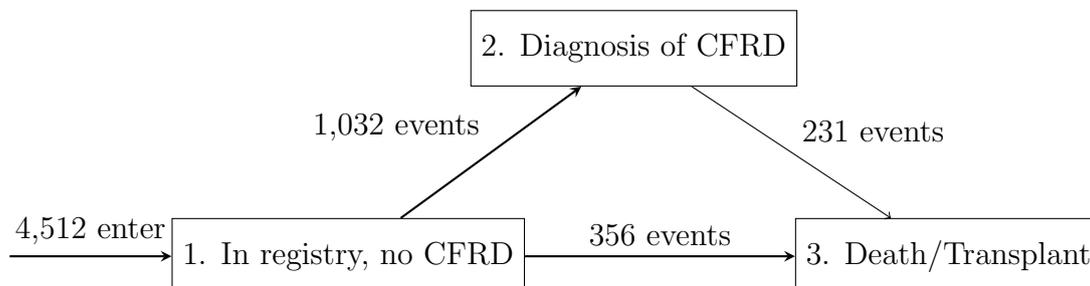
homozygous status, neonatal screening (Y/N), use of corticosteroids, liver disease, and infection with *B. cepacia* or *S. aureus*. The hazard ratio for CFRD was 1.40 [95% CI:1.24, 1.57] indicating a 40% increased hazard of mortality or transplant for someone with CFRD versus someone without CFRD, holding other variables constant. Repeating the previous analysis using longitudinal data in an extended Cox regression with the incident cohort, CFRD was again a risk factor for mortality with a hazard ratio of 1.62 [95% CI:1.36, 1.94]. Whereas genotype was identified as a risk factor in the baseline values analysis, the hazard ratio is estimated to be 1.03 [95% CI:0.86, 1.23] when using time-updated values and the confidence interval includes 1.0. In a third Cox regression analysis, the association of duration of CFRD with the hazard of mortality was investigated by incorporating the time in years since CFRD diagnosis as a time-varying covariate. This was modelled as a categorical variable with 4 categories: not diagnosed with CFRD, CFRD duration < 2.5 years, 2.5-5.0 years, and ≥ 5 years. The hazard ratio estimates for duration of CFRD increased slightly from 1.57 for CFRD duration of less than 2.5 years to 1.61 for a duration of 2.5-5 years and to 1.79 for 5-8.4 years with CFRD. The confidence intervals on the hazard ratios for higher durations completely encompass the confidence intervals on the hazard ratios for shorter durations indicating that uncertainty of the estimate increases with duration. A likelihood ratio test of the two Cox regressions with time-varying predictors indicates that the model includ-

Table 6.3: Cox regression analyses of risk factors associated with the composite outcome of death or lung transplant in adults with CF. Hazard ratio estimates (HR) along with 95% confidence intervals (CI) are shown for each covariate. “Baseline values” refers to the analysis using the period prevalent cohort where only risk factor measurements and CFRD (Y/N) from the first recorded annual review during the study period (2008-2017) were used. “Time-varying values” refers to the analyses using the incident cohort in an extended Cox model with longitudinal data on covariates and CFRD status from all annual reviews for each individual during the study period.

Risk Factor	Baseline values	Time-varying values	
	HR (95% CI)	HR (95% CI)	HR (95% CI)
CFRD	1.40 (1.24, 1.57)	1.62 (1.36, 1.94)	
CFRD (dur < 2.5yrs)			1.57 (1.24, 1.99)
CFRD (dur 2.5-5yrs)			1.61 (1.22, 2.14)
CFRD (dur 5-8.4yrs)			1.79 (1.19, 2.68)
Sex (female)	1.27 (1.13, 1.42)	1.16 (0.98, 1.37)	1.14 (0.96, 1.36)
BMI	0.94 (0.92, 0.96)	0.93 (0.90, 0.95)	0.93 (0.90, 0.95)
FEV1%	0.96 (0.96, 0.97)	0.95 (0.94, 0.96)	0.95 (0.94, 0.96)
FVC%	0.98 (0.98, 0.99)	0.98 (0.97, 0.99)	0.98 (0.97, 0.99)
F508del homozygous	1.24 (1.10, 1.40)	1.03 (0.86, 1.23)	1.03 (0.87, 1.24)
Neonatal screening	1.18 (1.03, 1.36)	0.93 (0.75, 1.15)	0.93 (0.75, 1.15)
<i>B. cepacia</i>	1.49 (1.21, 1.83)	1.65 (1.28, 2.12)	1.67 (1.30, 2.15)
<i>S. aureus</i>	0.80 (0.71, 0.90)	0.72 (0.61, 0.87)	0.73 (0.61, 0.87)
Hospital IV Days	1.16 (1.13, 1.20)	1.32 (1.26, 1.38)	1.32 (1.26, 1.38)
Corticosteroid use	1.05 (0.93, 1.18)	1.06 (0.88, 1.28)	1.07 (0.89, 1.29)
Liver disease	1.24 (1.05, 1.45)	1.07 (0.86, 1.35)	1.07 (0.85, 1.35)
Global chi-square test for PH	p=0.56	p=0.70	p=0.41

ing CFRD duration does not fit the data better than the model including only the binary indicator of CFRD yes/no ($p=0.14$). The proportional hazards assumption was satisfactory in both analyses with time-updated data with no p-values less than 0.08 on individual covariate tests and no p-value less than 0.41 on the global chi-square test for proportional hazards. In the baseline values analysis, there is some evidence that the log hazard ratio for *B. cepacia* varies with time ($p=0.03$) but a plot of the scaled Schoenfeld residuals appears reasonable. This result may be due to the small number of individuals infected with this bacteria.

Figure 6.5: Depiction of the multi-state model for the CFRD setting using the incident cohort. 4,512 people were first seen in the registry without CFRD and 1,032 of them subsequently developed CFRD. 231 people died or had a transplant after a diagnosis of CFRD and 356 died or had a transplant without CFRD.



Multi-state modelling analysis

Figure 6.5 illustrates the three-state illness-death model adapted for our CFRD setting. The 4,512 individuals in the incident cohort all began in state 1 without a diagnosis of CFRD. 1,032 people developed CFRD and 231 of those subsequently died or had a lung transplant. An additional 356 people without CFRD also progressed to the absorbing state of death or transplant. Table 6.4 shows the estimated hazard ratios for each covariate in our model of the transition hazards between states. Results are presented for the three models: clock-forward, clock-forward + duration, and clock-reset. Note that the clock-forward and clock-forward + duration models produce the same hazard ratio estimates for the transition from state 1 to 2 and state 1 to 3 as the only difference is the addition of the CFRD time-varying covariate in the model for the transition from state 2 to 3. In the clock-forward + duration model, the hazard ratio estimate for CFRD duration for the transition from state 2 (CFRD) to state 3 (death/transplant) is 0.90 [95% CI:0.86,0.94] suggesting that a longer length of time with CFRD is associated with a lower risk of death or transplant. This counter-intuitive result is likely explained by the fact that healthier people (i.e. people with higher FEV1% and normal weight) survive longer with CFRD; adjusting for duration is masking the effect of the two covariates BMI and FEV1% and giving the improbable result that a longer duration of CFRD is protective for progression to death or transplant. It is also challenging to interpret estimates and predict transition probabilities with time-updated covariates and so the clock-forward + duration model is not considered further.

Comparing the clock-forward and clock-reset models, there is little difference in the hazard ratios for the covariates. In the clock-reset model, the hazard ratio estimates for the age terms in the two transitions to death or transplant indicate a complex, at least quadratic relationship between age and risk of the composite

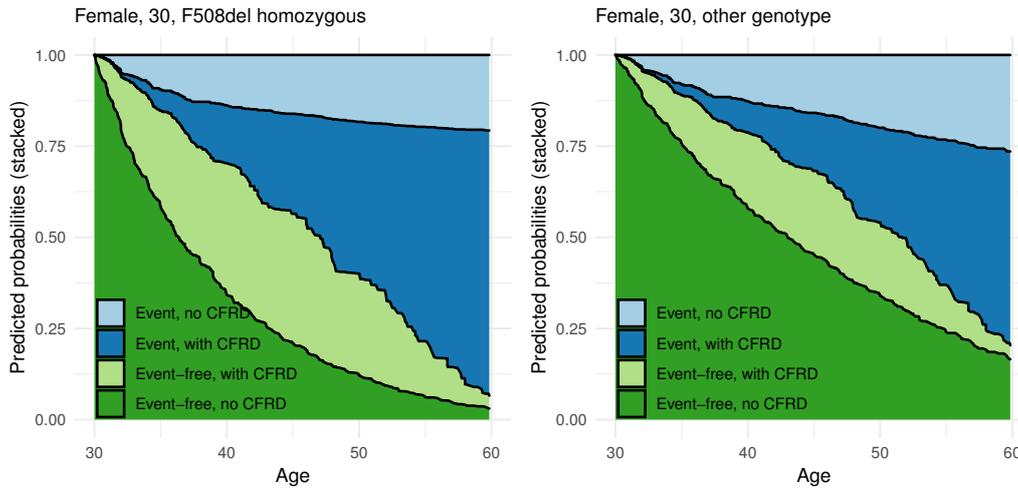


Figure 6.6: Predicted transition probabilities for two hypothetical 30-year old female patients with normal BMI and FEV1% equal to 50, one with F508del homozygous genotype and one with other genotype. The transition probabilities are stacked so that the height of each shaded area represents the transition probability from state 1 (in registry, no CFRD) at time 0 (age 30). An event is defined as either death or transplant.

outcome. Because the results of the clock-reset model are similar to the clock-forward model and the clock-forward model offers the opportunity to use the more natural timescale of age, the remainder of the analysis focuses on the clock-forward model. In that model, for the transition from state 1 (no CFRD) to state 2 (diagnosed with CFRD), a male who is F508del homozygous has a greater risk of being diagnosed with CFRD than a male with a different genotype, all other covariates held constant. The risk increases even more for a female who is F508del homozygous. Genotype continues to play a role in the risk of transition from state 1 (no CFRD) to state 3 (death/transplant) but was not significantly associated with the transition from state 2 to state 3. Higher FEV1% was associated with a lower risk for all transitions and being underweight was associated with a greater risk for all transitions.

Figure 6.6 shows predicted transition probabilities for two 30-year old females with normal BMI and FEV1%=50, starting from state 1 without a diagnosis of CFRD and either remaining in state 1 or transitioning to one of the other three possible states. One has a severe genotype, F508del homozygous (left), and one, labelled “other genotype”, is not F508del homozygous (right). In this graphic, the predicted transition probabilities are stacked so the height of each shaded region represents the predicted probability, $P_{1h}(30, t)$ for the transition from state

Table 6.4: Hazard ratio (HR) estimates with 95% confidence intervals (CI) for each covariate in the three semi-parametric multi-state models. The estimates from the clock-forward + duration model are the same as the estimates from the clock-forward model for the transitions from state 1 to 2 and state 1 to 3 and, for readability, are not repeated.

Covariate		Clock-Fwd HR (95% CI)	Clock-Fwd + Dur HR (95% CI)	Clock-Reset HR (95% CI)
<u>Transition 1 → 2: No CFRD → CFRD</u>				
Gender	Male			
	Female	1.06 (0.87,1.29)		1.06 (0.87,1.29)
Genotype	other			
	F508del homozygous	1.62 (1.36,1.94)		1.61 (1.35,1.92)
Interaction				
	Female × F508del homozygous	1.37 (1.06,1.76)		1.37 (1.06,1.76)
FEV1%		0.98 (0.98,0.99)		0.98 (0.98,0.99)
BMI	Normal weight			
Category	Underweight	1.28 (1.06,1.53)		1.23 (1.03,1.48)
	Overweight/obese	0.85 (0.71,1.03)		0.86 (0.71,1.04)
Age (years/10)				0.87 (0.71,1.08)
Age ² (years ² /100)				1.03 (0.96,1.10)
<u>Transition 1 → 3: No CFRD → Death/transplant</u>				
Gender	Male			
	Female	1.28 (1.03,1.59)		1.30 (1.05,1.61)
Genotype	other			
	F508del homozygous	1.35 (1.08,1.69)		1.36 (1.09,1.70)
FEV1%		0.93 (0.93,0.94)		0.93 (0.92,0.93)
BMI	Normal weight			
Category	Underweight	1.43 (1.10,1.86)		1.35 (1.04,1.76)
	Overweight/obese	0.73 (0.49,1.09)		0.77 (0.52,1.15)
Age (years/10)				0.51 (0.36,0.73)
Age ² (years ² /100)				1.21 (1.09,1.34)
<u>Transition 2 → 3: CFRD → Death/transplant</u>				
Gender	Male			
	Female	1.44 (1.10,1.89)	1.46 (0.94,1.26)	1.36 (1.04,1.77)
Genotype	other			
	F508del homozygous	0.96 (0.73,1.26)	1.09 (0.94,1.26)	0.91 (0.69,1.19)
FEV1%		0.95 (0.94,0.96)	0.96 (0.95,0.96)	0.95 (0.94,0.96)
BMI	Normal weight			
Category	Underweight	1.41 (1.02,1.95)	1.18 (0.98,1.41)	1.23 (0.89,1.71)
	Overweight/obese	0.42 (0.23,0.76)	0.54 (0.41,0.72)	0.42 (0.24,0.75)
CFRD Duration			0.90 (0.86,0.94)	
Age (years/10)				0.58 (0.37,0.90)
Age ² (years ² /100)				1.28 (1.12,1.46)

1 to state h . The sum of the light and dark blue shaded regions represents the predicted probability of a transition to death or transplant (“Event”) irrespective of CFRD diagnosis. At all ages, the transition to death or transplant is more likely for the F508del homozygous individual. For the F508del homozygous (other) genotype, after age 38 (45), there is a greater predicted probability of an event after a positive CFRD diagnosis than without CFRD. The burden of CFRD is represented by the combined area of the dark blue and light green regions and the predicted probability of a transition to or through the CFRD state reaches 70% by age 50 for the F508del homozygous individual. Although the predicted transition probabilities involving the CFRD state are lower for the individual of other genotype, the predicted transition probabilities that include state 2 (CFRD) exceed 50% after age 53.

In figure 6.7, the transition probability to death or transplant with CFRD is plotted in greens and the transition probability to death or transplant without CFRD is plotted in blues for three females (left) and three males (right) who are in the healthy state at age 20, 30 and 40. These individuals are F508del homozygous with normal weight and FEV1% of 50. For the transition to death or transplant without CFRD, the probability of transition initially rises quickly for both sexes at all three ages and then the curves begin to flatten out as they approach 20%. In contrast, the predicted transition probabilities to death or transplant with CFRD rise steadily a few years after the individual is seen in the registry without CFRD. In this case, individuals must first transition to the CFRD state and then to the event state.

Finally, the predicted transition probability to death or transplant for an individual who begins in state 1 (no CFRD) is compared to that of an otherwise equivalent individual who begins in state 2 (CFRD). Figure 6.8 shows these predicted probabilities for hypothetical 30-year old females who are F508del homozygous (top) and for 30-year old females of other genotype (bottom). Orange lines plot the transition probability to death or transplant given that the individual already has a diagnosis of CFRD at age 30, i.e. given that they begin in state 2. This is $P_{23}(u = 30, t)$. The corresponding blue lines show transition probabilities given that the individual has not been diagnosed with CFRD at age 30. Their probability of transition to death or transplant is the sum of the probability of transition directly from state 1 to state 3, $P_{13}^1(u = 30, t)$, plus the probability of transition to state 3 via state 2, $P_{13}^2(u = 30, t)$. For both genotype groups, for a person of intermediate health (normal weight, FEV1%=50) or better health (normal weight, FEV1%=80), the predicted probability of transition to an event from the CFRD state exceeds the predicted probability of death from the non-CFRD state until at least age 60. The hypothetical individuals with worse health (underweight, FEV1%=30) show a similar pattern but before the age of 60, the probability of

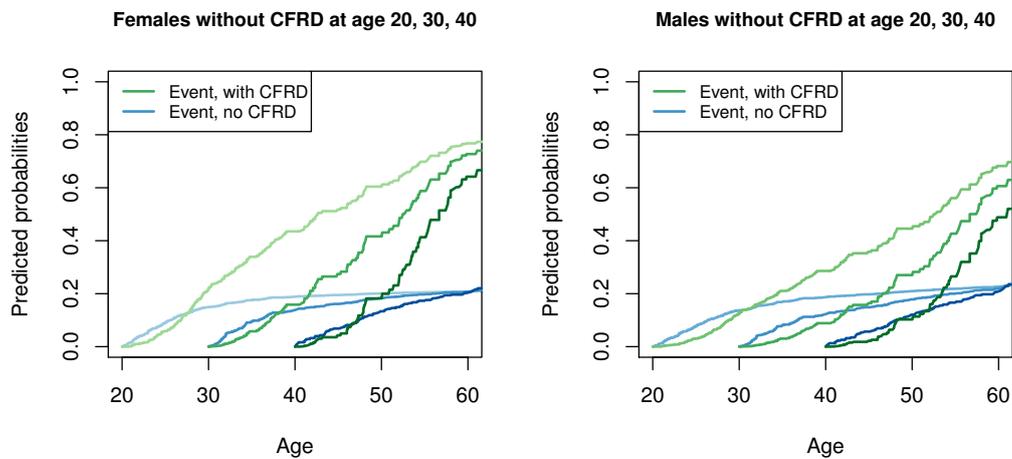


Figure 6.7: Predicted probability of transition to death or transplant (Event) for three hypothetical female patients (left) and three hypothetical male patients (right) who begin without CFRD. The individuals are aged 20, 30 and 40 years, have an F508del homozygous genotype, normal weight and FEV1% of 50. Green lines show the predicted transition probability up to age 60 of both being diagnosed with CFRD and having an event. Blue lines show the analogous predicted transition probability of having an event without being diagnosed with CFRD.

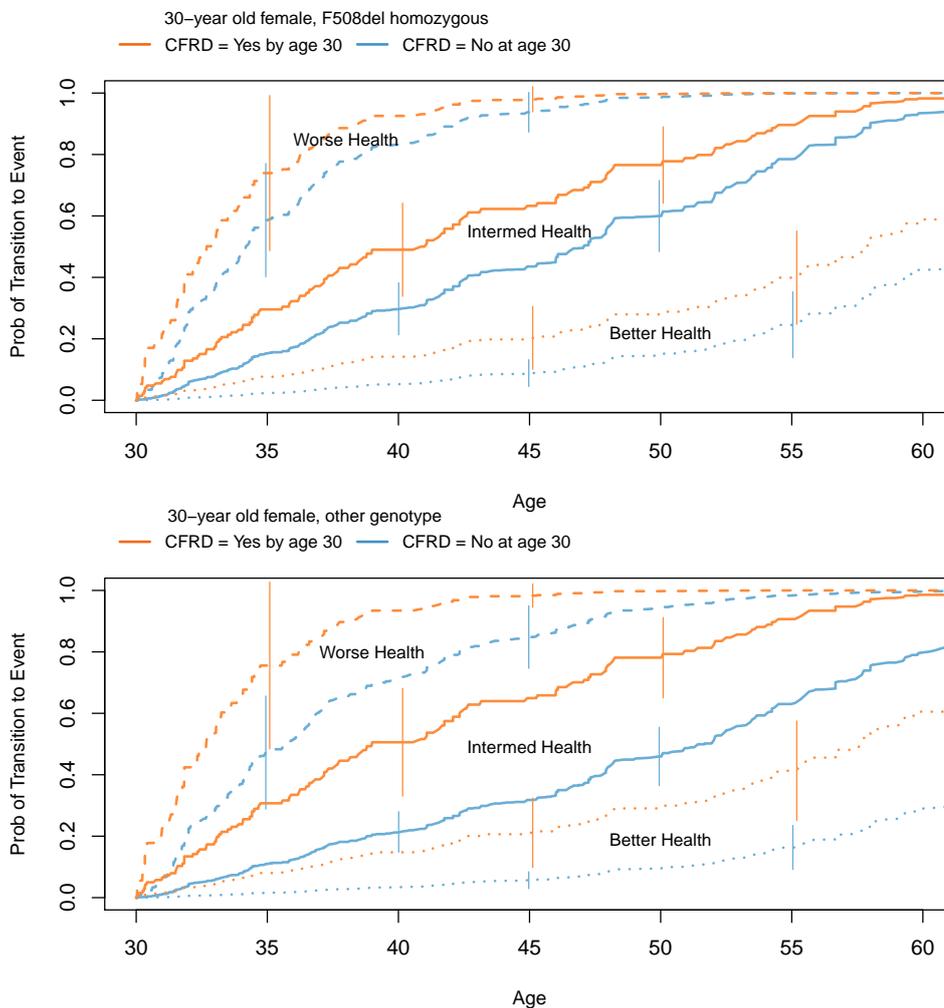


Figure 6.8: Predicted transition probabilities are drawn in blue for the transition from no CFRD to death or transplant and in orange for the transition from CFRD to death or transplant. In the top graphic, predictions are for three hypothetical 30-year old F508del homozygous female patients with intermediate health (solid lines), better health (dotted lines) and worse health (dashed lines). In the bottom graphic, the same cases are presented but for a 30-year old female of other genotype. Intermediate health is defined as normal weight and FEV1%=50, better health as normal weight and FEV1%=80 and worse health as under-weight and FEV1%=30. 95% confidence intervals are displayed as vertical error bars for selected ages.

transition to an event from either state 1 or state 2 both approach 1.0. 95% confidence intervals overlap for the F508del homozygous females with intermediate or better health but do not for the analogous individuals of other genotype indicating a significant difference in outcomes. Error bars are large for all individuals with worse health, especially at younger ages.

6.5 Discussion

In these analyses of UK CF Registry data, CFRD was associated with worse outcomes for adults with CF after adjustment for sex, genotype and baseline health. I found that adults with CF and CFRD experienced significantly higher mortality rates than adults without CFRD and that a diagnosis of CFRD was associated with a 62% increased risk of death or transplant using a Cox regression with time-varying covariates. Multi-state modelling provided further evidence of the negative relationship between CFRD and survival with predicted transition probabilities to death or transplant from the CFRD state exceeding the predicted transition probabilities to event from the no CFRD state.

Quantifying how CFRD affects survival in people with CF is deceptively difficult to answer. In addition to left censoring complications that arise from the clinical ambiguities in diagnosing CFRD, numerous statistical obstacles make a simple summary number of CFRD's effect elusive. One might wish to begin with a Kaplan-Meier plot comparing those with and without CFRD but difficulties in constructing the two comparator groups immediately arise. For those with CFRD we could assemble their survival data using age as the time scale and age at diagnosis as the start time but it is not clear how to define the non-CFRD group. Simply censoring individuals at the time of their CFRD diagnosis leads to censored people having different survival probabilities than uncensored people and this violates an assumption of the Kaplan-Meier method. Nor can we select the group of individuals who never develop CFRD during the study period as our non-CFRD comparator group as this would be conditioning on the future.

Age-adjusted mortality rates provide a summary measure to contrast two groups in a relative sense but these are difficult to compare across studies as the population used to standardise the rates generally differs by study and in my review of the literature, I found that it is often not provided with the manuscript. Also, because confounder adjustment is limited in the models, we cannot preclude the explanation that CFRD is simply a marker of more severe disease. More flexibility is offered with the Cox regression and it was found that CFRD is significantly associated with mortality after adjustment for gender, genotype and a range of time-varying health status measurements. This result is interesting because use of time-varying covariate data did not attenuate the estimated hazard ratio of CFRD

as compared to the baseline analysis. If CFRD's key impact on survival is through increased respiratory infections, poor lung function and worse nutritional status, as surmised in section 6.2, then by adjusting for these covariates using time-updated values, we would expect the estimated hazard ratio of CFRD to be close to 1.0 as it's effect on survival would be captured through other downstream measures. This is thought to be why the hazard ratio estimates of genotype are not significantly different from 1.0 in our models with time-varying data but it was significant in the baseline analysis. (See table 6.3) Genotype affects the health of the individuals and the time-varying health indicator measurements capture that effect. The same is not true for CFRD which suggests that perhaps another mechanism is at work that has not been measured / included in the analysis. (Please see chapter 7 for a further investigation of the relationship between CFRD, lung function, nutritional status, respiratory infections and mortality).

The proportional hazards assumption in the Cox regression analysis may also be questioned. By modelling CFRD as a time-varying covariate within an extended Cox analysis, the restrictive assumption is made that CFRD has a multiplicative effect on the baseline hazard. Although tests of the proportional hazards assumption conducted during the Cox regression analysis revealed no violations, in the multi-state modelling analysis, I concluded that proportional hazards could not be assumed between the transition from state 1 (no CFRD) to state 3 (event) and from state 2 (CFRD) to state 3 (event). If an assumption of proportionality could be justified, this would have facilitated the estimate of a transition hazard ratio for CFRD. The multi-state model allowed for the investigation of CFRD not just as a covariate that may change value, but as an event that may be modelled as part of an individual's history. This rich analysis allows us to compare the prognosis for two individuals with different event histories and/or covariate values but cannot provide a one-number summary of the effect of CFRD on survival. There are also trade-offs between including more covariates and the resulting need for more covariate-specific predictions as well as challenges in interpretation as time-updated covariates are included.

Because CFRD is associated with severe genotypes and severe genotypes are associated with decreased survival [Lewis et al., 2015], genotype was adjusted for in all analyses using a binary covariate for F508del homozygous status in order to maximise our sample size. The F508del mutation is the most common mutation in the UK with 89.7% of people having at least one copy and 48.6% having two copies (homozygous) [UK Cystic Fibrosis Registry, 2020]. However, the group defined as not F508del homozygous, contains people with severe genotypes characterised by other mutations. Ideally, the exact mutation on each CFTR allele for all individuals would be known and each individual's genotype could be classified but for many one or both mutations are unknown. In the incident cohort used here, 2,563 individuals

were not F508del homozygous. Of these, 849 were known to have a severe genotype (i.e. Class I, II or III mutations), 750 were known to have a non-severe genotype and the remainder were unable to be fully classified. Because of the mixed nature of the mutation types included in the not F508del homozygous group, care must be taken not to interpret findings for this group as if it were a group of homogeneous individuals with less severe genotypes. I hypothesise that the differences reported here by genotype would likely be even greater if the comparison were between all individuals with severe genotypes versus all individuals with less severe genotypes.

Multi-state models are also capable of producing dynamic survival predictions of the sort discussed in part I of this thesis. The predictions presented in the previous section began from a fixed age (i.e. a fixed u) and extended over multiple times t . Alternatively, a prediction horizon could have been chosen, (say) 5 years, and computed dynamic predictions of 5-year survival from various starting ages given covariates and an event history. van Houwelingen and Putter [2008] explored using landmarking as an alternative method to multi-state modelling for producing such dynamic predictions. Both approaches gave similar results for their application and they note that a key advantage with landmarking is its simple model and ability to easily incorporate time-varying covariates. Although multi-state models may provide insight into the biological processes, once the multi-state model is no longer a Markov model, prediction becomes difficult. Because the primary interest in this section was in contrasting the transition probabilities between those with and without CFRD, multi-state modelling was the natural choice. In part I, however, landmarking provided a way to include time-varying data on a large number of predictors and facilitated the combination of a machine learning ensemble with dynamic survival prediction.

In summary, in this chapter, I have highlighted the challenges in interpreting results from traditional analyses when the aim is to estimate the effect of an intermediate time-dependent variable on survival. In the next chapter, approaches designed for measuring effects and the mechanisms through which an exposure may affect an outcome will be applied to further illuminate the relationship between CFRD and survival or transplant.

Chapter 7

Investigation of Mediators of the Effect of CFRD on Survival

7.1 Introduction

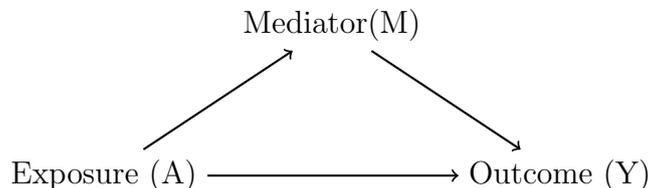
The goal of causal inference is distinct from the goals of descriptive analyses in that we aim to go beyond identifying an association between an exposure and an outcome. Rather, the aim is to make a statement about whether an exposure or treatment is a cause of the outcome. Estimating the total effect is typically the focus of investigations into causal relationships. If there is evidence of an effect of exposure on outcome, it can then be of interest to investigate the mechanisms through which this effect arises. Mediation analysis is used to investigate whether another variable(s) may be mediating the effect of an exposure on an outcome. In this chapter, I review mediation analysis using both traditional methods and a counterfactual framework. I then describe in detail two recently proposed methods for mediation analysis with a time-to-event outcome and a longitudinal mediator. The second part of the chapter is dedicated to an application of these methods to data from the UK CF Registry to investigate whether the effect of CFRD on survival is mediated through lung function, nutritional status and/or respiratory infections.

7.2 Mediation

7.2.1 Overview of traditional mediation methods

Mediation analysis helps answer the question of how much of the total effect of an exposure on an outcome is mediated through some other process. Figure 7.1 illustrates a simple mediation scenario with a single exposure A , a single mediator

Figure 7.1: In this simple mediation scenario, the exposure (A) may affect the outcome (Y) directly and/or it may affect a third mediating variable (M) which in turn affects the outcome.



M and an outcome Y . The mediator lies on a potential causal path between the exposure and the outcome and may explain all, none or part of the total effect. The aim of a mediation analysis is to quantify both the direct effect of the exposure on the outcome as well as the indirect effect of the exposure on the outcome that is relayed through the mediator.

One of the early and most widely used approaches for assessing whether mediation is present was proposed by Baron and Kenny [1986]. This ‘causal steps’ qualitative approach consists of first testing for the existence of an effect of A on Y by looking for a significant coefficient α_A in an unadjusted regression of the outcome on the exposure:

$$Y = \phi_1 + \alpha_A A + e_1 \quad (7.1)$$

Second, the existence of an association between A and M is evaluated by looking at the regression coefficient β_A in a regression of the mediator on the exposure:

$$M = \phi_2 + \beta_A A + e_2 \quad (7.2)$$

Presence of an association between M and Y when A is controlled for can be checked with the coefficient α_M in the regression:

$$Y = \phi_3 + \alpha'_A A + \alpha_M M + e_3 \quad (7.3)$$

In the above equations, the ϕ 's are intercept terms and the e 's are errors. Finally, if the regression coefficient α_A (equation 7.1) is greater in magnitude than α'_A (equation 7.3), then M is a mediator of the effect of A on Y according to the ‘causal steps’ method. Later, it was recognized that the effect of the exposure on the outcome as measured by α_A need not be significant because the direct effect and indirect effect may have opposite signs [Valeri and VanderWeele, 2013].

To quantify the effect of the exposure on the outcome via the mediator, the above regression equations may be used in either the ‘product of coefficients method’ [Alwin and Hauser, 1975] or the ‘difference method’. The product of coefficients method makes use of equations 7.3 and 7.2. The direct effect of A on

Y is simply α'_A and the indirect effect of A on Y through M is the combination of the effect of the exposure on the mediator and the effect of the mediator on the outcome, $\beta_A \times \alpha_M$. In contrast, the difference method makes use of equations 7.1 and 7.3. Again, the direct effect is given by α'_A . The indirect effect is the difference between the total effect of the exposure on the outcome, α_A , and the effect of the exposure on the outcome when the mediator is controlled for, α'_A . Intuitively, this captures how much of the total effect is explained by the mediator. MacKinnon et al. [1995] showed the equivalence of these two methods using the above regression equations for a continuous outcome when the models are fitted by ordinary least squares and maximum likelihood estimation. For binary outcomes analysed using logistic regressions and analyses with survival outcomes however, the product and difference methods may yield different estimates [VanderWeele, 2016, Fulcher et al., 2017]. Further, for binary outcomes, because odds ratios are not collapsible, direct and indirect effect estimates calculated as above may be invalid unless the outcome is rare [VanderWeele, 2016].

7.2.2 Causal Estimands for Mediation

A limitation of the method of Baron and Kenny [1986] and other similar methods that fall broadly into the category of linear structural equation models is that the estimand definitions are tied to a particular parametric model. Also, it is difficult to account for exposure-mediator interaction and non-linearities [Richiardi et al., 2013]. A second approach to mediation suggested by Robins and Greenland [1992] and Pearl [2001] is based on counterfactuals and this facilitates model-free estimand definitions and allows more flexibility in the parametric modelling assumptions. Briefly, given a setting where the exposure A can take two values, 1 (exposed) or 0 (not exposed), define $Y_i(1)$ to be the outcome of individual i if he were exposed and $Y_i(0)$ is the outcome if he were not exposed. Only one of these two potential outcomes can be observed for each individual; the other is counterfactual. Similarly, there are two potential values of the mediator M , $M_i(1)$ and $M_i(0)$, the value of the mediator if individual i was exposed and not exposed, respectively. Let $Y_i(a, M_i(a^*))$ be the potential outcome of individual i if we set the exposure $A = a$ and the corresponding mediator value $M_i(A = a^*)$ to the level it would take if A were set to a^* where a may or may not equal a^* . Using this notation in our binary exposure setting, the natural direct effect (NDE) and natural indirect effect (NIE) are defined as:

$$NDE = E[Y(1, M(0)) - Y(0, M(0))] \quad (7.4)$$

$$NIE = E[Y(1, M(1)) - Y(1, M(0))] \quad (7.5)$$

[VanderWeele and Vansteelandt, 2009]. The NDE captures the effect of the exposure on the outcome if the mediator remained unchanged at the level it would have taken without being exposed. The NIE captures the change in outcome that would result from changing the mediator value from the level it would take if unexposed to the level it would take if exposed while holding the exposure status constant. The sum of the NDE and NIE is the total causal effect.

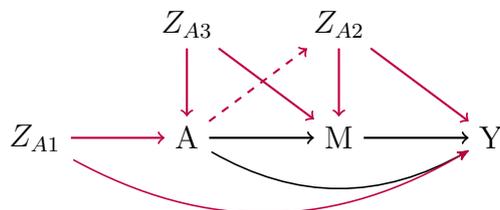
Identification of the NIE and NDE requires a set of strong assumptions:

- A1. no unmeasured confounding of the exposure-outcome relationship
- A2. no unmeasured confounding of the mediator-outcome relationship
- A3. no unmeasured confounding of the exposure-mediator relationship
- A4. no confounder of the mediator-outcome relationship that is affected by the exposure

Assumptions A1 - A3 can be visualised by modifying figure 7.1 with the addition of three confounders Z_{A1} , Z_{A2} , Z_{A3} as in figure 7.2 [Valeri and VanderWeele, 2013, VanderWeele and Vansteelandt, 2013]. Each of these must be measured in order to control for confounding. If the exposure is randomised, assumptions A1 and A3 would be satisfied [Valeri and VanderWeele, 2013]. If not, the analysis must be controlled for common causes of the exposure and outcome (as would be required for estimation of the total effect) and common causes of the exposure and mediator. Because it is unlikely, even in a randomised trial, that the mediator will be randomised, assumption A2 requires measurement of and control for all common causes of the mediator and the outcome for an unbiased estimate of the direct effect. Assumption A4 is equivalent to there being no arrow from A to Z_{A2} in figure 7.2 and requires that Z_{A2} is not itself a mediator of the effect of the exposure on the outcome. Assumption A4 is problematic because measuring such a confounder does not avoid violation of this assumption; rather, A4 says that such a confounder cannot exist. For cases when the time between exposure and mediator measurement is very short, this assumption may be reasonable [VanderWeele, 2016].

VanderWeele and Vansteelandt [2009] show how these definitions of direct and indirect effects coincide with the method of Baron and Kenny [1986] when there is no exposure-mediator interaction and, further, how their regression models can be extended for use even in the presence of exposure-mediator interaction. Putting a causal interpretation on the estimates of direct and indirect effects estimated using the traditional method with regression equations 7.1-7.3 requires the same four assumptions as in the causal framework.

Figure 7.2: The single mediator model with the addition of three confounders. Z_{A1} is an exposure-outcome confounder, Z_{A2} is a mediator-outcome confounder and Z_{A3} is an exposure-mediator confounder. If Z_{A2} is affected by the exposure, then it is also a mediator of the effect of A on Y .



The concepts of NDE and NIE require strong assumptions; the presence of a second mediator, for example, violates the assumptions necessary for identification. One idea for identification of direct and indirect effects under less restrictive assumptions is interventional direct and indirect effects [VanderWeele et al., 2014]. Instead of setting the mediator to the pre-defined counterfactual level, the interventional effects approach involves setting the mediator to a random draw from the distribution of mediator values at a given exposure level, conditional on observed covariates. Vansteelandt and Daniel [2017] extended this idea to the case of multiple mediators and, by decomposing the total effect into path-specific effects together with a ‘mediated dependence’ term, they overcame the problem of interventional effects not summing to the total effect.

7.2.3 Mediation analysis for time-to-event outcomes

Mediation analysis with survival outcomes requires additional considerations. Using the traditional framework for time-to-event outcomes, the Cox regression model may not be used to directly compare the hazard ratio of the exposure when the mediator is included to the hazard ratio when it is not included due to non-collapsibility of the hazard ratio [Martinussen and Vansteelandt, 2013]. This may be alleviated for rare outcomes [VanderWeele, 2011] or, alternatively, an additive hazards model [Lange and Hansen, 2011] or accelerated failure time (AFT) model [VanderWeele, 2011] may be used. For example, assuming survival times T follow a Weibull distribution with scale parameter ν , we can rewrite equation 7.1 for the association of the exposure A with the outcome T :

$$\log(T) = \phi_4 + \alpha_A A + \nu e_4 \quad (7.6)$$

Equation 7.3 may be modified for the association between A and T when M is controlled for as:

$$\log(T) = \phi_5 + \alpha'_A A + \alpha_M M + \nu e_5 \quad (7.7)$$

where e_4 and e_5 are random variables from an extreme value distribution. Equation 7.2 for the mediator remains unchanged. Given this AFT model, VanderWeele [2011] has shown that the estimate of the indirect effect from the product method, $\beta_A \alpha_M$, equals the estimate from the difference method, $\alpha_A - \alpha'_A$.

The above method may be applied for a time-fixed mediator but with the motivating application considered here, CFRD, we must consider not only a time-to-event outcome, but also a repeatedly measured mediator. Specifically, there is a joint longitudinal process (the mediator) and a survival process. A key implication of this is that after an individual has an event, the mediator can no longer be measured; in other words, survival is also a post-treatment confounder [Didelez, 2018]. The exposure is able to affect the mediator directly but also indirectly via its effect on the survival time [Zheng and van der Laan, 2012]. This also complicates the notion of intervention on a mediator. Consider the case of an individual who would survive to time t_1 if they were exposed ($A=1$), but would have the event prior to t_1 if they were unexposed ($A=0$). The value of the mediator at time t_1 in the unexposed scenario is undefined and, therefore, such an intervention may lead to ill-defined natural effects [Didelez, 2018].

The repeatedly-measured mediator further complicates the analysis as measurements of the mediator and other confounders made prior to time t likely confound the association between the outcome and the mediator measured at time t [Vansteelandt et al., 2019]. Confining ourselves to a summary measure of the mediator or the most recently measured mediator is not a satisfactory solution as interest lies in the overall mediation effect which may change over time dynamically. In fact, estimating the indirect effect using only the last mediator measurement may underestimate the effect via the mediator as it only reflects one part of the process that has developed over time (Ibid.).

Two recent approaches in the literature provide techniques for mediation analysis in this setting. Aalen et al. [2020] propose a method based on treatment separation that extends the approach of Didelez [2018] to a time-to-event outcome. Conceptually, instead of imagining intervening on the mediator, the exposure is split into two variables: one that accounts for the effect of the exposure directly on the survival time and one that acts on the outcome via the mediator. A second approach, which can accommodate time-varying confounders and is based on path-specific effects, is presented in Vansteelandt et al. [2019]. The latter two are described and applied to data from the UK CF Registry in the next sections.

7.3 The Method of Aalen et al. (2020)

7.3.1 Overview

Aalen et al. [2020] propose a method (“method of Aalen”) for our setting of a single exposure, a survival outcome and a repeatedly-measured mediator that avoids two key obstacles: an undefined intervention on a mediator due to longer survival in one counterfactual scenario than the other and survival acting as a post-treatment confounder. A key idea is treatment (exposure) separation [Robins and Richardson, 2011, Didelez, 2018]. The assumption is made that the exposure can be decomposed into two parts: one that affects survival directly and one that affects the mediator process. This is a biological assumption that the treatment or exposure has separate physiological mechanisms that could in principle be manipulated separately. Aalen et al. [2020] provide the example of a blood pressure treatment that is hypothesised to have two separate components: one affecting diastolic blood pressure and the other affecting the outcome, kidney failure, through other pathways.

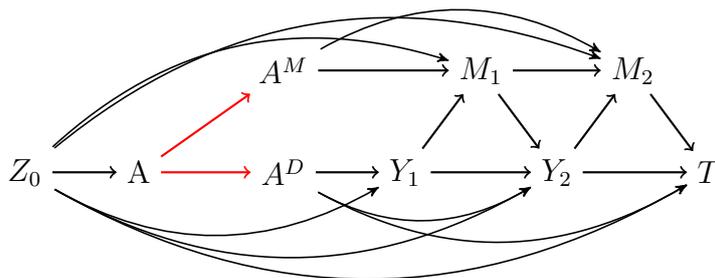
7.3.2 Setting and Estimands

Following the notation of Aalen et al. [2020], let A^M and A^D be variables representing the separate components of the exposure A . Whereas A^M affects the survival time via the mediator, A^D affects survival directly via pathways not through the mediator. For discrete mediator process $M(t)$, let $M(t) = M_k$ for $t_k \leq t < t_{k+1}$ where k indexes the visit at which the mediator is measured and $M(0) = M_0$ is the baseline mediator measurement prior to exposure. There is a set of baseline covariates Z_0 that includes M_0 . Denote the history of the mediator values for times $t \leq t_k$ as \bar{M}_k . For event time T , define $Y_t = I(T > t)$, an indicator of survival time being greater than t . The hypothesised relationships between exposure, mediator and outcome can be visualised in the DAG of Figure 7.3. In this figure, the separation of the exposure into two components is highlighted in red.

Because an individual is either exposed or not exposed, we can only observe the case where $A = A^D = A^M$. However, the separation of A into two parts allows for a hypothetical intervention where $A^D \neq A^M$. The intervention is on A^M not on the mediator and, therefore, the problem of an undefined mediator due to different survival times in the two counterfactual scenarios does not arise [Didelez, 2018]. Using these variables, three contrasts are defined to measure the effects of interest:

1. Total effect = contrast between $A^D = A^M = 1$ and $A^D = A^M = 0$
2. Direct effect = contrast between $A^D = 1, A^M = 0$ and $A^D = 0, A^M = 0$

Figure 7.3: Data-generating mechanism hypothesised in the method of Aalen that includes 2 post-exposure visits at times $t = 1, 2$. The effect of the exposure A is split into one component, A^M , that affects the mediator and one, A^D , that directly affects survival (highlighted in red). Y_1, Y_2 indicate survival to time 1, 2, respectively and T represents the survival outcome past time 2. M_1 and M_2 are the mediator measurements at times 1 and 2, respectively. Z_0 is the set of baseline covariates and includes the baseline mediator measurement.



3. Indirect effect = contrast between $A^D = 1, A^M = 1$ and $A^D = 1, A^M = 0$

where a value of 1 indicates exposed and a value of 0 indicates not exposed.

To estimate these contrasts, a continuous-time mediational g-formula is derived based on the discrete mediator process. The mediator is assumed to follow a linear model. Let $j = 1, \dots, J$ index the ordered event times t_j for the n individuals where $J \leq n$. Then, M_{ij} represents the mediator measurement for the i^{th} individual at the j^{th} event time and can be written:

$$M_{ij} = a_j A_i + \delta_j^\top Z_{0,i} + \sum_{k:t_k < t_j} b_{jk} M_{ik} + e_{ij} \quad (7.8)$$

where a_j, δ_j and b_{jk} are regression coefficients. M_{ij} is undefined for those individuals no longer at risk at time t_j .

Aalen et al. [2020] show how a corresponding model, which is marginal over the past history of M can be derived. It is this model that will be used in the mediational g-formula:

$$M_{ij} = m_{0,j} + \beta_{A,j} A_i + \beta_{Z_0,j} Z_{0,i} + \eta_{ij} \quad (7.9)$$

where $m_{0,j}$ is an intercept term, η_j is an independent error term, and $\beta_{A,j}$ and $\beta_{Z_0,j}$ are the coefficients for the effect of the exposure and baseline confounders on the mediator at time t_j , respectively.

Because the mediator model above is defined in discrete-time and the survival model in continuous-time, it is useful to define $r(t) = k$ for $t_k \leq t < t_{k+1}$. With

$\overline{M}_{r(t)}$ representing the history of the mediator values for times $t \leq t_{r(t)}$, the survival process is modelled with an additive hazards model with hazard of the form:

$$\lambda_i(t \mid \overline{M}_{i,r(t)}, A_i, Z_{0,i}) = \alpha_{0,t} + \alpha_{A,t}A_i + \alpha_{M,t}M_{i,r(t)} + \alpha_{Z_0,t}Z_{0,i} \quad (7.10)$$

where $\alpha_{0,t}$ is an intercept term and $\alpha_{A,t}$, $\alpha_{M,t}$ and $\alpha_{Z_0,t}$ are coefficients for the effect of the exposure, mediator and baseline confounder on the hazard at time t , respectively. Although the hazard at time t is conditional on the history of the mediator up to time t , it is assumed that only the most recent value of the mediator is necessary to model the hazard. In Figure 7.3, this assumption is represented by the absence of a line from M_1 to T .

Using equations 7.9 and 7.10, the mediational g-formula may be written:

$$\begin{aligned} Q(t; A^D = a, A^M = a^*, Z_0) &:= \Pr\{T(A^D = a, A^M = a^*) > t \mid Z_0\} \\ &= f(t, Z_0) \exp \left\{ -a \int_0^t \alpha_{A,u} du - a^* \int_0^t \alpha_{M,u} \beta_{A,r(u)} du \right\} \end{aligned} \quad (7.11)$$

where $T(A^D = a, A^M = a^*)$ is the survival time under an intervention where A^D was set to level a and A^M was set to level a^* . $f(t, Z_0)$ is a function capturing the covariate effects.

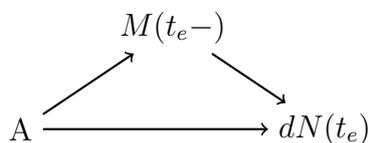
The special form of equation 7.11, made possible by the assumptions made in the mediator and hazard models, allows us to write the survival direct and indirect effects simply. The survival indirect effect (*SIE*) and survival direct effect (*SDE*) are defined using the contrasts above and the probabilities, $Q(t; A^D, A^M, Z_0)$:

$$\begin{aligned} SDE(t) &= Q(t; A^D = 1, A^M = 0, Z_0) / Q(t; A^D = 0, A^M = 0, Z_0) \\ &= \exp \left\{ (0 - 1) \int_0^t \alpha_{A,u} du \right\} \end{aligned} \quad (7.12)$$

$$\begin{aligned} SIE(t) &= Q(t; A^D = 1, A^M = 1, Z_0) / Q(t; A^D = 1, A^M = 0, Z_0) \\ &= \exp \left\{ (0 - 1) \int_0^t \alpha_{M,u} \beta_{A,r(u)} du \right\} \end{aligned} \quad (7.13)$$

Because the $f(t, Z_0)$ cancels, these are marginal effects with respect to the set of baseline covariates. The similarity of the above formulas with the product approach of Baron and Kenny [1986] is apparent. In that approach, the direct effect was given by α'_A and the indirect effect by $\beta_A \alpha_M$ from equations 7.2 and 7.3.

Figure 7.4: Simple graphical model for dynamic path analysis. The relationship between the exposure A at baseline, the change in the event time counting process $dN(t_e)$ at event time t_e and the mediator just prior to the even time $M(t_e)$ are depicted.



7.3.3 Estimation

Estimation is via dynamic path analysis suggested by Fosen et al. [2006] and applied to the setting of a survival outcome and repeatedly measured mediator in Strohmaier et al. [2015]. Dynamic path analysis is an estimation method based on a set of graphical models that define the relationships between variables over time. A counting process models the event time and the mediator is defined as a time-varying process. At every event time, t_e , there is a graphical model specific to that t_e (see Figure 7.4) relating the exposure at baseline to the mediator value just prior to the event time and to the event at t_e as noted by a jump from 0 to 1 in the counting process $dN(t_e)$. The dynamic path analysis proceeds by regressing the change in the counting process onto its parent nodes, the mediator and the exposure, and regressing the mediator onto its parent, the exposure. The additive hazards model is used for the former (equation 7.10) and linear regression for the latter (equation 7.9). This is repeated for each event time and the cumulative estimates over time are used.

As in Aalen et al. [2020], estimation can be described in 4 steps:

1. The coefficients $\alpha_{0,t}$, $\alpha_{A,t}$, $\alpha_{M,t}$ and $\alpha_{Z_0,t}$ from equation 7.10 are estimated using additive hazards regression. Refer to the cumulative estimate of $\alpha_{A,t}$ as $\alpha_{A,t}^{cum}$ and the cumulative estimate of $\alpha_{M,t}$ as $\alpha_{M,t}^{cum}$.
2. The coefficients $\beta_{A,j}$ and $\beta_{Z_0,j}$ from equation 7.9 are estimated using linear regression at each event time t_j for all those surviving up to time t_j .
3. The indirect effect of A on the hazard is then estimated by a cumulative regression function: $(a - a^*) \int_0^t \beta_{A,r(u)} d\alpha_{M,t}^{cum}$.
4. Similarly, the direct effect of A on the hazard is estimated using $(a - a^*) \alpha_{A,t}^{cum}$.
5. The SIE and SDE of equations 7.13 and 7.12 are estimated as the exponential of minus 1 times these effects on the hazard rate.

7.3.4 Assumptions

For the method of Aalen to be valid, a number of assumptions must be made. Independent censoring conditional on Z_0 is assumed in addition to the four assumptions stated in A1-A4 above. Note that confounding must be controlled solely by Z_0 , the set of baseline exogenous covariates. A linear model is assumed for the mediator with no interaction between the exposure and past values of the mediator. Further, only the most recent value of the mediator is allowed to impact the hazard. The analysis assumes an additive hazards model and that the exposure is fixed at time $t=0$. The method of Aalen relies crucially on the assumption that the exposure or treatment can be split into two separate physiological pathways with one acting on the mediator and one acting via other pathways on survival. Given survival to time t_k , the mediator history to t_{k-1} , covariates Z_0 and A^M , M_k is independent of A^D , even in the hypothetical world in which $A^D \neq A^M$. That observed survival when $A = a$ equals modelled survival when $A^M = A^D = a$ is also required. Another consequence of exposure separation is that all probabilities estimated using this approach are conditional on survival to the first mediator measurement. Individuals who had the event prior to the first mediator measurement are excluded from the analysis. See the discussion accompanying Figure 8.10 in Section 8.7 for more on this. We also assume that the parametric models are correctly specified. As with any analysis, violations of these assumptions may lead to biased estimates of the quantities of interest.

7.4 The Method of Vansteelandt et al. (2019)

7.4.1 Overview

In contrast to the method of Aalen, which has foundations in dynamic path analysis and the additive hazards model, Vansteelandt et al. [2019] propose a method (“method of Vansteelandt”) that infers the effect of an exposure on the outcome via combinations of path-specific effects. Because exposure-splitting is not used in this method, counterfactuals are defined in terms of an intervention on the mediator. As noted previously, one difficulty with mediation analyses involving time-to-event outcomes is how to set the level of the mediator for an exposed person to the level it would have been if they were not exposed if they survive longer in the scenario where they receive the exposure. Their mediator measurement is undefined after death. Whereas Aalen et al. [2020] avoided this problem by intervening on the separated exposure variable instead of the mediator, Vansteelandt et al. [2019] consider the mediator level the patient would have had if their death had been prevented.

There are two other notable differences between the methods. First, the method

of Vansteelandt overcomes the obstacle of time-varying confounding by other potential mediators, i.e. confounders of the mediator-outcome relationship that are themselves affected by treatment, by accommodating such confounders explicitly in the analysis. Also, estimation in this approach does not rely on parametric specifications.

7.4.2 Setting and Estimands

In this setting, as before, let M_k be the value of the time-varying mediator measured at visit k and Z_0 be the set of baseline confounders which may include M_0 . We introduce L_k , the value of the repeatedly-measured confounder(s) measured at visit k . The data-generating mechanism and causal ordering for this setting are visualised in Figure 7.5 for a case with two post-exposure visits, $k = 1, 2$. The assumed causal ordering is important: L_k is not influenced by M_k but L_k may influence M_k, M_{k+1}, \dots . Also, L_k may be influenced by M_{k-1}, M_{k-2}, \dots . Additionally, L_k includes an indicator taking the value 1 if the individual is at risk at visit k and 0 otherwise. The association between L and Y is allowed to be confounded by unmeasured variables, U_l , and the association between the mediator measurements may be affected by unmeasured variables U_m . U_m and U_l are permitted because the effect of the exposure along the combination of paths where the mediator is directly influenced by the exposure is computed. These are indicated in green in Figure 7.6-left. The combination of effects along all such paths is the indirect effect via the mediator. The effect along the combination of brown paths in Figure 7.6-right represents the effect of the exposure on survival that is *not* via the mediator. Note also that paths where the exposure first affects L and subsequently affects M will also be part of the direct effect not via M .

Given this setup, it is possible to estimate $S_{A(1),M(0)}(t)$, the probability of survival to time t for an individual if they were exposed and the mediator levels were set to the levels they would have been if the individual was unexposed. The values of the time-varying confounders remain unchanged at their exposed levels. Because the levels of L and M depend on each other, the level of M_k would be set at $M_k(A = 0, L_k(A = 1))$. $S_{A(1),M(1)}(t)$, the probability of survival to time t if exposed and the mediator levels are set to levels we would have seen if exposed, and $S_{A(0),M(0)}(t)$, the survival probability if unexposed and mediators are set to their unexposed levels are also estimated.

The indirect effect (IE) and direct effect (DE) are defined as contrasts between these three survival curves.

$$IE(t) = S_{A(1),M(1)}(t)/S_{A(1),M(0)}(t) \quad (7.14)$$

$$DE(t) = S_{A(1),M(0)}(t)/S_{A(0),M(0)}(t) \quad (7.15)$$

Figure 7.5: Data-generating mechanism assumed in the method of Vansteelandt for post-exposure visits, $k = 1, 2$ at times $t = 1, 2$. The assumed causal ordering is that the time-varying confounder L_k may influence M_k and M_k may influence L_{k+1} . The definition of L_k includes survival to the time of visit k . T indicates survival past visit 2 and Z_0 is the set of baseline covariates, which includes the baseline mediator measurement. U_m and U_l are unmeasured but permitted.

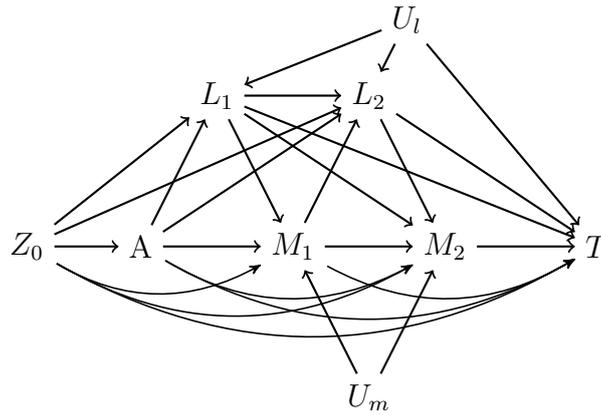
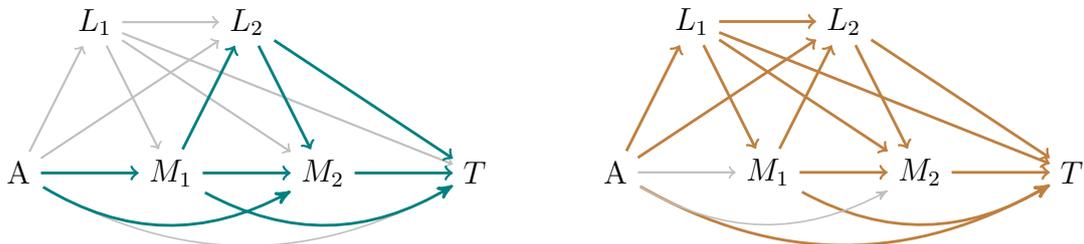


Figure 7.6: Colour-coded path-specific effects. On the left, the combination of pathways in green between A and T represent the indirect effect of the exposure on survival via the mediator. On the right, the combination of pathways in brown between A and T represent the effect of the exposure on survival *not* via the mediator. Note that these pathways may involve the mediator, but A does not first impact M . For convenience, this will be referred to as the direct effect.



The DE captures the effect of the exposure on the outcome via all paths in which the exposure does not first affect one of the mediator measurements. Like the method of Aalen, the estimand is a ratio of survival probabilities. Vansteelandt et al. [2019] describe their approach as “a generalisation of dynamic path analysis”. In fact, if all individuals survive to the first mediator measurement, there are no time-varying confounders and the mediator and event times follow additive models, the approaches are equivalent and the estimands in the method of Aalen effectively correspond to path-specific effects.

The proportion mediated may be calculated as the ratio of the indirect effect via the mediator to the total effect and can be written as:

$$\frac{S_{1,1}(t) - S_{1,0}(t)}{S_{1,1}(t) - S_{0,0}(t)} \quad (7.16)$$

Vansteelandt et al. [2019] show that $S_{A(a),M(a^*)}(t)$ is identified by

$$\begin{aligned} S_{A(a),M(a^*)}(t) &= \int f(T > t \mid T > [t], \bar{M}_{[t]}, \bar{L}_{[t]}, A = a, Z_0) \\ &\quad \times \prod_{s=1}^{[t]} f(M_s \mid T > s, \bar{L}_s, \bar{M}_{s-1}, A = a^*, Z_0) \\ &\quad \times f(L_s \mid T > s - 1, \bar{L}_{s-1}, \bar{M}_{s-1}, A = a, Z_0) f(Z_0) dM_s dL_s dZ_0 \end{aligned} \quad (7.17)$$

where \bar{M}_s (\bar{L}_s) is the history of the mediator (time-varying confounders) up to time s and $[t]$ is the visit time at or before time t . This result is based on the edge g-formula, which is applicable to nested counterfactuals [Shpitser and Tchetgen, 2016]. The time-varying confounder, L , can be separated into components: the individual’s clinical characteristics, V , and an indicator of that individual continuing to be at risk. Therefore, the final term of equation 7.17 can be rewritten:

$$\begin{aligned} f(L_s \mid T > s - 1, \bar{L}_{s-1}, \bar{M}_{s-1}, A = a, Z_0) &= \\ f(V_s \mid T > s - 1, \bar{L}_{s-1}, \bar{M}_{s-1}, A = a, Z_0) \Pr(T > s \mid T > s - 1, \bar{L}_{s-1}, \bar{M}_{s-1}, A = a, Z_0) \end{aligned} \quad (7.18)$$

7.4.3 Estimation

Estimation of $S_{A(a),M(a^*)}(t)$ is accomplished via repeated regressions. The following is an illustration of the procedure described in Vansteelandt et al. [2019]. Consider estimation at time t , where $[t] \geq 2$ (i.e. after the second visit) The first term of Equation 7.17 can be estimated using any survival modelling approach, such as Cox regression, restricted to the individuals with exposure $A = a$. This fitted

model is then used to predict the expected survival probability to time t for all individuals still at risk at visit $\lfloor t \rfloor$ using observed values of M and L . This yields a vector, $Q^{\lfloor t \rfloor}(t)$. Rewriting equation 7.17 after this first step, we have:

$$S_{A(a),M(a^*)}(t) = \int Q^{\lfloor t \rfloor}(t) \times \prod_{s=1}^{\lfloor t \rfloor} f(M_s | T > s, \bar{L}_s, \bar{M}_{s-1}, A = a^*, Z_0) \\ \times f(L_s | T > s - 1, \bar{L}_{s-1}, \bar{M}_{s-1}, A = a, Z_0) f(Z_0) dM_s dL_s dZ_0 \quad (7.19)$$

Next, models are fit for each of the remaining components, $f(M_s)$, $f(V_s)$ and $P(T \geq s)$, from equations 7.17 and 7.18 and predict fitted values, working backwards by visit number as described next. Beginning with visit $k = \lfloor t \rfloor$ and repeating for each subsequent visit $(\lfloor t \rfloor - 1), \dots, 1$:

1. Regress $Q^k(t)$ on $\bar{M}_{k-1}, \bar{L}_k, Z_0$ for those with exposure $A = a^*$ who were at risk at visit k . This model is used to predict expected values, $Q_M^k(t)$ for all individuals. After this step, we can write:

$$S_{A(a),M(a^*)}(t) = \int Q_M^k(t) \times f(L_k | T > k - 1, \bar{L}_k, \bar{M}_{k-1}, A = a, Z_0) \\ \times \prod_{s=1}^{k-1} f(M_s | T > s, \bar{L}_s, \bar{M}_{s-1}, A = a^*, Z_0) \\ \times f(L_s | T > s - 1, \bar{L}_{s-1}, \bar{M}_{s-1}, A = a, Z_0) f(Z_0) dM_s dL_s dZ_0 \quad (7.20)$$

2. $Q_M^k(t)$ is then regressed on $\bar{M}_{k-1}, \bar{L}_{k-1}, Z_0$ for those with exposure $A = a$ who were at risk at visit k and fitted values are predicted for all who were still at risk at visit $k - 1$, using observed data. Call this result $Q_L^{k-1}(t)$ and write:

$$S_{A(a),M(a^*)}(t) = \int Q_L^{k-1}(t) \Pr(T > k | T > k - 1, \bar{L}_{k-1}, \bar{M}_{k-1}, A = a, Z_0) \\ \times \prod_{s=1}^{k-1} f(M_s | T > s, \bar{L}_s, \bar{M}_{s-1}, A = a^*, Z_0) \\ \times f(L_s | T > s - 1, \bar{L}_{s-1}, \bar{M}_{s-1}, A = a, Z_0) f(Z_0) dM_s dL_s dZ_0 \quad (7.21)$$

3. $Q_L^{k-1}(t)$ will be multiplied by the expected values obtained from a survival model fitted to the group of individuals with $A = a$ who were at risk at visit

$k - 1$ using covariates $\bar{M}_{k-1}, \bar{L}_{k-1}, Z_0$. This product is called $Q^{k-1}(t)$. After rewriting the equation for $S_{A(a), M(a^*)}(t)$ as:

$$\begin{aligned}
S_{A(a), M(a^*)}(t) &= \int Q^{k-1}(t) \times \prod_{s=1}^{k-1} f(M_s | T > s, \bar{L}_s, \bar{M}_{s-1}, A = a^*, Z_0) \\
&\quad \times f(L_s | T > s - 1, \bar{L}_{s-1}, \bar{M}_{s-1}, A = a, Z_0) f(Z_0) dM_s dL_s dZ_0
\end{aligned}
\tag{7.22}$$

the process is repeated for the next visit time.

After computations for visit time 1 are complete, we obtain estimated survival probabilities to time t for each individual, $Q_0(t)$. Averaging over these predicted probabilities provides an estimate of $S_{A(a), M(a^*)}$ at time t . The method is not tied to any particular parametric model, therefore, a variety of suitable models may be used at each step but Vansteelandt et al. [2019] suggest Cox regression and quasi-binomial models with a logit link.

7.4.4 Assumptions

Consistent with most mediation analysis methods, the method of Vansteelandt also assumes no unmeasured confounding as in assumptions A1-A3 above. However, assumption A4, which forbids the presence of other mediators regardless of whether they are measured, is not a requirement. Rather, time-varying confounders are accommodated as long as the prescribed causal ordering is observed. It is assumed that there are no unmeasured common causes of the mediator and the time-varying confounders but they permit unmeasured common causes of the mediators and unmeasured common causes of the time-varying confounders with the outcome (U_m and U_l in Figure 7.5). Non-informative visit times and censoring are required as well as the absence of competing risks. Finally, identification of the path-specific effects also requires that the causal structure represents a non-parametric structural equation model with independent errors including the unmeasured variables U_m and U_l . Many of these assumptions are untestable and, as with all causal analyses, care must be taken.

7.5 Analysis of the UK CF Registry

7.5.1 Overview

In this section I describe an application of the methods of Aalen and Vansteelandt to the CF Registry data to investigate the role of three potential mediators of the

effect of CFRD on survival. The method of Vansteelandt was recently applied by Buse et al. [2020] to explore possible mediators of the effect of treatment with liraglutide on cardiovascular events and by Kalogeropoulos et al. [2020] to investigate their hypothesis that the drug spironolactone affects hospitalisation with heart failure via its diuretic effect. The method of Aalen does not yet appear to have been used in practice.

7.5.2 Study population

The study population comprised all individuals aged 18 - 60 years in the UK CF Registry with at least two annual reviews between 1/1/2008 and 31/12/2017. Individuals without at least two measurements of FEV1% or BMI were omitted as well as those without information about whether or not they have two copies of the F508del CFTR mutation. After these restrictions, there were 6,374 individuals with a combined 37,896 records between 2008 and 2017 remaining. Pancreatic insufficiency is closely associated with CFRD and, in this population, there were only 198 people who were simultaneously pancreatic sufficient and had CFRD. Because of this small number, the study population was further restricted to include only records where the individual was pancreatic insufficient leaving 5,453 individuals with 32,304 annual review records. After removing prevalent cases of CFRD, the final study population consisted of 3,708 individuals with 18,963 annual review records. The key differences between this study population and the incident cohort of the previous chapter are (1) the requirement that each individual have at least two annual reviews between 2008 and 2017 and (2) the restriction to pancreatic insufficient individuals; the incident cohort contained individuals with only one annual review and people who were pancreatic sufficient. 52% of the individuals in this study population had eight or more annual reviews recorded.

7.5.3 Mediators, confounders and outcome

Based on current hypotheses of the possible mechanism(s) for CFRD affecting survival (see section 6.2), three candidate mediators are considered in this analysis: lung function, nutritional status and respiratory infections. The hallmark of CF is obstructive lung disease and the severity of that lung disease is best characterised by FEV1%. FEV1% was reviewed in section 3.3. The standard of measuring nutritional status via BMI is adopted here. In contrast to the volatile trajectories of FEV1 seen in Figure 3.2, BMI tends to have less variability over time. Figure 7.7 plots the trajectories of BMI for 80 individuals selected at random from the study population. Each measurement of BMI was taken at the individual's annual review. The cycle of infections and inflammation in CF harms the lung tissue over time and infections are often treated aggressively with intravenous antibiotics. To

BMI Trajectories for 80 Random Patients

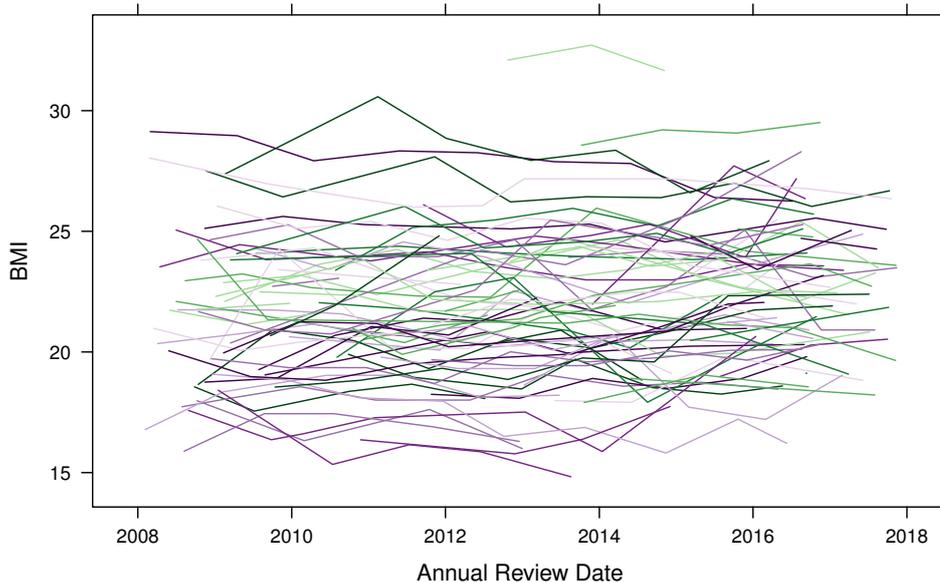
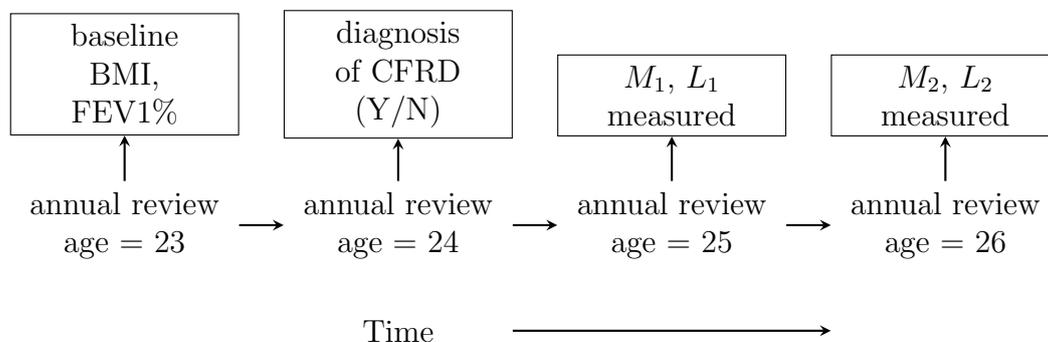


Figure 7.7: Each line plots the approximately annual measurements of BMI for one randomly selected person with CF from the study population. Measurements were taken between 1/1/2008 and 31/12/2017. Both right- and left-truncation of the data can be seen.

capture occurrence of respiratory infections, the number of days in hospital receiving IV antibiotics (IV days) is used. According to UK Cystic Fibrosis Registry [2020], 38.0% of people in the UK CF population received at least one course of IV antibiotics in hospital. The IV days data was categorised into 6 categories as: 0 days, 1-7 days, 8-14 days, 15-21 days, 21-28 days and >28 days. 1,911 different individuals spent more than 4 weeks in hospital receiving IV antibiotics in one or more years with 1,211 of those individuals having CFRD in at least one of those years.

In each analysis, one of these three candidate mediators is selected as the mediator of interest. All analyses are adjusted for five baseline confounders: gender, genotype, calendar year, baseline FEV1% and baseline BMI. (See Table 7.1) Measurements for baseline confounders are taken at the review prior to the one where evaluation for CFRD occurred and the first mediator and time-varying confounder measurements are taken at the annual review after the one where the CFRD diagnosis was recorded. A sample timeline is provided in Figure 7.8. In this example, suppose an individual's annual review at age 24 shows a first-time diagnosis with CFRD. This indicates that a diagnosis of CFRD occurred either at the age 24

Figure 7.8: Sample timeline for data collection. Here, the diagnosis of CFRD (yes or no) was recorded at the annual review at age 24. Baseline confounder levels are taken from the annual review at age 23 and the first and second mediator measurements, M_1 and M_2 , are taken at the age 25 and 26 annual reviews.



annual review or some time between the prior annual review (at age 23) and this one. Similarly, the days in hospital receiving IV antibiotics is a measurement representing the previous one year. FEV1% and BMI are measured at the annual review. Therefore, to ensure that baseline values of FEV1% and BMI are taken prior to the exposure, measurements are used from the prior annual review; in our example, measurements would be taken from the age 23 annual review. BMI and FEV1% from the annual review when diagnosis was recorded are not included. Rather, the first measurement for the mediator and time-varying confounders is taken from the following annual review, age 25 in this example. Note also that CFRD is treated as a fixed exposure although it is time-dependent as individuals are screened annually for CFRD.

Consistent with Part I, “Dynamic Prediction of Survival”, the outcome is a composite of age at all-cause death or lung transplantation. For brevity, I will use the term mortality to refer to this composite outcome.

7.5.4 Implementation of the mediation methods

The dataset used in the mediation analyses bears some resemblance to a stacked landmark dataset in that measurement times are assumed to occur at integer-valued “landmark” ages and the analysis dataset comprises many smaller stacked datasets. An age-specific dataset was created at each integer age a from 18-50 that contains those individuals at risk at age a who either have not been diagnosed with CFRD or who have been diagnosed with CFRD within the past year. Because only data from the first time of diagnosis is used, each individual will only contribute data as an exposed person once. In contrast, people without CFRD may contribute

Table 7.1: Data from the UK Cystic Fibrosis Registry dataset used in the mediation analyses. The exposure is having a new diagnosis of CFRD. Three potential mediators are investigated and, in the method of Vansteelandt analysis, one mediator is selected and the other two become time-varying confounders. Baseline confounders are measured in the annual review prior to the assessment of CFRD.

<i>Category</i>	<i>Description</i>	<i>Type</i>
Exposure	Cystic fibrosis-related diabetes, incident case (CFRD)	Binary
Mediators / Time-Varying Confounders	Forced expiratory volume in 1 second as percentage of predicted (FEV1%)	Numeric
	Body mass index (BMI)	Numeric
	Days (past year) in hospital for IV antibiotics	Categorical
Baseline Confounders	Gender	Binary
	Genotype (F508del homozygous)	Binary
	Calendar year at measurement time	Numeric
	Prior year FEV1%	Numeric
	Prior year BMI	Numeric

Original Data:							Analysis Data:								
id	age	CFRD	FEV1	sex	age at		id	age	start	stop	event	CFRD	FEV1	sex	Baseline
					event	event									Confounders
A	23	0	43.5	M	27.0	0	A	24	0.0	1.0	0	0	56.56	M	43.5
A	24	0	56.6	M	27.0	0	A	24	1.0	2.0	0	0	48.33	M	43.5
A	25	0	48.3	M	27.0	0	A	24	2.0	3.0	0	0	38.94	M	43.5
A	26	0	38.9	M	27.0	0	A	25	0.0	1.0	0	0	48.33	M	56.6
B	23	0	43.3	F	26.2	1	A	25	1.0	2.0	0	0	38.94	M	56.6
B	24	0	36.9	F	26.2	1	A	26	0.0	1.0	0	0	38.94	M	48.3
B	25	1	29.5	F	26.2	1	B	24	0.0	1.0	0	0	36.90	F	43.3
B	26	1	35.0	F	26.2	1	B	24	1.0	2.0	0	0	29.51	F	43.3
							B	24	2.0	2.2	1	0	34.99	F	43.3
							B	25	0.0	1.0	0	1	29.51	F	36.9
							B	25	1.0	1.2	1	1	34.99	F	36.9

Not used, only incident cases of CFRD are considered

Figure 7.9: Construction of the analysis dataset. The table on the left represents the original data with one row of data per year the individual was seen in the registry in the study period. Individuals contribute age-specific sets of data, formatted with start and stop times for the mediator measurement, which are then vertically stacked to form the analysis dataset (right). Individual A contributes 3 age-specific sets of data (ages= 24, 25, 26); the age 23 data is used for baseline measurements to ensure proper causal ordering from $Z_0 \rightarrow A \rightarrow M_1$. Only people newly diagnosed with CFRD and people without CFRD contribute data.

data at multiple ages.

Figure 7.9 illustrates the dataset creation process. The data on the left is a subset of the data for two individuals, A and B, at ages 23 to 26. Each row contains information from the annual review that most closely precedes the date the individual turned age a . For example, if individual A's 23rd birthday occurred on 4 March 2008, and he had annual reviews on 2 July 2007 and 14 June 2008, the row in the data corresponding to age 23 would be taken from 2 July 2007 review. As explained in the previous section, an indication that an individual has been diagnosed with CFRD may reflect a diagnosis that occurred some time prior to the annual review. To ensure the correct causal ordering, the values of FEV1% and BMI from age $a-1$ are used as the baseline confounders. Therefore, in the example of Figure 7.9, neither A nor B will contribute data to an age-specific dataset at age $a=23$ because there are no prior year values to use as baseline measurements. Their data from age 23 is used for baseline adjustment of the age $a=24$ data. To form the $a=24$ dataset for A, time is set to zero at age 24 and rows of data are created with start and stop ages indicating the age range over which the measurements are valid. The event indicator equals 1 where the stop age is the time that the person

had the event. The process repeats for each age a that the person is at risk and has not previously been diagnosed with CFRD. For A, the age $a=25$ and age $a=26$ datasets contain 2 rows and 1 row, respectively. Individual B had an event at age 26.16 so her data for the age $a=24$ dataset consists of three rows: time 0-1 (age 24 to 25), time 1-2 (age 25 to 26), and time 2 to 2.16 (age 26 to age at event). Note also that the exposure is set to zero (unexposed) in all rows of the age 24 dataset despite B being diagnosed with CFRD at age 25. The exposure is considered fixed at time 0; our analyses do not accommodate time-varying exposures. In B's age 25 dataset, the exposure is set to 1 (exposed) to match her diagnosis. B does not contribute an age 26 dataset because she is no longer an incident case at this age. Once the age-specific datasets have been created for all individuals they are vertically stacked to form one analysis dataset.

For each analysis, one of the candidate mediators, FEV1, BMI, or IV days was selected as the mediator to study. Additionally, because the method of Vansteelandt accommodates time-varying confounders, the other two candidate mediators were incorporated as time-varying confounders.

To facilitate comparison between the two mediation methods, results are presented for the total effect, direct effect and indirect effect on a relative survival scale (equations 7.12, 7.13, 7.14, 7.15). The effects presented are a ratio of probabilities per the previously defined contrasts. Ratios are presented instead of differences because calculation of differences using the method of Aalen would require estimation of the function capturing the effects of the baseline confounders, $f(t, Z_0)$, an item Aalen et al. [2020] have left for future research. The earliest estimate of indirect effect was at the first visit time in both analysis methods as this is when the first mediator measurement is taken after the exposure. In the method of Aalen, the earliest estimates of the total effect and direct effect were also at $t = 1$ because only individuals who have survived to the first mediator measurement time are included. This is a requirement in the method of Aalen and, therefore, the estimate is conditional. In contrast, the earliest estimate of the total effect and direct effect in the method of Vansteelandt are immediately after the exposure. The regressions from the dynamic path analysis in the method of Aalen are performed at each unique event time. In the method of Vansteelandt, effect estimates are computed every 0.1 years beginning at time $t = 0.05$.

Confidence intervals at each visit time were computed via non-parametric bootstrap with 500 bootstrap samples. Because the analysis dataset includes multiple rows per age per person, resampling was done at the individual level. 95% bootstrap confidence intervals are determined using the percentile method.

All analyses were performed using R v4.0.2 [R Core Team, 2020]. The method of Aalen was implemented using R code provided by the authors in their supplementary materials [Aalen et al., 2020]. I wrote code for the method of Vansteelandt

and made use of the R package `survival` [Therneau and Grambsch, 2000, Therneau, 2015] for fitting a Cox proportional hazards model as the survival model. Resulting survival probabilities were modelled using a quasi-binomial regression with a logit link as recommended in Vansteelandt et al. [2019].

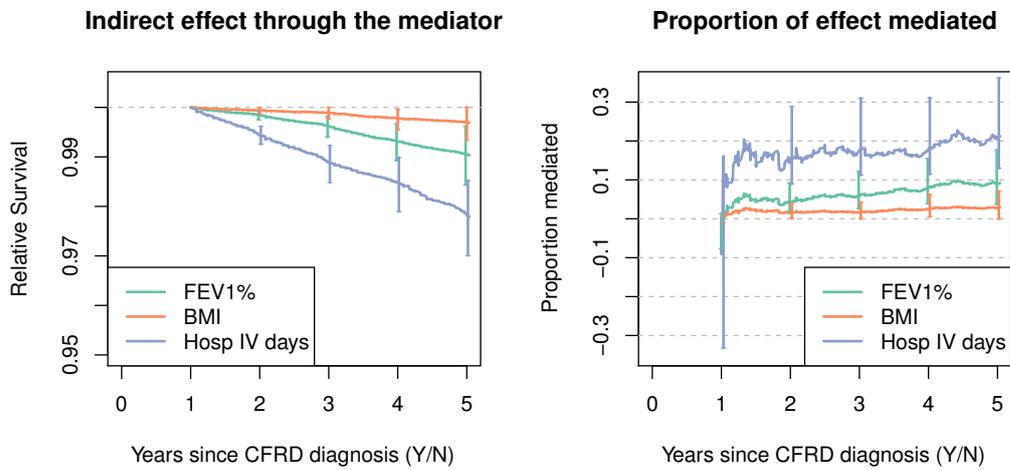
7.5.5 Results

Overview

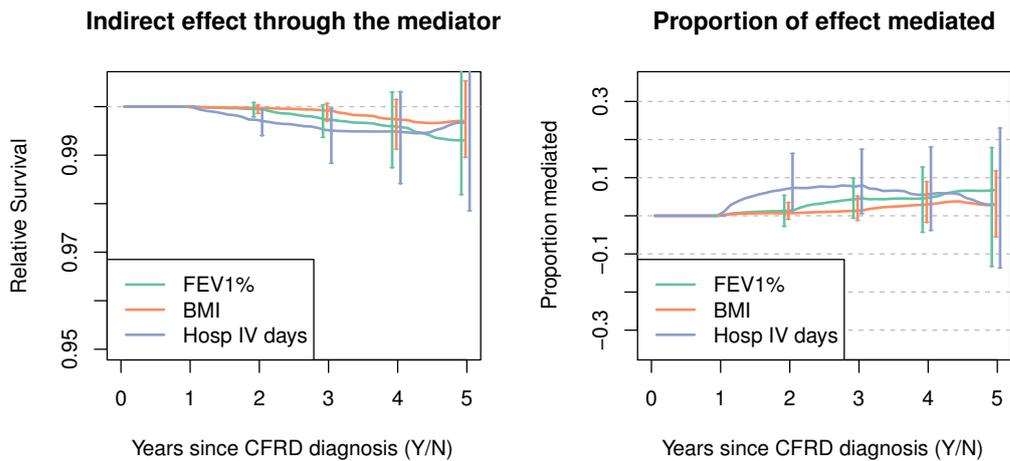
Mediation analyses using the method of Aalen and the method of Vansteelandt to explore the three potential mediators both resulted in similar findings. A small portion of the total effect of CFRD on mortality may be mediated through lung function and respiratory infections but nutritional status was not found to be a significant mediator. Figure 7.10 shows the estimated indirect effect via the mediator and the proportion of the total effect that is mediated for each of the three potential mediators using the method of Aalen (top row) and the method of Vansteelandt (bottom row). For the method of Aalen, the estimated indirect effect increases over time for all three mediators but is modest in size. For the method of Vansteelandt, the estimated indirect effect is smaller and bootstrap confidence intervals contain 1.0 at each visit time for both BMI and FEV1%. The proportion of the effect mediated is estimated to be greatest for respiratory infections, as proxied by IV days when using the method of Aalen. Respiratory infections also reach a higher proportion mediated using the method of Vansteelandt but the estimated proportion decreases beyond three years after CFRD diagnosis. In the next sections, these results are explored in more depth for both methods.

Lung function as a mechanism for the effect of CFRD on mortality

Figure 7.11 shows the results of the method of Aalen mediation analysis with lung function, as measured by FEV1%, as the mediator. The four graphs plot the total effect, indirect effect via FEV1%, direct effect not via FEV1% and the proportion of the effect mediated by FEV1%. The x-axis time scale begins at time $t = 0$ when the diagnosis of CFRD/ no CFRD was indicated in the annual review record and continues to 5 years post-CFRD evaluation. No estimates are shown prior to 1 year after the CFRD evaluation because the method of Aalen analysis is conditional on survival to the first mediator measurement. The three effect graphics are plotted as relative survival, i.e. the plot depicts a ratio as defined in Equations 7.12 and 7.13. Each graph also shows confidence intervals determined by non-parametric bootstrap ($n = 500$) at each visit time. Consistent with the descriptive analysis presented in chapter 6, a significant total effect of CFRD on survival is seen and this effect increases with time since diagnosis. Note that the estimated total effect



(a) Method of Aalen effect estimates for the three candidate mediators.



(b) Method of Vansteelandt effect estimates for the three candidate mediators.

Figure 7.10: Results from the method of Aalen (top row) and the method of Vansteelandt (bottom row) illustrating the indirect effect via the mediator (left) for each of the three candidate mediators: lung function (FEV1%), nutritional status (BMI) and respiratory infections (IV days). The proportion of the total effect that is mediated by each candidate mediator is shown on the right. 95% bootstrap confidence intervals are shown for each visit time.

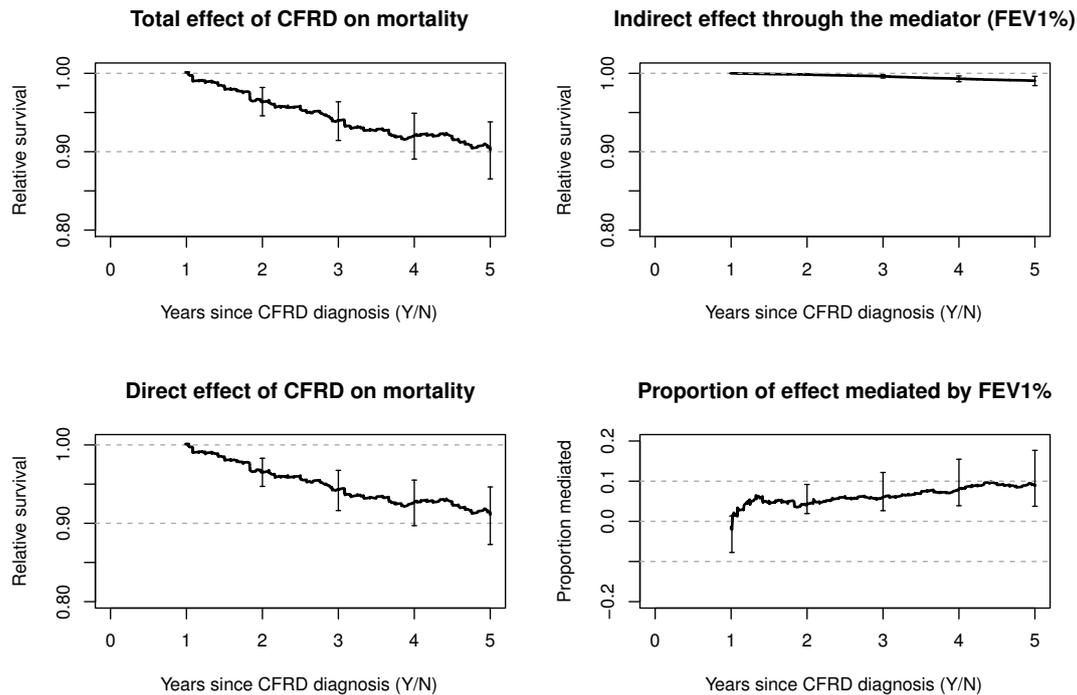


Figure 7.11: Effect estimates using the method of Aalen with FEV1% as the candidate mediator. Time 0 on the x-axis represents the year when diagnosis of CFRD / no CFRD was recorded at the annual review. Because this analysis is conditional on survival to the first mediator measurement at time $t = 1$ year, effect estimates begin at 1 year post-evaluation. 95% bootstrap confidence intervals are shown at each visit time.

is independent of the choice of mediator. The indirect effect of CFRD on survival via lung function is small; at 3 years after CFRD evaluation the indirect effect is 0.94 [95% CI: 0.91, 0.96] and at 5 years post-evaluation the indirect effect is 0.90 [95% CI: 0.87, 0.94]. The interpretation of the indirect effect estimate at 3 years is that if everyone had been diagnosed with CFRD, the ratio of the probability of survival at 3 years if everyone had their mediators set to levels they would be at with a positive diagnosis to the probability of survival if everyone had their mediators set to levels they would have been at if they had not been diagnosed with CFRD is 0.94. The proportion of the effect mediated by FEV1% slowly rises over time from 4% [95% CI: 2%, 9%] at $t = 2$ to 6% [95% CI: 3%, 12%] at $t = 3$ and 9% [95% CI: 4%, 18%] at $t = 5$.

Figure 7.12 shows the method of Vansteelandt estimated survival curves $S_{A(1),M(0)}(t)$, $S_{A(1),M(1)}(t)$ and $S_{A(0),M(0)}(t)$ when FEV1% is the candidate medi-

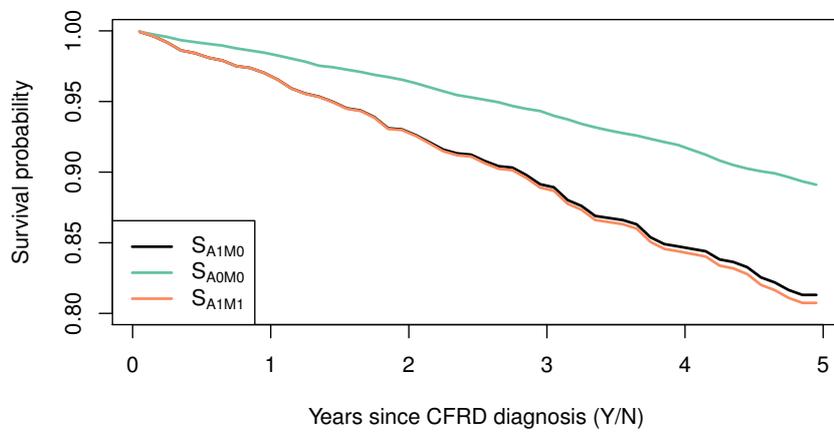


Figure 7.12: Survival curves estimated using the method of Vansteelandt with FEV1% as the candidate mediator. Contrasts between these three survival curves provide the estimates of total effect, indirect effect and direct effect. S_{Aa, Ma^*} is the estimated survival curve for a person selected at random with exposure set to a , mediator levels set to a^* , and all other covariates at the levels that would have been seen with an exposure of a .

ator. The ratio $S_{A(1),M(1)}(t)/S_{A(0),M(0)}(t)$ defines the total effect. If the curve $S_{A(1),M(0)}(t)$ (black) were to perfectly coincide with the $S_{A(1),M(1)}(t)$ curve (orange), this would indicate that no indirect effect via the mediator was found. Conversely, if the $S_{A(1),M(0)}(t)$ curve equals the $S_{A(0),M(0)}(t)$ curve (green), this would indicate that all of the total effect was indirect, via the pathways where the exposure acts first on the mediator. In figure 7.12, there is evidence of both an indirect effect and a direct effect, however, because the black and orange curves have little separation, the estimated indirect effect is small. The confidence intervals for the indirect effect from the method of Vansteelandt at each visit time include 1.0. (See Figure 7.13-upper right). Point estimates of indirect effect over time from the method of Vansteelandt are nearly identical to those from the method of Aalen. The proportion of the effect mediated by FEV1% is estimated by the method of Vansteelandt to increase slowly from 1% [95% CI: -3%, 5%] at 2 years post-evaluation to 4% [95% CI: -1%, 10%] after 3 years and to 7% [95% CI: -14%, 18%] after 5 years. (Figure 7.13-lower right).

In the method of Vansteelandt, measurement of the total effect and direct effect of CFRD on survival begins immediately after evaluation at time $t = 0$. After 1 year, these effects are significantly different than 1.0 throughout the study period. (See Figure 7.13-upper left, lower left). At 5 years post-evaluation, the total effect is estimated at 0.91. The interpretation is that, for a random person from the study population, the ratio of their probability of survival at 5-years after a positive CFRD diagnosis (with mediators and other covariates at the levels they would have been at with a positive diagnosis) to their probability of survival at 5-years had they not been diagnosed with CFRD at that age (with mediators and other covariates at the levels they would have been at with a negative diagnosis) is 0.91. If the necessary identification assumptions hold, we can conclude that the negative effect on survival observed for CFRD is not explained by known mediators and confounders and appears to be a direct result of the presence of CFRD.

Respiratory infections as a mechanism for the effect of CFRD on mortality

The two analyses with days in hospital receiving IV antibiotics as the mediator produced somewhat different results. Figure 7.14 (top row) plots the estimates obtained from the method of Aalen. The estimated proportion mediated via IV days reaches 15% [95% CI: 9%, 29%] at 2 years post-evaluation and 21% [95% CI: 13%, 36%] after 5 years. The method of Vansteelandt analysis estimates the path-specific effects via the mediator to be smaller and, at all times, the 95% bootstrap confidence interval includes 1.0 as shown in Figure 7.14 (bottom row). The maximum estimated proportion mediated, 8% [95% CI: 1%, 17%], occurs at 3 years post-evaluation and declines to 3% [95% CI: -14%, 23%] after 5 years. The confidence intervals include 0.0 at 4 and 5 years post-evaluation and the lower

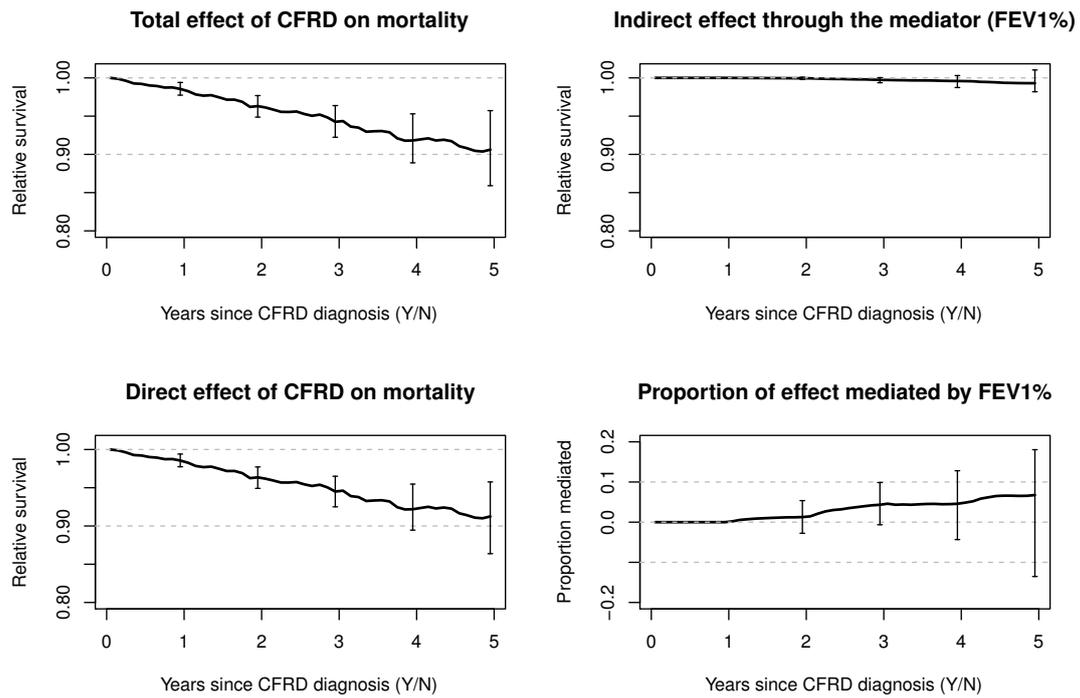


Figure 7.13: Effect estimates using the method of Vansteelandt with FEV1% as the candidate mediator. Time 0 on the x-axis represents the year when diagnosis of CFRD / no CFRD was recorded at the annual review. 95% bootstrap confidence intervals are shown at each visit time.

bound is 1% at 2 and 3 years post-evaluation indicating borderline significance of mediation by IV days.

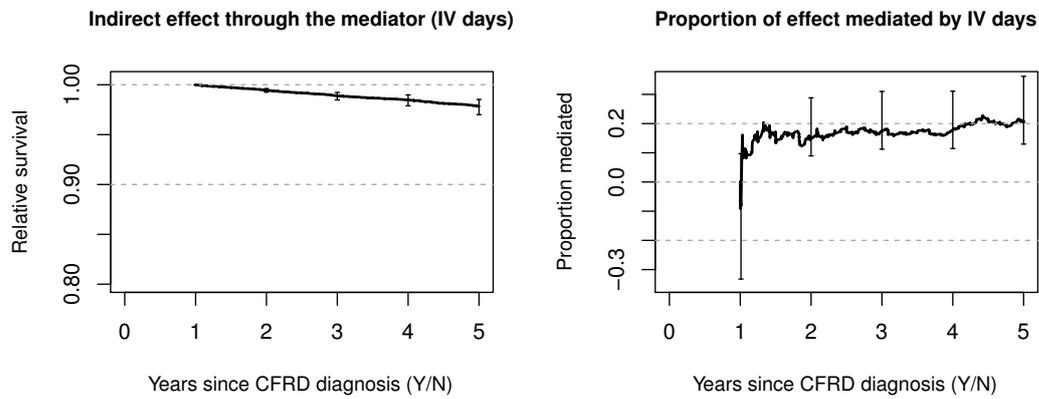
Nutritional status as a mechanism for the effect of CFRD on mortality

Neither the method of Aalen nor the method of Vansteelandt found evidence of nutritional status as measured by BMI mediating the effect of CFRD on mortality. The estimated indirect effect from the method of Aalen is 1.00 at all times beginning with 1 year post-diagnosis. Although the estimated proportion mediated by BMI is greater than 0%, the confidence intervals include 0. Using the method of Vansteelandt, again the indirect effect is estimated to be 1.00 for all times post-evaluation. The confidence intervals for proportion mediated include 0 at all times. (See Appendix figure E.1)

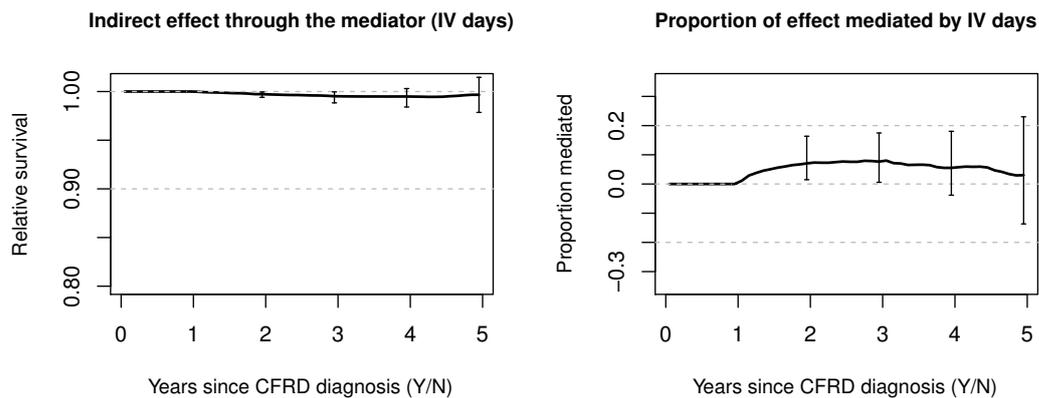
7.6 Discussion

Two recently proposed methods for mediation analysis in the setting of a time-varying mediator and a survival outcome have been applied to data from the UK CF Registry. Based on current clinical research hypotheses, I investigated whether the total effect of CFRD on mortality or transplant is mediated by lung function, respiratory infections and/or nutritional status. Despite some methodological and data requirement differences, both methods produced similar effect estimates. In particular, both methods estimated the proportion of the total effect mediated by lung function to be less than 10% at all times up to 5 years post CFRD diagnosis. A small portion of the total effect may be mediated by respiratory infections but neither method found evidence of nutritional status being a mediator. All of the effect estimates were associated with a large degree of uncertainty. Because causal mediation analyses are complex and generally involve accepting untestable assumptions, care must be taken when interpreting the results.

Both methods studied offer different advantages and disadvantages. Computationally, the method of Aalen is simple to implement using code provided by the authors and analyses run in a matter of minutes. Conceptually, however, the assumption that the treatment or exposure can be split into two biologically separate parts could be difficult to justify. Also, the need for confounding to be controlled by exogenous variables measured at baseline will not be plausible in many scenarios. In our study, CFRD has a complex pathophysiology where multiple systems are affected by the diabetes. It is conceivable, at least in theory, that these different effects could be separated into an effect on lung function, an effect on nutritional status, an effect on respiratory infections and other effects. What seems less plausible, however, is that sufficient control for confounding can



(a) Method of Aalen effect estimates with IV days as the candidate mediator.



(b) Method of Vansteelandt effect estimates with IV days as the candidate mediator.

Figure 7.14: Estimates of indirect effect of CFRD on mortality via days in hospital receiving IV antibiotics for the two mediation analysis methods. Time 0 on the x-axis represents the year when diagnosis of CFRD / no CFRD was recorded at the annual review. 95% bootstrap confidence intervals are shown at each visit time.

be made using only baseline covariates. In our setting, a relationship between BMI and lung function has been established [Stephenson et al., 2013] as well as the decline in lung function associated with repeated cycles of respiratory infection and inflammation [Stanford et al., 2021]. These relationships suggest that time-varying confounding is likely present. Didelez [2018] points out that because this method does not involve an intervention on the mediator itself, the role of adjustment for time-varying confounders is to ensure that the mediator and survival processes are conditionally independent of both separated treatments.

In contrast, the method of Vansteelandt is computationally intensive. The authors have provided SAS code but users must implement it themselves if using other languages. Also, more data is required as the number of visit times increases; the number of covariates in each model increases with visits because we regress on the history of the covariates. For example, in our study, at visit time 4 the survival model had 22 covariates. There are fewer survivors to later visit times and having sufficient data to fit the models could be a concern. The key advantage to the method of Vansteelandt is the ability to accommodate time-varying confounders. Although this makes the models more complicated, it allows the method to be applied in a broader range of settings where control for confounding cannot be accomplished through baseline covariates alone. Inclusion of time-varying confounders requires assumptions about the causal ordering between L_k and M_k . As seen in Figure 7.5, a time-varying confounder may be another mediator. In this study, results suggest that two of the three candidate mediators may mediate the total effect of CFRD on mortality. If that is true, then failing to account for the time-varying confounding of a second mediator could produce biased results. Finally, the method of Vansteelandt is flexible in allowing for different models to be applied at different steps and is not tied to a particular specification.

Implementing these methods using registry data presented several challenges. First, when the exposure is a randomly-assigned treatment, the need to adjust for baseline confounders of the exposure is alleviated. In this study, it was necessary to adjust not only for time-fixed covariates such as gender and genotype but also for baseline values of the mediator. Our exploratory data analysis suggested that incident cases of CFRD had a lower FEV1% than people without CFRD after adjustment for age, sex and genotype and, therefore, adjustment for a baseline value of FEV1% was required. As expected, the analysis without adjustment for baseline FEV1% estimated the indirect effect to be much greater: 0.96 at time $t = 5$ versus 0.99 after adjustment. Because the baseline value of FEV1% will be highly correlated with future measurements of FEV1%, it seems possible that this adjustment may introduce bias. This hypothesis will be investigated via a simulation study in chapter 8.

An additional challenge when using the onset of disease as an exposure is the

definition of time 0. For individuals who are diagnosed with CFRD, the diagnosis date serves as time 0 but there is no natural time 0 for those who haven't been diagnosed. A stacked analysis dataset is proposed to exploit all of the information in the registry dataset. As a consequence, most individuals will be represented multiple times in the dataset. Although this causes no issues with the point effect estimates, it must be accounted for in the inference. I opted for a non-parametric bootstrap with resampling at the individual level to construct confidence intervals. Ideally, 1,000 bootstrap samples would be used but the method of Vansteelandt is computationally intensive and the estimates with 500 samples appear stable based on finding no difference in the confidence intervals generated with only 250 bootstrap samples.

To ensure proper causal ordering of the data, the first mediator measurement was taken from the annual review after the review where CFRD was first diagnosed – a time gap of one year. Because lung function is measured at the annual review when CFRD is first diagnosed, another option would have been to use that FEV1% measurement as the first mediator measurement and set the time to some small increment after time 0. This would allow more individuals to remain in the method of Aalen analysis as there is a requirement of survival to the first mediator measurement and the direct effect could be separated from the indirect effect sooner. This approach is not recommended although it has been used often in the literature. In a review of mediation studies with survival outcomes, Lapointe-Shaw et al. [2018] found that 60% used an exposure and mediator that were both measured at baseline. The failure to guarantee separation in time between exposure and mediator risks the possibility of reverse causation.

In mediation studies, the proportion mediated is a commonly reported statistic. In the Lapointe-Shaw et al. [2018] review of mediation analysis with time-to-event outcomes, they found that while an estimate of indirect effect was only reported in 37% of studies, an estimate of proportion mediated was reported in 56% of the studies. Nearly all of the aforementioned studies reporting indirect effect also provided a measure of uncertainty but under half of the studies reporting proportion mediated included an estimate of uncertainty. The proportion mediated has intuitive appeal and it seems to answer the question being asked in a mediation analysis: how much of the total effect is mediated by M ? Unfortunately, using this measure has some drawbacks. Because the proportion mediated is a ratio of the indirect effect to the total effect, it is sensitive to the magnitude of the total effect estimate. Richiardi et al. [2013] point out that the proportion mediated will not be useful in situations where the exposure positively affects both the outcome and a mediator which negatively affects the outcome, resulting in a total effect of 0. In other words, when the direct and indirect effects are in opposite directions, the proportion mediated will not be intuitively interesting. In this study, the

proportion mediated was associated with large confidence intervals. Particularly with the method of Aalen, there were also volatile estimates close to the first visit time. (See in particular the top right graphic of Figure 7.14) This occurs because the analysis is conditional on survival to the first visit time and, therefore, both the total effect and the indirect effect just after the first visit are close to zero. Small changes in the effect estimates can lead to large changes in the proportion mediated in this case.

Although the results from the method of Aalen and the method of Vansteelandt have many similarities, their differences are important. For example, the confidence intervals produced by the method of Vansteelandt are wider than those from the method of Aalen. I believe this difference reflects the additional flexibility in the models being fit in the method of Vansteelandt. Especially at larger visit times, the number of covariates in each model is much larger than in the method of Aalen. I explore this further in the simulation study of chapter 8.

In section 7.5.4, it was explained that the method of Aalen analysis produces effect estimates conditional on survival to $t = 1$. Because of the stacked analysis dataset used, people who do not have CFRD can contribute data for every year they are at risk. If they experienced an event in the final year they are at risk, however, the age-specific dataset for that age will not be included because they didn't survive a full year to the next annual review. This equates to 2,708 records not being used in the method of Aalen because an event occurs prior to the first mediator measurement. These records are included in the method of Vansteelandt estimates of total and direct effects prior to $t = 1$. This difference means we do not expect the results of the two analyses to be equivalent in terms of direct effect, total effect or proportion mediated. Because the first estimate of indirect effect is at the same time with both methods, we do expect those effect estimates to be the same. This effect is explored further in section 8.7 in the context of a simulation study.

Both analyses suggest that less than 10% of the effect of CFRD on mortality is mediated through lung function. It is known that CFRD negatively impacts both survival and lung function and that respiratory failure is the primary cause of mortality in CF. However, the relationship between CFRD and lung function may be more complicated than our model allowed for. As discussed in chapter 6, other factors such as respiratory infection and BMI that are affected by CFRD may in turn affect lung function. An interesting future project may be to perform an analysis accounting for multiple mediators simultaneously. Another area for future work is to understand whether subgroup analysis may shed light on the mechanism for the total effect of CFRD on survival and transplant.

In the investigation of respiratory infections as a possible mediator of the effect of CFRD on mortality, the two analysis methods produced different indirect effect

estimates. A possible reason for this difference is that the method of Aalen only controls for confounding at the baseline while the method of Vansteelandt incorporates time-varying confounders. It seems likely that FEV1% and possibly BMI confound the relationship between respiratory infections and survival over time. FEV1% can vary dramatically year-by-year and is not only an important health indicator but is also known to be a key predictor of survival for people with CF. Failing to control for FEV1% after the time of evaluation of CFRD almost certainly violates the no unmeasured confounding assumption. This issue should be relevant in the analyses of all three candidate mediators but similar results were seen from both analyses for BMI and FEV1% as candidate mediators. Perhaps the indirect effect through BMI and FEV1% was too small for the residual confounding to add significant bias.

In the previous chapter, several more traditional analysis methods were employed to describe the association between CFRD and mortality. Although our goal in this chapter was to quantify mediation, estimates of the total effect of CFRD on death and transplant were also obtained. Using relative survival, the total effect is interpreted as the ratio of the probability of survival if everyone had CFRD to the probability of survival if no one had CFRD. Before such an interpretation is made, however, further thought should be given to the nature of this exposure, a disease diagnosis, and whether this is the appropriate exposure to use for a causal total effect interpretation. For future work, I suggest going beyond this average treatment effect measure and investigating the average treatment effect in the treated. Translated to our setting, this is the average effect of disease on the diseased.

Another interesting area for future work would be to compare the results from an analysis based the method of Zheng and van der Laan [2012]. Their approach is based on stochastic interventions, a concept introduced in Didelez et al. [2006]. In this method, mediator values are random draws from a given counterfactual distribution. Although natural effects can be defined in a similar way, they are different causal parameters. As this approach is methodologically quite different from the method of Aalen and method of Vansteelandt, it was out of scope for this thesis but a comparison between these three approaches designed for the setting of a time-to-event outcome with a repeatedly-measured mediator would be valuable.

The goal of this mediation study was to apply and compare two recently proposed methods for mediation in the setting of a time-varying mediator and a survival outcome and to highlight their relative strengths, weaknesses and differences. Overall, both methods produced similar effect estimates for our three candidate mediators of the total effect of CFRD on mortality. While the method of Aalen is easier to implement and produces estimates within minutes, the method of Vansteelandt may be more broadly applicable as it is able to accommodate time-

varying confounders in the analysis. Future research into including time-varying confounders into the method of Aalen would be welcome. Another interesting extension to both methods would be the inclusion of a time-varying exposure. These additions would make already complex methods more complicated and it is important to weigh the trade-offs between complexity, interpretability and transparency.

Chapter 8

Simulation Study: Comparison of Mediation Methods for Survival Outcomes and Time-Updated Mediators

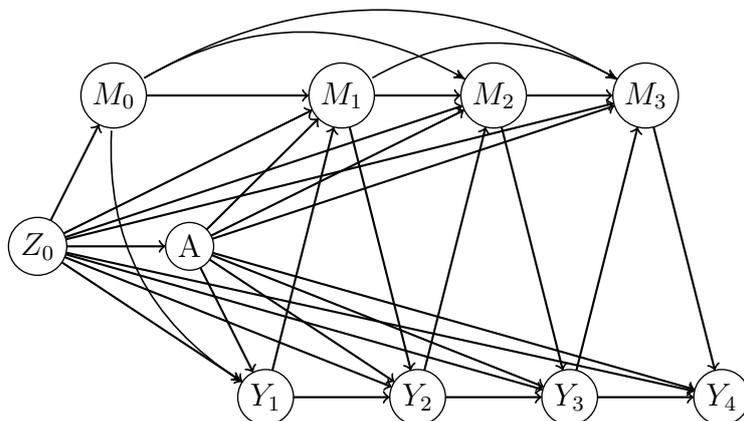
8.1 Introduction and aims

In chapter 7, two techniques for mediation analysis were applied to the UK CF Registry data to investigate possible mediators of the effect of CFRD on survival. In this chapter, a simulation study is used to further probe the two methods under controlled data scenarios. The aims of this simulation study are twofold. First, we wish to assess the performance of the two methods for mediation analysis used in the previous chapter, the method of Aalen et al. [2020] and the method of Vansteelandt et al. [2019]. Comparison of the two methods will be made in a baseline scenario where we expect good performance from both as well as in settings with model misspecification, unmeasured confounding and data-related issues characteristic of observational datasets. Second, the results of the simulation analysis will be used to better understand the limitations and assist in interpretation of the CFRD analysis presented in section 7.5 and make recommendations for future analyses of observational datasets using these techniques.

8.2 Data-generating mechanisms

For each individual there is a binary exposure A occurring at time $t = 0$, a single binary baseline confounder Z_0 and an event time T . The study period is assumed to last 4 years with measurement of the continuous mediator M_k at baseline,

Figure 8.1: Illustration of the data-generating mechanism. There are measurements of the baseline confounder Z_0 and the pre-exposure mediator level M_0 . The exposure A occurs at time $t = 0$. Event times are generated in waves. At each $k = 1, 2, 3$, those with event time $T > k$ have a further mediator measurement taken at visit time k . Y_k is an indicator defined as $Y_k = I(T > k)$.



M_0 , and at visit times $k = 1, 2, 3$. Administrative censoring is applied at $t = 4$ years. This setting is illustrated in Figure 8.1 and represents the baseline scenario. For this baseline scenario, three sub-scenarios are considered: (1) there is both a direct and an indirect effect (“DE+IE”); (2) there is no direct effect because the exposure does not affect the outcome (“NoDE”); and (3) there is no indirect effect because the exposure does not affect the mediator (“NoIE”). The NoDE sub-scenario corresponds to removal of all lines from A to Y_t in Figure 8.1. The NoIE sub-scenario corresponds to removal of all lines from A to M_t in the same figure.

The following steps outline the procedure for data generation for each individual $i = 1, \dots, n$ under the baseline scenario:

1. Generate A_i , the exposure at time $t = 0$, from a Bernoulli distribution with probability $p_A = 0.5$.
2. Generate $Z_{0,i}$ from a Bernoulli distribution with probability $p_{Z_0|A}$. Because the probability is allowed to depend on the exposure, we are effectively simulating the situation where the baseline covariate affects the exposure. In the baseline scenario, $p_{Z_0|A=1} = 0.6$ and $p_{Z_0|A=0} = 0.4$ and, therefore, $p_{Z_0} = 0.5$ in the simulated data.
3. Generate a random intercept $\mu_{m0,i}$ and random time slope $\mu_{m1,i}$ for the mediator M where $(\mu_{m0}, \mu_{m1}) \sim MVN(\mu_M, \Sigma_M)$. Obtain values for the baseline

mediator measurement as random draws where

$$M_{0,i} \sim N(\mu = \mu_{m0,i} + \beta_{Z_0}^s Z_{0,i}, \sigma = 1) \quad (8.1)$$

The effect of the baseline covariate on the mediator, $\beta_{Z_0}^s$, is assumed to be constant. The superscript s is added to distinguish this coefficient in the simulation study from similar coefficients used in the methods of Aalen et al. [2020] and Vansteelandt et al. [2019].

4. Obtain values $M_{t,i}$ for $t = 1, 2, 3$ as random draws where

$$M_{t,i} \sim N(\mu = \mu_{m0,i} + \mu_{m1,i}t + \beta_{Z_0}^s Z_{0,i} + \beta_{A_t}^s A_i, \sigma = 1) \quad (8.2)$$

The effect of the exposure on the mediator, $\beta_{A_t}^s$, is allowed to be time-varying but is set to a constant value in the baseline scenario.

5. Define the conditional hazard

$$\lambda_i(t \mid A_i, Z_{0,i}, M_{0,i}, M_{[t],i}) = \begin{cases} \alpha_0 + \alpha_A^s A_i + \alpha_{Z_0}^s Z_{0,i} + \alpha_{M_0}^s M_{0,i} & \text{for } t < 1 \\ \alpha_0 + \alpha_A^s A_i + \alpha_{Z_0}^s Z_{0,i} + \alpha_{M_{[t]}}^s M_{[t],i} & \text{for } 1 \leq t < 4. \end{cases} \quad (8.3)$$

where $[t]$ refers to the visit time at or prior to t and α_0 is a constant baseline hazard. The effect of the mediator on the hazard, $\alpha_{M_{[t]}}^s$, may vary at each visit time while the effects of the baseline covariate, $\alpha_{Z_0}^s$ and the exposure, α_A^s are time-fixed. Parameter values were chosen so that the resulting hazard is positive. The conditional hazard is given as an additive hazards model because, unlike the method of Vansteelandt, the method of Aalen relies on the assumption of additive hazards.

Time-to-event outcomes are then generated in waves $w = 1, 2, 3, 4$ corresponding to the intervals between visit times: $[0, 1)$, $[1, 2)$, $[2, 3)$, $[3, 4)$. For each wave where the individual remains at risk:

- (a) Draw u_i where $u_i \sim \text{Uniform}(0, 1)$
- (b) Calculate the event time for wave w as $T'_i = -\log(u_i)/\lambda_i(t \mid A_i, Z_{0,i}, M_{w,i})$
- (c) If $(T'_i + (w - 1)) < (w)$, then the event time is set to $T'_i + (w - 1)$. Otherwise, no event time is assigned in wave w and a new event time is generated in the next wave.
- (d) All individuals still at risk at time $t = 4$ are administratively censored.

The result is a dataset with values A , Z_0 , M_0, M_1, M_2, M_3 and T for each individual. Further, an event indicator is set to 1 if $T < 4$ and 0 otherwise. Because

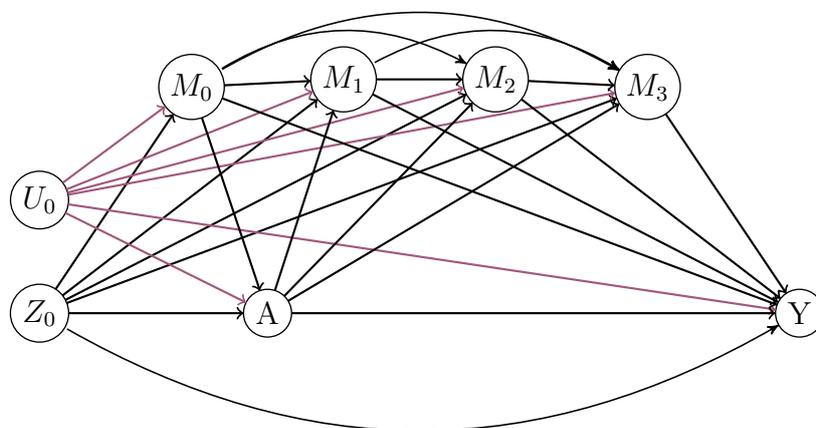
Table 8.1: Parameter values used in the mediator model (equations 8.1, 8.2) and hazard model (equation 8.3) to generate simulated data for the baseline scenario.

Mediator Model		Hazard Model	
Param	Value	Param	Value
$\mu_M =$	$\begin{bmatrix} 2.9 \\ 0.0 \end{bmatrix}$	$\lambda_0 =$	0.75
$\Sigma_M =$	$\begin{bmatrix} 0.250 & -0.015 \\ -0.015 & 0.010 \end{bmatrix}$	$\alpha_{M_{[t]}}^s =$	0.3 for $[t] = 0,1,2,3$
$\beta_{Z_0}^s =$	0.5	$\alpha_{Z_0}^s =$	0.35
$\beta_{A_t}^s =$	2* for $t = 1,2,3$	$\alpha_A^s =$	0.35**
* or 0 for NoIE scenario		** or 0 for NoDE scenario	

the method of Aalen requires survival up to the first mediator measurement, all simulated individuals with event times $T \leq 1$ do not contribute to the method of Aalen analyses. Table 8.1 summarises the parameter values used to generate the simulated data for the baseline scenario.

Both the method of Aalen et al. [2020] and the method of Vansteelandt et al. [2019] are expected to perform well with the baseline scenario as it was designed to be consistent with the assumptions of both methods. In addition to this baseline scenario, alternative scenarios were created to explore a variety of situations that may be encountered in an observational dataset. The first set of scenarios includes unmeasured confounding by generating simulated data with an additional baseline covariate, U_0 , that is unknown to the mediation analysis. A second set of scenarios use event times generated with multiplicative hazards models as opposed to additive hazards. This results in model-misspecification in the method of Aalen, which relies on an additive hazard model, but not in the method of Vansteelandt, which allows for any survival model. Three further scenarios were considered: (1) the situation where the baseline mediator measurement affects the exposure, (2) cases where time-varying covariates are present, and (3) the impact of infrequent mediator measurements. As described above, for each scenario, three sub-scenarios were created: “DE+IE”, “NoDE” and “NoIE”. A table detailing each scenario and the parameter values used to generate the simulated data can be found in Appendix F. A reference list of all simulation scenarios can be found in table 8.3 at the end of this section. Because some of these simulation scenarios will consider additional data or make different assumptions about the underlying data, minor modifications to the data generation procedure may be required as outlined below.

Figure 8.2: Purple lines show the possible relationships resulting from the addition of an unmeasured baseline confounder, U_0 , to the baseline scenario. For simplicity, the outcome Y is represented by a single node.



8.2.1 Unmeasured confounding

For scenarios examining the impact of including an unmeasured baseline confounder, U_0 is generated from a Bernoulli distribution with probability $p_{U_0|A}$ after generation of A but before generating the mediator and event time values. By adding an additional term to equations 8.2 and/or 8.3, U_0 may affect the mediator and the event time via parameters $\beta_{U_0}^s$ and $\alpha_{U_0}^s$, respectively. To create scenarios with exposure-outcome confounding (U1,U2), an effect of U_0 is added to the hazard model and the values of U_0 and A are correlated. For the scenarios with mediator-outcome confounding (U3,U4), an effect of U_0 is added to both the mediator model and the hazard model but the values of U_0 and A are uncorrelated. Finally, in the scenarios with exposure-mediator confounding (U5, U6), an effect of U_0 is added to the mediator model and the values of U_0 and A are correlated. Figure 8.2 illustrates the addition of an unmeasured confounder U_0 with possible effects indicated with purple lines.

8.2.2 Multiplicative hazards for event time generation

Scenarios with multiplicative hazard models (M1-M3) were explored using a Weibull model for the hazard. In these scenarios, step 5 described above in the data generation procedure for the baseline scenario is replaced with the following. First,

define the conditional hazard as:

$$\lambda_i(t \mid A_i, Z_{0,i}, M_{0,i}, M_{[t],i}) = \begin{cases} \kappa b t^{\kappa-1} \exp(\alpha_A^s A_i + \alpha_{Z_0}^s Z_{0,i} + \alpha_{M_0}^s M_{0,i}) & \text{for } t < 1 \\ \kappa b t^{\kappa-1} \exp(\alpha_A^s A_i + \alpha_{Z_0}^s Z_{0,i} + \alpha_{M_{[t]}}^s M_{[t],i}) & \text{for } 1 \leq t < 4. \end{cases} \quad (8.4)$$

where κ is the shape parameter (set to 1 [exponential] or 2) and b is the scale parameter. Because the values of M are time-varying, event times are generated in waves $w = 1, 2, 3, 4$ corresponding to the intervals between visit times: $[0, 1)$, $[1, 2)$, $[2, 3)$, $[3, 4)$. For brevity, let $\exp(\alpha X_{[t],i})$ represent $\exp(\alpha_A^s A_i + \alpha_{Z_0}^s Z_{0,i} + \alpha_{M_{[t]}}^s M_{[t],i})$. The cumulative hazard, H , is then:

$$H_i(t \mid X_{[t],i}) = \begin{cases} b t^\kappa \exp(\alpha X_{0,i}), & \text{for } 0 \leq t < 1 \\ \sum_{j=0}^{\lfloor t-1 \rfloor} b t^\kappa \exp(\alpha X_{j,i}) + b(t^\kappa - \lfloor t \rfloor) \exp(\alpha X_{\lfloor t \rfloor, i}), & \text{for } 1 \leq t < 4. \end{cases} \quad (8.5)$$

Following the method of Bender et al. [2005], set $H_i(T_i)$ equal to $-\log(u_i)$ where $u_i \sim \text{Uniform}(0, 1)$. Solving for T_i provides an equation for generation of simulated survival times under a Weibull distribution assumption:

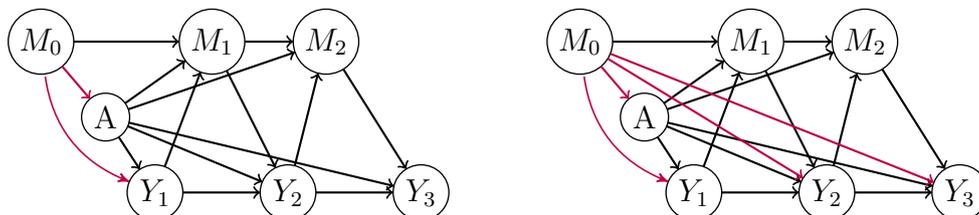
$$T'_{[t],i} = \left[\frac{-\log(u_i)/b - \sum_{j=0}^{\lfloor t-1 \rfloor} \exp(\alpha X_{j,i})}{\exp(\alpha X_{[t],i})} + \lfloor t \rfloor \right]^{1/\kappa} \quad (8.6)$$

After drawing u_i , an event time is calculated for each wave, $T'_{w,i}$ using equation 8.6. If this event time $T'_{w,i}$ is less than the upper time bound for wave w , then the simulated event time T_i is set to $T'_{w,i}$; otherwise, a new event time is generated in the next wave. As before, any individuals still at risk are administratively censored at time $t = 4$.

8.2.3 Baseline mediator affects the exposure

To study the case where the baseline mediator measurement affects the exposure, two scenarios were created using the relationships depicted in Figure 8.3. Note that M_0 is the only baseline covariate for these scenarios. For simplicity, only visit times 1 and 2 are shown in the figure but data was generated for visit times 1, 2 and 3 as in the previous scenarios. In the first scenario, B1, (figure 8.3-left), the baseline mediator measurement affects the exposure and the hazard up to the first visit time. In contrast, in the second scenario, B2, (figure 8.3-right), the baseline mediator measurement affects the exposure and also explicitly affects the hazard at all future times. We can imagine a situation where the level of a biomarker at diagnosis impacts the hazard in addition to the current biomarker measurement by

Figure 8.3: Relationships between the mediator, exposure and outcome in the two scenarios where the baseline mediator measure affects the exposure. In scenario B1 (left), M_0 directly affects Y_1 , the survival time in the first wave. In scenario B2 (right), M_0 directly affects the survival time in all waves. For readability, only two post-exposure mediator measurements are shown.



indicating a measure of disease severity. The difference can be seen by comparing the equations for the conditional hazard used to generate the simulated data. The conditional hazard for scenario B1 is the same as that used to generate data for the baseline scenario (Equation 8.3). For scenario B2, an additional term is added for the contribution of the baseline mediator measurement at times $t \geq 1$:

$$\lambda_i(t \mid A_i, M_{0,i}, M_{[t],i}) = \begin{cases} \alpha_0 + \alpha_A^s A_i + \alpha_{M_0}^s M_{0,i} & \text{for } t < 1 \\ \alpha_0 + \alpha_A^s A_i + \alpha_{M_0}^s M_{0,i} + \alpha_{M_{[t]}}^s M_{[t],i}, & \text{for } 1 \leq t < 4. \end{cases} \quad (8.7)$$

In addition to this revised definition of the hazard, these scenarios also require a modification to the procedure for generating the exposure, A . For these two scenarios only, A_i is generated from a Bernoulli distribution with probability $p_{A|M_0}$ where

$$p_{A|M_0} = \begin{cases} 0.8 & M_0 > \mu_{m0} \\ 0.2 & M_0 \leq \mu_{m0} \end{cases} \quad (8.8)$$

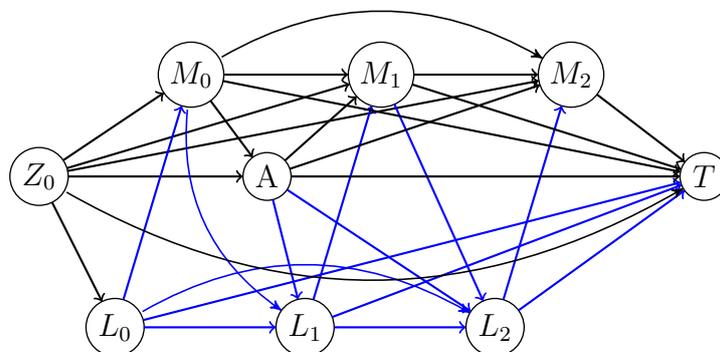
8.2.4 Time-varying confounders present

To include a time-varying confounder, L_t , measured at $t = 1, 2, 3$, a random intercept μ_{l0} and random time slope μ_{l1} are generated where $(\mu_{l0}, \mu_{l1}) \sim MVN(\mu_L, \Sigma_L)$. Following the analysis in Vansteelandt et al. [2019], a causal ordering is assumed such that L_t , measured at the same time as M_t , may influence M_t and M_t may influence L_{t+1} . Figure 8.4 illustrates these relationships. The baseline values L_0 and M_0 are generated first as:

$$L_{0,i} \sim N(\mu = \mu_{l0,i} + \psi_{Z_0}^s Z_{0,i}, \sigma = 1) \quad (8.9)$$

$$M_{0,i} \sim N(\mu = \mu_{m0,i} + \beta_{Z_0}^s Z_{0,i} + \beta_{L_0}^s L_{0,i}, \sigma = 1) \quad (8.10)$$

Figure 8.4: Blue lines show the relationships resulting from the addition of the time-varying confounder, L , to the baseline scenario. As in Vansteelandt et al. [2019], L_k includes survival to the time of visit k . T indicates survival past visit 2 and, for readability, only 2 visit times are depicted.



where $\psi_{Z_0}^s$ is the parameter for the effect of Z_0 on L_0 and $\beta_{L_0}^s$ is the parameter for the effect of L_0 on M_0 . M_0 does not affect L_0 because of the assumed causal ordering.

For $t = 1, 2, 3$, L_t and M_t are successively generated as:

$$L_{t,i} \sim N(\mu = \mu_{l0,i} + \mu_{l1,i}t + \psi_{Z_0}^s Z_{0,i} + \psi_{A_t}^s A_i + \psi_{M_{t-1}}^s M_{t-1,i}, \sigma = 1) \quad (8.11)$$

$$M_{t,i} \sim N(\mu = \mu_{m0,i} + \mu_{m1,i}t + \beta_{Z_0}^s Z_{0,i} + \beta_{A_t}^s A_i + \beta_{L_t}^s L_{t,i}, \sigma = 1) \quad (8.12)$$

with parameters $\psi_{A_t}^s$ for the effect of the exposure on L and $\psi_{M_{t-1}}^s$ for the effect of the previously measured mediator on L . An additional term is added to equation 8.3 to model the effect of the time-varying covariate on survival via a parameter, $\alpha_{L_{[t]}}^s$.

Six scenarios with time-varying confounders were generated. Scenarios L1-L3 allow for investigation of different random slopes used in the generation of L while scenarios L4 and L5 have larger variability in that random slope. Scenario L6 studies the case where the exposure A has a large effect on the values of L . Table 8.2 summarises the differences between each scenario, and complete parameter specifications for all scenarios are found in appendix table F.2.

8.2.5 Infrequent mediator measurements

Finally, for the two scenarios investigating the impact of infrequent mediator measurements, values of the mediator were generated using equation 8.2 for $t = 0, 0.25, 0.5, \dots, 3.75$ (i.e. at intervals of 0.25). Instead of only four values of the mediator, 16 values are generated and event times are generated using each one

Table 8.2: Parameter values used to generate simulated data for the scenarios with time-varying confounders.

Scenario	$E[\mu_{11}]$	$\sigma_{\mu_{11}}$	$A \rightarrow L$	$M \rightarrow L$	$L \rightarrow M$
L1	0.5	2.0	No	No	Yes
L2	5.0	2.0	No	No	Yes
L3	15.0	2.0	No	No	Yes
L4	2.5	5.0	No	No	Yes
L5	2.5	10.0	No	No	Yes
L6	0.5	0.1	Yes	No	No

in waves of length 0.25. However, when estimating the effects, only the mediator measurements at $t = 0, 1, 2, 3$ were made available. Two scenarios were created, F1 and F2, distinguished only by the direction of the effect of the exposure on the mediator.

8.3 Methods

Both the method of Aalen and the method of Vansteelandt were used to estimate effects in the simulated datasets for each scenario. The estimands are specified below. These methods are described in detail in sections 7.3 and 7.4, respectively. All analyses and generation of data were performed using R v4.0.2 [R Core Team, 2020]. The method of Aalen was implemented in the same way for all simulation scenarios using R code provided in the supplementary materials of Aalen et al. [2020]. The method of Vansteelandt allows for different models to be “plugged in” to each step of the estimation. To facilitate comparison of the two methods, an additive hazards model was used to fit the survival model for all scenarios except the multiplicative hazards scenarios where a Cox proportional hazards model was used. The following R packages were used for the computations: `timereg` [Martinsen and Scheike, 2006] (additive hazards model) and `survival` [Therneau and Grambsch, 2000, Therneau, 2015] (Cox regression). A quasi-binomial regression with a logit link was implemented in all simulation scenarios to model the survival predictions.

Because the two methods are based on different assumptions, they are expected to result in biased estimates of the estimands of interest under different scenarios. In the presence of time-varying covariates (scenarios L1-L6), the method of Aalen is misspecified as it is a requirement that all confounder adjustment be accomplished through baseline confounders and that there is no unmeasured

Table 8.3: Listing of all simulation scenarios, the abbreviated name used in the Results section, the percent of simulated individuals experiencing an event prior to time $t=4$, and the table number where full results are provided for that scenario.

Type	Description	Name	Events %	Results table
Baseline				
	Constant effect of M on hazard	Baseline	87-89	G.1
	Increasing effect of M on hazard	-	87-89	-
	Immediate effect of M on hazard	-	90-92	-
	Delayed effect of M on hazard	-	84-87	-
Unmeasured confounding				
	Exposure-outcome, $\alpha_{U_0} > 0$	U1	89-91	G.2
	Exposure-outcome, $\alpha_{U_0} < 0$	U2	78-81	G.3
	Mediator-outcome, $\alpha_{U_0} > 0$	U3	87-89	G.4
	Mediator-outcome, $\alpha_{U_0} < 0$	U4	77-81	G.5
	Exposure-mediator, $\beta_{U_0} > 0$	U5	85-88	G.6
	Exposure-mediator, $\beta_{U_0} < 0$	U6	82-85	G.7
Multiplicative hazards for event times				
	Exponential, fewer events	W1	43-49	G.8
	Exponential, more events	W2	79-83	G.9
	Weibull model	W3	89-91	G.10
Baseline mediator affects exposure				
	M_0 affects the first survival time	B1	77	G.11
	M_0 affects all survival times	B2	81	G.12
Time-varying confounders				
	L with $E[\mu_{l1}] = 0.5$	L1	84-86	G.13
	L with $E[\mu_{l1}] = 5.0$	L2	87-88	G.14
	L with $E[\mu_{l1}] = 15.0$	L3	91-92	G.15
	L with $\sigma_{\mu_{l1}} = 5.0$	L4	85-87	G.16
	L with $\sigma_{\mu_{l1}} = 10.0$	L5	85-86	G.17
	L where A affects L	L6	78-82	G.18
Infrequent mediator measurements				
	$\beta_{A_t}^s > 0$	F1	87-89	G.19
	$\beta_{A_t}^s < 0$	F2	78-87	G.20

mediator-outcome confounding. Further, as the method of Aalen relies on an assumption of additive hazards, it is misspecified for scenarios where event times were generated using multiplicative hazards models (scenarios W1-W3).

8.4 Estimands

The two estimands studied are the direct effect of the exposure on the outcome (not via the mediator) and the indirect effect of the exposure on the outcome via the mediator. The total effect of the exposure on the outcome is also reported as it provides insight into the sources of bias but we would not expect these methods to be used if the goal were solely estimation of the total effect. All three of these causal effects are defined as contrasts between two survival curves. Let $\hat{S}_{A(1),M(0)}(t)$ represent the probability of surviving at time t when exposed but each patient has a mediator value equal to the value it would be if they were not exposed. Values of time-varying confounders, if present, remain unchanged. Define $\hat{S}_{A(1),M(1)}(t)$ ($\hat{S}_{A(0),M(0)}(t)$) as the survival probability at time t when exposed (not exposed) and the mediator value is equal to the value it would be if the person were exposed (not exposed). The effects of interest are represented by the following contrasts:

$$TE(t) = \frac{\hat{S}_{A(1),M(1)}(t)}{\hat{S}_{A(0),M(0)}(t)} \quad (8.13)$$

$$DE(t) = \frac{\hat{S}_{A(1),M(0)}(t)}{\hat{S}_{A(0),M(0)}(t)} \quad (8.14)$$

$$IE(t) = \frac{\hat{S}_{A(1),M(1)}(t)}{\hat{S}_{A(1),M(0)}(t)} \quad (8.15)$$

Our estimands are conditional effect estimates because only survival probabilities conditional on survival to the first mediator measurement are considered. These conditional estimands are used for both methods to facilitate comparison even though the method of Vansteelandt does allow for the estimation of the effects without conditioning on survival to the first mediator measurement. The proportion mediated is not used here because when the total effect estimate is small, the denominator approaches zero and the values become unstable.

The two methods take such different approaches to the analysis of mediation, it is not initially apparent that they are estimating the same quantity in certain circumstances. However, Vansteelandt et al. [2019] explain the equivalence of the two approaches when the patients survive to the first mediator measurement, the event time follows an additive hazard model and the mediator follows a linear regression model. In other words, both methods are estimating path-specific effects under these assumptions.

8.5 Performance measures

Bias in the estimates of TE(t), DE(t) and IE(t) are reported at three time points: the times corresponding to the 20th, 50th (median) and 80th percentile of the event times. Let θ be the true value of the estimand (either total effect, direct effect or indirect effect) and $\hat{\theta}$ be the estimated value. Bias is computed as the average difference between the true and estimated values.

$$\text{Bias} = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \theta) \quad (8.16)$$

where i refers to the i^{th} simulated dataset and n_{sim} denotes the number of simulated data sets. The corresponding Monte Carlo standard error (MCSE) is:

$$\text{MCSE} = \sqrt{\frac{\frac{1}{n_{sim}-1} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \theta)^2}{n_{sim}}} \quad (8.17)$$

[Morris et al., 2019].

In addition to bias, the empirical standard error, a measure of efficiency, is reported for the baseline scenario. Empirical standard error is calculated as:

$$\text{Empirical SE} = \sqrt{\frac{1}{n_{sim}-1} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \bar{\theta})^2} \quad (8.18)$$

where $\bar{\theta}$ is the mean of $\hat{\theta}_i$. The efficiency of the method of Aalen relative to the method of Vansteelandt is also reported. The percentage increase in efficiency of one method over the other is the squared ratio of their empirical standard errors [Morris et al., 2019]. Note that if a method is biased, its empirical standard error may be impacted by this bias; estimates biased towards zero will have a smaller empirical standard error simply because of their smaller magnitude. For this reason, the focus is on empirical standard error and relative efficiency in the baseline scenario where both methods are unbiased by design.

As bias is the primary performance measure for this study, the calculation of number of simulated datasets is based on bias. For this study, we require that the MCSE be below 0.005. In a small simulation run, the observed variance of $\hat{\theta}$ was less than or equal to 0.015. This suggests the minimum number of simulated datasets needed is 600 based on the relationship:

$$n_{sim} = \frac{\text{Var}(\hat{\theta})}{\text{MCSE}_{\text{bias}}^2} \quad (8.19)$$

To allow for cases where the variance may be slightly higher, $n_{sim} = 1,000$ datasets were generated with $n_{obs} = 2,000$ individuals per dataset for each scenario investigated.

8.5.1 Generation of the truth

To compute the bias, the true value of the estimand is required. These values were estimated using a large ($n=3,500,000$) simulated dataset. Following the approach of [Keogh et al., 2021], data were generated for each individual under four conditions: exposed with mediator values set to the values they would have taken if exposed, $A(1), M(1)$; not exposed with mediator values set to the values they would have taken if not exposed, $A(0), M(0)$; exposed with mediator values set to the values they would have taken if not exposed $A(1), M(0)$; and not exposed with mediator values set to the values they would have taken if exposed, $A(0), M(1)$. The data was generated using the relationships described in section 8.2 except that no variables affect the exposure. Rather, the aim was to replicate a randomised trial setting, where the exposure would be randomised. After event times are generated, individuals with event times occurring before the first mediator measurement are removed. For each of the four cases, the probability of survival was computed at three evaluation times, corresponding to the 20th, 50th and 80th percentiles of event occurrence in the $A(1), M(1)$ group. Survival probabilities at each time point are simply the proportion still at risk at that time point as there is only administrative censoring at time $t = 4$. Denote the true survival probability at time t as $\tilde{S}_{A(a),M(m)}(t)$ where a and m are 0/1 for unexposed/exposed and $M(m)$ refers to mediator values as if the exposure was set to level m . The true value of the TE, DE and IE at each time point can be calculated as:

$$\theta_{TE}(t) = S_{A(1),M(1)} / S_{A(0),M(0)} \quad (8.20)$$

$$\theta_{DE}(t) = S_{A(1),M(0)} / S_{A(0),M(0)} \quad (8.21)$$

$$\theta_{IE}(t) = S_{A(1),M(1)} / S_{A(1),M(0)} \quad (8.22)$$

Figure 8.5 shows plots of the three survival curves $S_{A(1),M(1)}(t)$, $S_{A(0),M(0)}(t)$ and $S_{A(1),M(0)}(t)$ representing the true values for the baseline scenario under the three sub-scenarios. In the NoDE sub-scenario, $S_{A(1),M(0)}(t) = S_{A(0),M(0)}(t)$ because without a direct effect, the total effect equals the indirect effect. Similarly, in the NoIE sub-scenario, $S_{A(1),M(0)}(t) = S_{A(1),M(1)}(t)$ because none of the total effect goes through the mediator. In the DE+IE sub-scenario, the dotted black line representing the survival curve $S_{A(1),M(0)}(t)$ lies in between the other two curves indicating the presence of both a direct and an indirect effect.

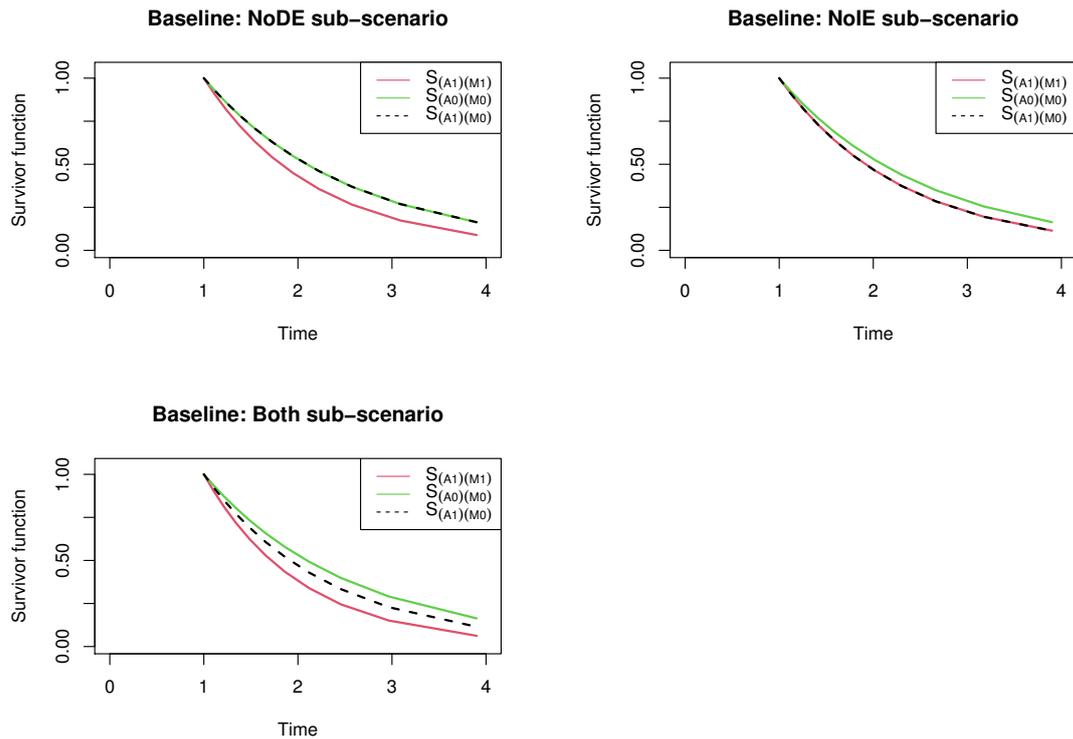


Figure 8.5: Survival curves based on a large simulated data set (the “truth data”) used to generate true values of the estimands. The NoDE sub-scenario (upper-left) is the case where the total effect equals the indirect effect. The NoIE sub-scenario (upper-right) corresponds to the case where none of the effect of the exposure on the outcome goes through the mediator. The DE+IE sub-scenario (bottom) includes both a direct and an indirect effect.

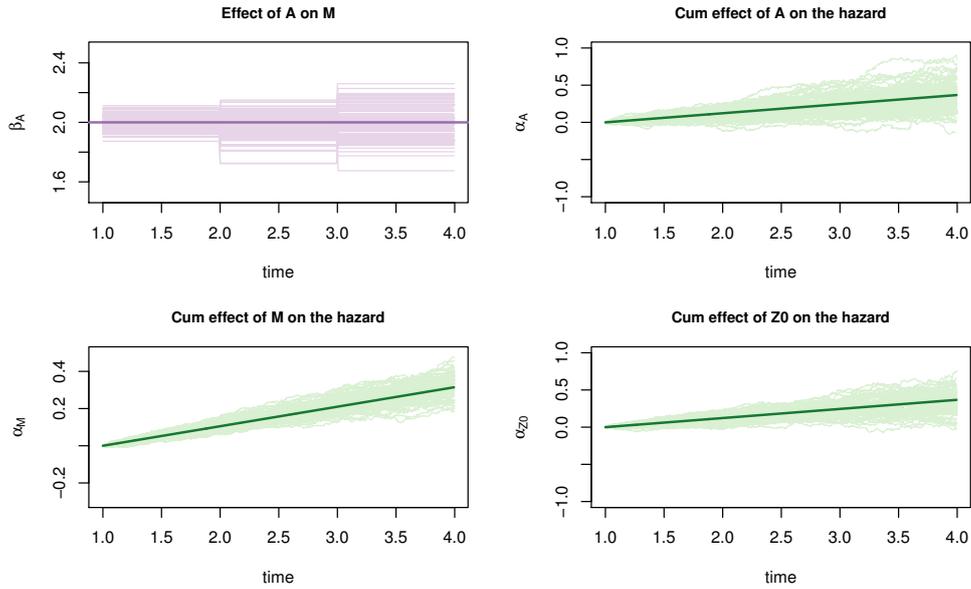


Figure 8.6: Diagnostic plots used to verify generated truth data. The darker solid line is the value of the parameter over time if there were no random variation in the data generation. Each set of lighter lines represents the values of the same parameters across simulation runs. β_A , the effect of the exposure on the mediator (top left in purple) is a parameter of the mediator model. The other three graphs in green show α_A , α_M and α_{Z_0} , parameters of the hazard model from the method of Aalen.

When approximating the true values of the estimands using a large simulated data set (the “truth data”), I recommend constructing plots and thoroughly exploring this data to ensure the characteristics of the data are as expected. For example, Figure 8.6 shows a set of four diagnostic plots I created to verify the truth data using the method of Aalen. Each plot corresponds to one estimated parameter: β_A^s from equation 8.2 (top-left), α_A^s (top-right), $\alpha_{M_{|t_j}}^s$ (bottom-left), and $\alpha_{Z_0}^s$ (bottom-right) from equation 8.3. The thicker solid line shows the parameter value used to generate the truth data and the lighter-colour thin lines each represent the parameter estimate from one simulation run. For correctly generated truth data, the true parameter value should equal the average of the parameter estimates from the simulations. I also suggest verifying that the link between any baseline covariates and the exposure has been broken in the data generation process so that the exposure is randomised in the truth data. R code for generating truth data and simulated data as described in this chapter is available from <https://github.com/KamTan/MediationSimulation>.

Table 8.4: Summary of performance of the methods of Vansteelandt and Aalen for mediation analysis using simulated scenarios. A “Yes” indicates that the absolute value of the bias exceeded 0.01 in one or more of the scenarios/sub-scenarios for that category at the time when 50% of events had occurred. Across scenarios, the MCSE of the bias was consistently <0.005 .

Type	Presence of bias in any scenarios					
	Method of Vansteelandt			Method of Aalen		
	TE	DE	IE	TE	DE	IE
Baseline	No	No	No	No	No	No
Unmeasured confounding	Yes	Yes	Yes	Yes	Yes	Yes
Multiplicative hazards for event times	No*	No*	No*	No	Yes	Yes
Baseline mediator affects exposure	No	No	No	No	No	No
Time-varying confounders	No	No	No	No	Yes	Yes
Infrequent mediator measurements	No	Yes	Yes	No	Yes	Yes

* when implemented with Cox proportional hazards survival model

8.6 Results

8.6.1 Overview and baseline scenario

Across scenarios, the two methods performed differently in terms of bias in the estimates of TE, DE and IE. Table 8.4 summarises the bias found in each simulation scenario using each method. In this table, a method was considered to show some bias if the magnitude of the bias exceeded 0.01. Unmeasured confounding and infrequent mediator measurements led to bias in both methods while the presence of time-varying covariates was associated with bias in the method of Aalen. These results are described in detail in the subsequent sections. Note that the Vansteelandt method was implemented with an Aalen additive hazards model (Vansteelandt^{add}) in all scenarios except those exploring event times generated via a multiplicative hazards model where it was implemented with a Cox survival model (Vansteelandt^{cox}).

Under the baseline scenario, the estimated TE, DE and IE are approximately unbiased for all three sub-scenarios for both the method of Aalen and the method of Vansteelandt. The MCSE of the bias estimate was < 0.005 for both methods for all estimates. Detailed results tables from all scenarios are available in Appendix G and results from the baseline scenario are shown in Table G.1. The baseline scenario is consistent with the assumptions of both methods and, therefore, we

Table 8.5: The empirical standard error and relative efficiency for the baseline scenario simulation analyses. Empirical standard error is a measure of efficiency of the effect estimators. Precision is reported for the method of Aalen relative to the method of Vansteelandt; values greater than 1 indicate greater precision by the method of Aalen. Results are shown at times corresponding to the 20th, 50th and 80th percentile of event occurrence.

Events	Aalen			Vansteelandt ^{add}			Rel. Precision		
	TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>									
20%	0.02	0.03	0.02	0.02	0.03	0.03	0.99	1.60	2.62
50%	0.03	0.05	0.03	0.03	0.06	0.05	0.98	1.75	2.94
80%	0.04	0.08	0.04	0.04	0.12	0.09	0.98	2.14	4.31
<i>No Direct Effect</i>									
20%	0.02	0.03	0.02	0.02	0.03	0.03	0.99	1.47	2.25
50%	0.03	0.05	0.03	0.03	0.06	0.05	0.99	1.55	2.43
80%	0.05	0.09	0.05	0.05	0.13	0.08	1.00	1.92	3.31
<i>No Indirect Effect</i>									
20%	0.02	0.02	0.00	0.02	0.02	0.00	0.99	0.98	0.87
50%	0.03	0.03	0.00	0.03	0.03	0.00	1.00	0.98	0.96
80%	0.06	0.06	0.01	0.06	0.06	0.01	1.00	0.97	0.99

expected good performance. Table 8.5 summarises the empirical standard error, a measure of efficiency, for the baseline sub-scenarios. Overall, the empirical standard error increases over the study period for both methods. In the DE+IE and NoDE sub-scenarios, the empirical standard error of the method of Aalen is less than the method of Vansteelandt in the estimates of DE and IE meaning that the relative efficiency is greater than 1. In the NoIE scenario, the efficiency of the two methods is approximately equal.

Figure 8.7 shows the effect estimates from the 1,000 simulation runs of the Baseline-DE+IE scenario using the method of Vansteelandt. The effects are defined using the contrasts from section 8.4. Each gray line corresponds to one simulated dataset and the thick black line plots the average effect estimate. The variability of the direct and indirect effect estimates is much greater than that of the total effect estimate. Effect estimates begin at time $t=1$, the time of the first mediator measurement.

As both methods also allow for a time-varying effect of the mediator on the hazard, three versions of the baseline scenario were also examined where the ef-

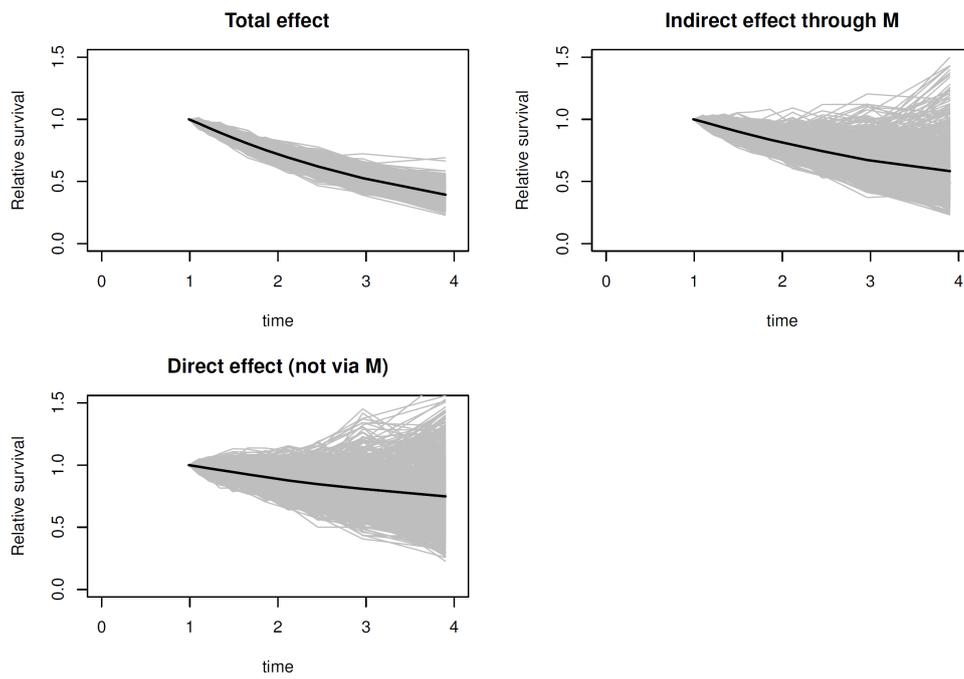


Figure 8.7: Effect estimates from the method of Vansteelandt for the Baseline-DE+IE scenario. Results from each simulated dataset are plotted in gray and the thick black line shows the average across the datasets.

fect of the mediator on the hazard was: increasing over time, only non-zero at $t = 1$ (immediate effect), only non-zero at $t = 2, 3$ (delayed effect). Details of the parameters used in all scenarios can be found in Appendix F. Again, both methods produced unbiased estimates at all time points in all sub-scenarios indicating that both methods reliably produce estimates of the TE, DE and IE for correctly specified models.

8.6.2 Unmeasured confounding

In the scenarios with unmeasured confounding, both methods produced biased results. Table 8.6 summarises the absolute bias seen in each of these six scenarios for the DE+IE sub-scenario at the time when 50% of events had occurred. Throughout these scenarios, both analysis methods were biased in the same direction and by approximately the same magnitude. When an unmeasured confounder U_0 of the exposure and the outcome was introduced, bias was seen in both the estimate of the direct effect and the total effect. The direction of this bias could be either negative or positive depending on the direction of the effect of U_0 on the hazard. (Tables G.2 and G.3). When mediator-outcome confounding is not controlled, bias is seen in the estimates of the direct and indirect effect but not the total effect for both methods. The NoIE scenario estimates were unbiased. Again, direction of the bias depends on the direction of the effect of the unmeasured confounder on the hazard. (Tables G.4 and G.5). Finally, in the presence of exposure-mediator confounding, bias is seen in the indirect effect and the total effect. Because the total effect is the sum of the direct and indirect effects, the bias in the total effect is of the same magnitude and direction as the bias in the indirect effect. Similar to the previous two unmeasured confounding scenarios, the direction of the bias depends on the direction of the effect of the unmeasured confounder on the mediator. (Tables G.6 and G.7)

8.6.3 Multiplicative hazards

When the time to event outcomes were generated using a Weibull hazard model with shape parameter equal to 1 (equivalent to exponentially distributed survival times), the method of Vansteelandt implemented with a Cox proportional hazards survival model produced unbiased results, regardless of how many events occurred (Tables G.8 and G.9). The method of Aalen also produced approximately unbiased results when events occurred in less than half of the individuals. When events occurred in 79-83% of individuals, the NoDE and NoIE sub-scenarios continued to show no bias but a small bias (2% on the estimate of DE and IE when 50% of events had occurred) was seen in the DE+IE sub-scenario. When the shape parameter was set to 2 for generating event times and 89-91% of individuals had

Table 8.6: The true effect and estimated absolute bias for the six simulation scenarios with unmeasured confounding when both a direct and an indirect effect were present. Results are shown at the time 50% of events had occurred. The estimated absolute bias is the difference between the estimated effect and the true effect. Full results for each of these scenarios can be found in appendix tables G.2 - G.7.

Name	Truth			Bias: Aalen			Bias: Vansteelandt ^{add}		
	TE	DE	IE	TE	DE	IE	TE	DE	IE
U1	0.81	0.93	0.87	0.06	0.07	0.00	0.06	0.07	0.00
U2	0.78	0.91	0.86	-0.06	-0.06	0.00	-0.05	-0.06	0.00
U3	0.81	0.92	0.87	0.00	-0.02	0.02	0.00	-0.02	0.02
U4	0.78	0.91	0.85	0.00	0.03	-0.02	0.00	0.03	-0.02
U5	0.80	0.92	0.87	0.03	0.00	0.03	0.03	0.00	0.03
U6	0.79	0.92	0.86	-0.03	0.00	-0.03	-0.02	0.01	-0.03

events, both methods returned approximately unbiased effect estimates up to the time when 50% of events had occurred in all three sub-scenarios. (Table G.10). Some bias in the estimates of DE and IE was seen from both methods in the NoDE sub-scenario at the time corresponding to 80% event occurrence. At the same time point in the DE+IE sub-scenario, the method of Vansteelandt returned biased DE and IE effect estimates. Overall, the bias seen was small and did not exceed 0.05 in magnitude.

8.6.4 Baseline mediator affects exposure

For the two scenarios investigating the impact of a baseline mediator measurement that affects the exposure, B1 and B2, higher values of M_0 led to a higher chance of exposure. Effect estimation was done for both scenarios, with and without adjustment for M_0 as a baseline confounder for the DE+IE sub-scenario. Table G.11 contains complete results for the analyses of scenario B1 where M_0 directly affects the hazard prior to the first visit and the exposure. Both methods were approximately unbiased when M_0 was adjusted for; when the analyses were not adjusted for M_0 as a baseline confounder, both methods produced somewhat biased estimates of the total effect and indirect effect when 80% of events had occurred. At earlier times, when 50% or fewer events had occurred, both methods were approximately unbiased. For scenario B2 where M_0 directly affected the exposure and the hazard at all times, when adjustment was made for M_0 , only a small amount of bias was seen in the estimate of the direct effect at the time when 80% of events had occurred. (Table G.12). However, when no control was made for

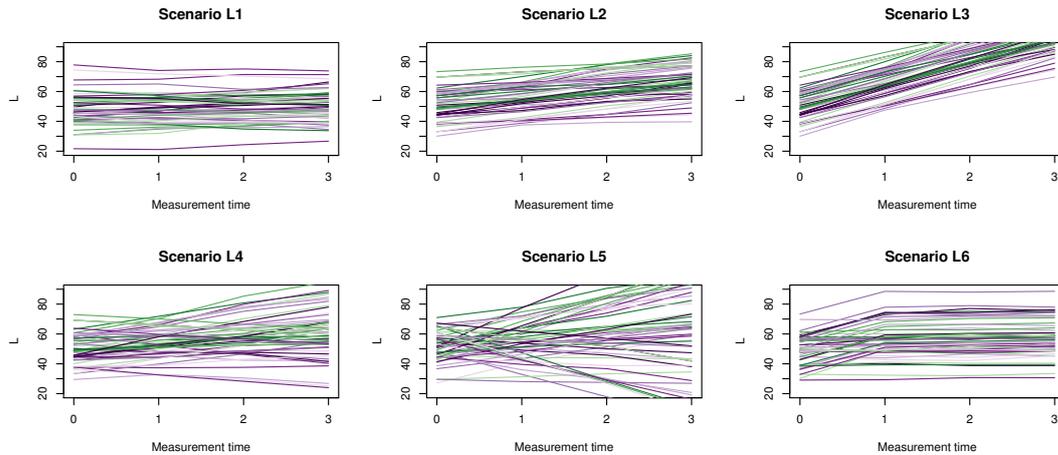


Figure 8.8: Values of the time-varying confounder L for 60 simulated individuals in each of the 6 scenarios investigated. Different colours are used to help distinguish individual trajectories. Top row: mean random slope used to generate L is 0.5 (left), 5.0 (centre), 15.0 (right). Bottom row: standard deviation of the random slope used to generate L is 5.0 (left) or 10.0 (centre) and the scenario when the exposure affects L (right).

M_0 , both methods returned biased estimates for all three effects for multiple time points.

8.6.5 Time-varying confounders present

Six scenarios were created to explore the impact of a time-varying confounder L on the estimation of direct and indirect effects. The scenarios are distinguished by (1) the slope of L , (2) the amount of variation in L over time, and (3) the impact of A on L as described in table 8.2. Figure 8.8 shows the values of L for 60 simulated individuals for each scenario. The top row illustrates how the values of L change as the mean random slope used to generate L is increased from 0.5 to 5.0 to 15.0 (left to right). The bottom row shows L when the standard deviation of the random slope is doubled (left and centre) and when there is a strong effect of the exposure on L (right).

When the time-varying confounder is relatively stable over time and has a constant effect on the mediator as in scenario L1, both methods produce approximately unbiased effect estimates. For scenarios L2 and L3, the values of L taken at later measurement times are, on average, higher than L_0 and there is small variability in the slope of L over time. Effect estimates from the method of Vansteelandt continue to be approximately unbiased as are the effect estimates from the method of

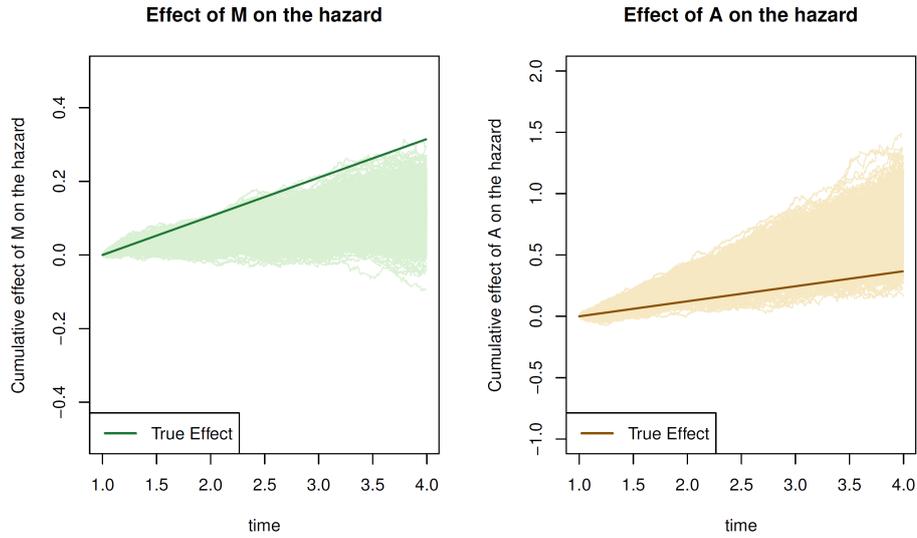


Figure 8.9: Results from 1,000 simulation datasets of scenario F1-DE+IE using the method of Aalen to estimate the parameters from equation 7.10. On the left, the true value of α_{M_t} is shown as a dark green line and estimated values of this parameter in lighter green. On the right, the true value of α_{A_t} is shown as a brown line with estimated values in beige.

Aalen. As the values of L varied more over time, as in scenarios L4 and L5, the method of Vansteelandt remains unbiased for the effect estimates but bias is seen in the effect estimates from the method of Aalen. There is more bias overall and more bias earlier in time in scenario L5 than in scenario L4; for example, in the DE+IE sub-scenarios, the bias in the direct (indirect) effect estimates when 50% of the events had occurred in scenario L4 was 0.01 (-0.01) versus 0.02 (-0.02) in scenario L5. At the time corresponding to the 80th percentile for event occurrence, the bias was 0.03 (-0.02) in scenario L4 versus 0.07 (-0.07) in scenario L5. Results were similar for the NoDE sub-scenario. In scenario L6, when the exposure has a relatively large impact on the level of L_1, L_2, L_3 , there is bias in the effect estimates of the DE+IE and NoDE sub-scenarios from the method of Aalen but the method of Vansteelandt effect estimates were approximately unbiased. Detailed results for each of these scenarios can be found in Tables G.13 - G.18.

8.6.6 Infrequent mediator measurements

The impact of an infrequently measured mediator was investigated in two scenarios which differed only in the direction of the effect of A on M . Mediator measurements

were generated four times per annum and used to calculate times-to-event. In the mediation analysis, however, only the measurements taken at the beginning of each year were used. No bias was seen with either method in the two NoIE sub-scenarios or in the estimation of total effect in any sub-scenario. In scenario F1 where A positively affects M , both methods yielded biased results for the direct and indirect effects in the DE+IE and NoDE scenarios (Table G.19). By the time 50% of events had occurred, the methods were over-estimating the IE by 7% and under-estimating the DE by approximately the same amount when both a DE and an IE were present. This bias more than doubled by the time 80% of events had occurred. Figure 8.9 displays results from the dynamic path analysis (method of Aalen) for scenario F1-DE+IE. Nearly all of the 1,000 simulation runs under-estimated the effect of the mediator on the hazard (α_{M_t} in equation 7.10). In scenario F2-DE+IE where A negatively affects M , the bias in the estimates of DE and IE was even greater (Table G.20). At the time when 50% of events had occurred, both methods underestimated the IE by 11% and overestimated the DE by 13%. The bias was greater still in the NoDE sub-scenarios.

8.7 Discussion

In this simulation study, I have assessed the performance of the two selected mediation methods under both conditions where performance was expected to be unbiased and conditions with misspecification, confounding or data constraints typical of observational datasets. Both methods produced approximately unbiased results under the scenarios consistent with their stated assumptions. The presence of uncontrolled confounding is problematic for any mediation method but if measurements are available for time-varying confounders, the method of Vansteelandt is equipped to accommodate them. No method can overcome data that is inappropriately or incompletely measured for analysis of the question at hand and thorough data exploration and discussions with subject matter experts is recommended prior to beginning a mediation analysis.

Presence of time-varying confounders L in a given setting is a key reason that one might select the method of Vansteelandt over the method of Aalen for a mediation analysis. For all scenarios with a time-varying covariate, the method of Vansteelandt returned unbiased effect estimates. In the method of Aalen, there is no mechanism to adjust for values of L post-exposure, even when those values are known. It is possible to adjust for the baseline value L_0 and when the values of L_1, L_2, L_3 are not too different from L_0 , this baseline adjustment may be sufficient to obtain results where the bias is fairly minimal. In our simulations, as L_0 became less representative of the future values of L , either because there was a greater variance in the random slope used to generate L or because L was affected

by the exposure, the bias in the estimates from the method of Aalen increased. Because L did not affect A in the scenarios considered, there is no exposure-outcome confounding and, therefore, the estimates of the total effect remain unbiased.

Substantial bias was found in the estimates of DE and IE in the scenarios where the mediator is measured infrequently. This scenario is likely to occur in many observational datasets where the mediator is a continuous biomarker measured intermittently at scheduled visits. Such a mediator affects survival in continuous time but the analysis is only able to incorporate discrete information. This is a type of measurement error and the result is an attenuation of the estimated effect. Here, because it is the mediator that is subject to measurement error, the indirect effect is attenuated and this bias accumulates over time. Vansteelandt et al. [2019] note that an analysis performed with a single mediator measurement may only capture a portion of the indirect effect. This supposition was confirmed for dynamic path analysis in the simulation study of Strohmaier et al. [2015]. Interestingly, they found that as more frequent mediator measurements were taken, estimation of IE and DE did not necessarily improve. Rather, performance depended more on sampling at times that were better representative of the average total effect. Intuitively, this makes sense as we need the timing of mediator measurements to correspond with the timing of the biological effects. In this simulation, it was found that the direct effect is over-estimated to compensate for the under-estimated indirect effect. It is possible that a better mediator model may mitigate the bias seen here and I suggest that as a direction for future research. In the previous chapter's mediation analysis of CFRD, only annual measurements were available for the mediator. For lung function, there can be substantial variability in an individual's FEV1% from month to month [Taylor-Robinson et al., 2012] but, on average, people over 25 experience a decline in FEV1% of 1.5% per year [Konstan et al., 2012]. It is possible that short-term fluctuations or a period of rapid decline in FEV1% could impact survival in a way that is not captured by annual measurements alone. Consistent with the results of this simulation study, we would expect such a bias to reduce the indirect effect estimate.

In an observational setting, investigators may find that the baseline level of the mediator differs between the exposed and the unexposed. In this case, adjustment for M_0 is required for conditional randomisation of the exposure. I had questioned whether adjustment for a baseline mediator measurement that is highly correlated to future mediator measurements could result in a biased estimate of the indirect effect. In this simulation study, no evidence was found of this; on the contrary, failure to adjust for M_0 is what leads to bias. In the context of the CFRD analysis with lung function as the mediator in chapter 7, the baseline measurement of FEV1% was adjusted for to achieve conditional randomisation of the exposure. In an identical analysis using the method of Aalen without adjustment for M_0 , the

Table 8.7: Estimates of indirect effect and proportion mediated with and without adjustment of the baseline mediator measurement M_0 for the analysis of the effect of CFRD on survival with lung function as a mediator.

Time	Indirect Effect Estimate				Prop. Mediated Estimate			
	1	2	3	4	1	2	3	4
With M_0 adj	1.00	1.00	1.00	0.99	-0.01	0.04	0.06	0.08
Without M_0 adj	1.00	0.99	0.98	0.97	-0.32	0.20	0.23	0.26

estimated proportion mediated was three or more times greater than estimated in the analysis with adjustment for M_0 . (Table 8.7). Because adjusting or not adjusting for the baseline mediator measurement can lead to very different conclusions, careful consideration is necessary when making this decision.

I investigated the ability of the two approaches to estimate effects when different mechanisms for generating time-to-event outcomes were used. The method of Aalen assumes an additive hazards model and the formulas for effect estimates rely on that assumption so times-to-event generated using a multiplicative hazards model may not meet this requirement. Some bias was seen in the estimates of direct and indirect effect in this case but, overall, the method was robust to this mismatch. Bias was more pronounced at later times and when the percent of individuals who experienced an event was greater. The method of Vansteelandt when implemented with a Cox model produced unbiased effect estimates in the same scenarios with one exception at a time when 80% of events had occurred. Because the theory underlying the method of Vansteelandt is non-parametric, the procedure may be implemented with any valid model at each step [Vansteelandt et al., 2019]. I chose an additive hazards model and a Cox regression model for the survival models but more complex structures could be supported by instead using a flexible splines model or other parametric survival model, e.g. Gompertz, log-logistic.

With the scenarios exploring unmeasured confounding, unbiased performance was not expected from either method. Both methods explicitly require control for exposure-outcome confounding, mediator-outcome confounding and exposure-mediator confounding. Even in simpler analyses not considering mediation, control of exposure-outcome confounding is required for an unbiased estimate of the total effect. In all six scenarios considered, if the unmeasured confounder was treated as a measured confounder and included in the analyses, the bias disappeared (results not shown). The key takeaway from exploration of these unmeasured confounding scenarios is that the bias can occur on any or all of the estimated effects and the direction of the bias is difficult to predict when the nature or even existence of the confounder is unknown. A number of sensitivity analysis methods have

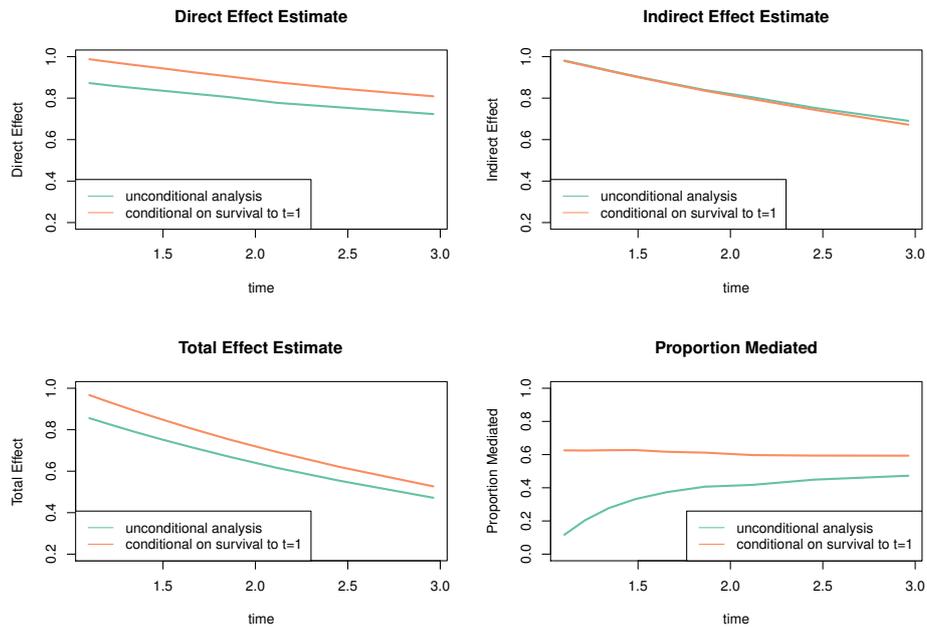


Figure 8.10: Comparison of effect estimates for an analysis restricted to individuals who survive to the first mediator measurement (orange) versus an analysis including all individuals (green). Estimates are plotted from the time at which 10% of events had occurred in the conditional analysis ($t = 1.1$) to the time at which 90% of events had occurred ($t = 2.96$).

been proposed to test robustness to unmeasured confounding. Because mediation analysis is frequently applied to data where the treatment was randomised, most focus on unmeasured mediator-outcome confounding. However, in observational datasets where the exposure is not likely to be randomised, all possible sources of confounding should be considered. VanderWeele [2015] (chapter 3) provide an excellent overview of several techniques.

One apparent limitation of the method of Aalen is that all individuals are assumed to have survived up to the time of the first mediator measurement. This changes the interpretation of the estimates to the effect conditional on survival to time $t = 1$. A natural question is what impact does this conditioning have on the analysis and how are the results different because of this requirement? Using the baseline scenario with both a direct and an indirect effect present, the results of two analyses, one with and one without individuals with events prior to the first mediator measurement were compared using the method of Vansteelandt. Figure 8.10 shows plots of the effect estimates over time for both analyses. Immediately after exposure and prior to the first mediator measurement, there is a direct effect

of A on Y in the analysis where events prior to $t = 1$ are included. At $t = 1$, the total effect and direct effect estimates are 1.0 for the conditional analysis (orange) but are different than 1.0 in the unconditional analysis (green). In contrast, because the indirect effect is defined as the effect that goes through M , it cannot be estimated until after the first mediator measurement. Therefore, the indirect effect is zero prior to $t = 1$ in both analyses. Based on this comparison, I conclude that performing an analysis conditional on survival to the first mediator measurement does not change the indirect effect estimate but the estimates of direct and total effect are impacted. A conditional analysis also results in a different estimate of the proportion mediated compared to an unconditional one because proportion mediated relies not only on the estimate of IE but also on the estimate of TE. From Figure 8.10, the estimated mediated proportion is quite different depending on whether the analysis was conditional on survival to $t = 1$ or not. In our baseline scenario at time $t = 1.5$, for example, one analysis says more than half of the effect is mediated while the other says that less than half is mediated. This subtlety suggests care needs to be taken with interpretation of this measure.

In the mediation analysis of the previous chapter and the simulation study of this chapter, I endeavoured to compare and contrast the method of Aalen and the method of Vansteelandt for mediation analysis with a survival outcome and a repeatedly-measured mediator. Both methods produce unbiased effect estimates in scenarios where they were designed to perform well and both were found to be robust to some misspecification. The method of Aalen is simple to implement and fast computationally but the requirement that the treatment be able to be split into two distinct biological processes may be limiting. It produces analyses that are conditional on survival to the first measurement time and is tied to particular parametric models for the hazard and mediator. In contrast, the method of Vansteelandt is more flexible in allowing different model types to be used but the price is a more computationally intensive procedure that requires more data as the number of visit times and covariates increases. It is also less efficient than the method of Aalen. The key advantage of the method of Vansteelandt is the ability to accommodate time-varying covariates. In this simulation study, it was found that when time-varying covariates are present but not included in the analysis, effect estimates can be biased. With both methods, many assumptions are required to make causal statements and care must be taken with the interpretation.

Discussion

Chapter 9

Discussion

9.1 Key Findings

The main aim of this thesis was to investigate and develop statistical methods for the analysis of longitudinal observational data with time-to-event outcomes. Methods for dynamic survival prediction were studied and a new technique was described that allows the power of machine learning algorithms to be combined with traditional statistical techniques in search of the best predictive performance. In addition to prediction, methods for understanding the mechanisms of a treatment or exposure effect were assessed. The focus here was on two recently proposed methods for causal mediation analysis with a survival outcome and time-varying mediator. A secondary aim of this thesis was to investigate how to communicate survival predictions to patients and clinicians in an accessible and emotionally sensitive way. These analyses were motivated by open questions in CF and the methods developed were applied and assessed using UK CF Registry data throughout.

9.1.1 Dynamic prediction of survival

In Part I of this thesis, a new method was described that combines landmarking with a machine learning ensemble. Although machine learning methods are increasingly used in prediction, their use in dynamic survival prediction is less common. This is in part because not all machine learning algorithms have been adapted for survival endpoints and/or longitudinal data. I showed how a discrete-time survival framework allows the use of any algorithm or model designed to predict the probability of a binary outcome and how this approach can be used in combination a machine learning ensemble, the Super Learner. A loss function defined on the v -year survival probability was also proposed. This technique was compared to joint modelling and landmarking with Cox regression on three

dimensions: calibration, discrimination and overall predictive performance using both cross-validation and a test dataset. In this study, using UK CF Registry data, performance of the Super Learner landmarking technique was found to be approximately the same as Cox landmarking but outperformed joint modelling. Using 10-fold cross-validation, the Super Learner landmarking method achieved 29%-52% reduction in prediction error and a C-index between 0.85 and 0.90 for 5-year survival prediction from landmark ages 20-40. One cannot know in advance which model or algorithm will yield the best predictive performance and the advantage of Super Learner landmarking is that it can combine numerous fully-parametric, semi-parametric and non-parametric techniques to create predictions based on the best performing combination of those techniques. The price for this flexibility is computational complexity. I showed in a simulation study that in the presence of interactions and/or non-linearities in the data that the analyst is unaware of, Super Learner landmarking outperformed joint modelling and Cox landmarking without interaction terms. Code for both dynamic survival prediction and the simulation was provided. The method described could in theory be applied to any dataset but when patient visits are irregular, additional modelling may be required. The conclusion of this work is that machine learning algorithms need not be at odds with traditional statistical methods; rather, they are another tool for us to use in understanding data and making predictions.

9.1.2 Communication of survival predictions

In Part II of this thesis, the focus shifted from methods for making survival predictions to methods for their presentation and communication. Having developed a method for survival prediction in Part I, it is important to consider how the results from such a prediction model could be used in practice. Unfortunately, many published clinical prediction models are never used to aid medical decision-making or inform clinicians and their patients about prognosis [Collins, 2021]. After reviewing best practices in graphical display of risk information, a prototype life expectancy presentation was trialled via interviews with PWCFs and HCPs. The presentation included different graphical formats and labelling, varying levels of personalisation of the information and consideration of how to incorporate uncertainty. Feelings were strong on the topic of access to this information with some believing they had a right to access any information or models germane to their situation. Others strongly believed that life expectancy information had the potential to be damaging or even dangerous to some vulnerable PWCFs. It was clear that there is a need for more tailored predictions of life expectancy in CF as disease progression can vary greatly and many felt that current information did not apply to them or their patient's specific situation. At the same time, we must be aware of the limits of our ability to make individualised predictions based on statistical modelling

and to adequately explain the predictions that are given. Based on this study and prior work in Keogh et al. [2019a], it is my opinion that this information should be made available to those who want it whether it be through their care team or via a website with possible access restrictions. With respect for the emotional nature of this topic and for those who do not want to know their life expectancy, there are many PWCFs who do want to know and who will be helped in their plans for the future by having it.

9.1.3 Mediation analysis for survival outcomes

Part III of this thesis studied methods for mediation analysis with a survival outcome and time-varying mediator using the motivating example of the extent to which the association between CFRD and survival is mediated through the effects of CFRD on several key predictors of mortality. Methods for mediation analysis that are capable of accommodating both a survival outcome and a repeatedly-measured mediator are relatively new and I focused on the methods of Vansteelandt et al. [2019] and Aalen et al. [2020]. Following descriptive work confirming the association of CFRD with increased risk of mortality, three hypothetical mechanisms of that effect were considered. Interestingly, lung function, BMI and respiratory infections were not significant mediators of the effect of CFRD on survival. To aid in the interpretation and understanding of this result, the two mediation methods were compared in a simulation study using 23 scenarios, with 3 sub-scenarios each. There is no work comparing these two recently published methods or offering advice on how to choose between them. One of the key differences between the methods is the ability to accommodate time-varying confounders and it was found that significant bias can be introduced if time-varying confounders exist but are not accounted for. Biased effect estimates were also found when mediators that affect survival are not measured sufficiently frequently to capture changes. Either or both of these issues could have led to biased results in the CFRD analysis and we suspect infrequent mediator measurements are common in observational datasets, especially registries.

9.2 Future Work

9.2.1 Dynamic prediction of survival

The work in this thesis highlights many different opportunities for future work, both in terms of the methodology and relating to the substantive applications to the CF data. First, in the area of dynamic survival prediction, the Super Learner landmarking approach should be tested in other datasets to better understand the

types of situation where it can improve predictive performance. Such a case was illustrated via simulation but the interaction was extreme. When using physiological measurements, it is perhaps unlikely that there would be major interactions without some a priori knowledge of them. Although, in any setting with a large number of predictors, it is plausible that there are interactions that have not been contemplated. Variable selection for data with a large number of predictors is an active area of research and it may be fruitful to better understand how to combine variable selection methods with ensemble learning for survival prediction. This is not straightforward for an ensemble as each component algorithm may use a distinct subset of predictors. Another fascinating area with respect to the ensemble involves error diversity. In this study, many different types of algorithms and models were included with the goal of achieving error diversity but our approach, based on considering the features of different methods, did not use a formal procedure. It would be helpful to have answers to questions such as: How many different learners are needed for error diversity? How many different hyperparameter combinations are needed per algorithm? What are the best types of algorithms to use in combination to achieve diversity? In light of computational time, how can we minimise the number of algorithms used in the Super Learner yet still achieve error diversity? Guidance in this area would allow researchers to make maximal use of the ensemble while minimising computation time.

A common concern with machine learning methods relates to inference. A prediction interval on the dynamic survival predictions would be informative and is a consideration in comparing the methods. Based on the work in Part II of this thesis, uncertainty information is also of interest to patients and clinicians. Unfortunately, methods for construction of a prediction interval from a machine learning ensemble are not well established. Bootstrapping, besides being computationally infeasible with the Super Learner given the large UK CF Registry dataset, has been shown to be not consistent for some machine learning methods such as linear classifiers [Laber and Murphy, 2011] and lasso [Chatterjee and Lahiri, 2011] due to non-regularity. Although algorithm-specific adaptations have been developed such as quantile regression for random forests [Meinshausen, 2006], for an ensemble, a general method that can be applied regardless of algorithm choices is needed. Conformal inference, originally developed in by Vovk et al. [2005], offers a promising framework. It can be applied to predictions created by any algorithm or model to create a 95% prediction region [Shafer and Vovk, 2008]. A key idea is computation of a nonconformity measure, which captures how well an observation ‘conforms’ to the model. The distribution of these nonconformity scores provides the foundation for constructing a prediction interval. This seems a promising approach and I am aware of ongoing research in this area.

A strength of this work was the use of a large and high quality data set from

the UK CF Registry for implementation of the methods. This dataset had a very low level of missingness overall but it is still important to consider the implications of missing data. There are several types of missing data in the registry dataset. For example, in all analyses I removed individuals with missing CFTR mutation information. With respect to this variable, a complete case analysis was performed. This complete case analysis would be considered valid if the reason for the missingness was not associated with the survival outcome [Little, 1992]. Here, an association is possible if the missingness was due to the genotype being rare or unknown because rare genotypes are often associated with less severe disease and, therefore, improved survival. Efficiency may be lost by not using all of the data but here less than 2% of individuals had missing mutation data and so the impact is expected to have been minimal. There was also some incomplete longitudinal data. Reasons for missingness range from the data not being keyed in at the person's annual review to the patient not having an annual review because they felt good, because they felt bad, or simply because they did not attend within 12 months of their previous visit. These data were assumed to be missing at random and, for FEV1%, FVC% and BMI, we were able to predict missing values using a mixed effects model for the dynamic prediction work, which is unbiased provided that measurements are missing at random, conditional on available data. For other variables and for the mediation analysis, last observation carried forward was relied on. This is reasonable for some variables such as BMI that do not change much over time. Missingness was low overall and, computationally, it was not clear that a multiple imputation approach was feasible. It is possible that multiple imputation could be implemented in Cox landmarking but the computation time with Super Learner would be prohibitive and other methods for handling missing data in combination with these methods should be explored. This is an important consideration as many other registries may have a higher degree of missingness or it may not be possible to assume missing at random.

Finally, improved treatment of CF has led to a more than doubling of the life expectancy in CF over the past 50 years. The relatively recent development of CFTR modulators has the potential to further increase expected survival for people with CF with certain mutation types. As more people begin using these treatments for longer periods of time, it will be important to incorporate this treatment information in survival predictions.

9.2.2 Communication of survival predictions

In the area of communication of survival predictions, it is my sincere hope that the work presented here leads to future work in the CF setting. First, devising a method for allowing access to life expectancy information for those PWCFs who want it without causing distress to those PWCFs who do not want it is

critical. This may involve a combination of interaction with clinicians and/or special technology for controlling access. A second important area for further work is to understand how these models can help clinicians answer questions and motivate patients in clinic. Also, another round of feedback from PWCFs on the web application could help refine it in preparation for broader use.

9.2.3 Mediation analysis for survival outcomes

The work presented in this thesis also highlighted several areas for future research in mediation analysis with survival outcomes and longitudinal data. Complete control of confounding is a fundamental assumption to ensure validity of mediation analysis. While methods for sensitivity analysis to the assumptions of no unmeasured confounding have been suggested [VanderWeele, 2015], I am not aware of any application of these techniques in this setting. Further, as many of the techniques assume randomisation of the treatment or exposure, there is a need for development of approaches for sensitivity analyses in the context of observational datasets where we can only hope to achieve conditional randomisation of the exposure. This will be essential if these methods are to be applied in broader contexts.

In the CFRD setting investigated here, although no compelling evidence of mediation was found, the confidence intervals were wide and it cannot be ruled out. The methods used here assumed an exposure fixed at baseline but many of the people who did not have CFRD at baseline went on to develop it later. Exploring extensions to the methodology to accommodate time-varying exposures could be useful in many settings. An analysis that considers multiple mediators jointly could also provide new information. Also of interest are several sub-group analyses. Because not all people with CFRD are taking insulin, different results may be seen by comparing the treated and the untreated subgroups. It will also be interesting to perform another mediation analysis after the physiological research has advanced in this area and additional hypotheses about the mechanism for the effect of CFRD on survival have been put forward. For example, there may be a group of CFTR mutations that affect certain hormones leading to different outcomes with CFRD.

9.3 Conclusions

In this thesis, I have developed and studied methods for addressing both prediction questions and causal questions involving time-to-event outcomes using longitudinal observational data. The complexity of this setting requires many decisions to be made during the analysis. There are trade-offs between utilising the information to the full and computational complexity; increased complexity may also lead to

difficulties with interpretability of findings and usability of a proposed method by a broader community. In some cases, additional methodological complexity may not equate to improved performance while in others, it may allow for more robust and rich conclusions to be drawn from an analysis. As it may not be possible to determine this in advance, any chosen technique must be applied with these trade-offs in mind, letting the research question in combination with knowledge about the data source and structure drive the analysis. It is hoped that my methods will enable other researchers to address their questions of interest using complex longitudinal observational data, and that my findings in relation to CF provide novel insights and motivation for further research. There remain many important areas for development, some of which I have highlighted, that will increasingly enable researchers to make appropriate and optimal use of routinely collected data.

Appendix A

Preparation of a Discrete-Time Landmark Super Dataset

Preparation of a discrete-time landmark super dataset follows these steps:

1. Obtain values for every time-dependent predictor at each landmark prediction time, $\tilde{y}(s)$, $s \in s_1, \dots, s_L$. Landmark prediction times should be independent of event times and may be an equidistant grid of time points over the period or ages of interest [van Houwelingen and Putter, 2012].
2. Create a sliding landmark dataset for each landmark prediction time s in s_1, \dots, s_L containing only those individuals still at risk at s and with administrative censoring at $s + v$. Add L dummy variables to the data indicating which prediction time s the sliding landmark dataset corresponds to.
3. Select the number of discrete intervals to divide the prediction time period v into. This number should be smaller than the number of events in the sample but large enough to model the differences in individual survival times without considerable loss of information. Section 3.4.3 discusses the impact of choice of number of intervals on computation time.
4. Reformat each sliding landmark dataset into person-period format by discretising the time period from s to $s + v$. Intervals may be of equal size or, alternatively, quantiles of event times may be used so that each interval contains an approximately equal number of events [Polley and van der Laan, 2011]. Each person will have one row of data for each discrete interval to which they contribute time-at-risk. An event indicator is set to 1 only in the intervals for which an individual has an event.
5. Stack the discretised person-period format sliding landmark datasets to form one single prediction dataset.

Appendix B

Super Learner Ensemble for Simulation Study

Table B.1: The algorithm / hyperparameter combinations used in the Super Learner ensemble for the simulation study. See table 3.7 for descriptions of the various hyperparameters.

<i>Category</i>	<i>Algorithm & Hyperparameters</i>
Random forest	Ranger: no. trees = 500; min node size = 1
	Ranger: no. trees = 500; min node size = 3
	Ranger: no. trees = 500; min node size = 5
Gradient boosting	xgBoost: eta=0.1, max depth = 3, min obs per node = 10
	xgBoost: eta=0.1, max depth = 4, min obs per node = 10
	xgBoost: eta=0.1, max depth = 5, min obs per node = 10
	xgBoost: eta=0.1, max depth = 4, min obs per node = 20
	xgBoost: eta=0.3, max depth = 6, min obs per node = 10
GLM	glm
GAM	gam: degrees of freedom = 3
	gam: degrees of freedom = 5
Reference	mean model (no covariates)

Appendix C

Dynamic Survival Prediction Simulation: Reference Scenario

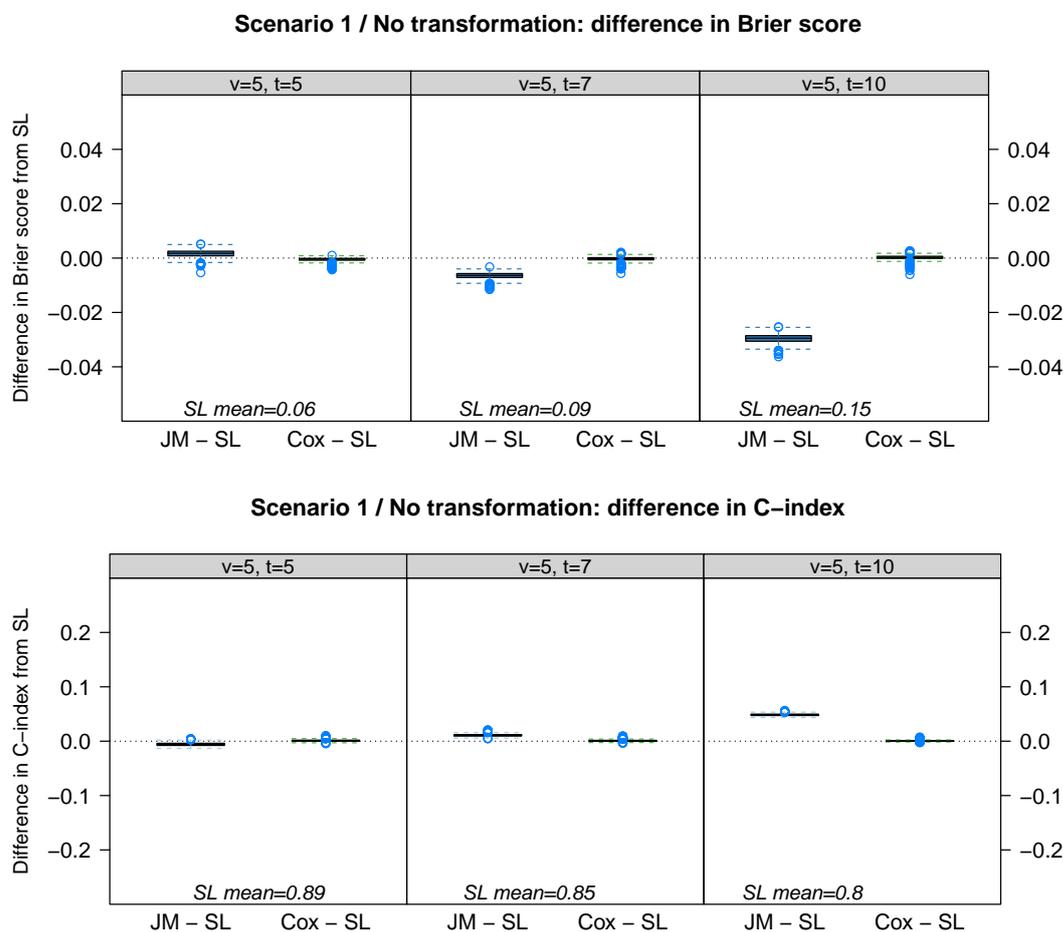


Figure C.1: Box plots of the difference in Brier score (top) and C-index (bottom) between the Super Learner landmark (SL) and the two comparison methods, Cox landmark (Cox) or joint modelling (JM), across the 1,000 simulated datasets. The longitudinal predictor was not transformed so that $y_i^*(t) = y_i(t)$. “SL mean” indicates the mean value for the Super Learner landmark approach. Each panel represents the difference at one landmark time, $t = 5, 7$, or 10 for $v = 5$ -year dynamic survival prediction.

Appendix D

Interview Topic Guide

Initial Interview Topic Guide – Semi-Structured Individual Interviews

Project Title: *Providing accessible and personalized information on life expectancy in cystic fibrosis for use in clinic and to enhance reporting from the UK Cystic Fibrosis Registry.*

Researcher: *Fahad Malik (Research Fellow MSD –EPH)*

Email: fahad.malik@lshtm.ac.uk

The interviews will be semi-structured allowing for a natural flow of conversation, however it will be grounded in the following topics to help guide the conversation and ensure collection of data in line with the aims and objectives of this study.

1. Introduction

- Thank participant for being interviewed
- Explain purpose of study

This project is a collaboration between LSHTM and the Cystic Fibrosis Trust (UK) in order to understand the current provision and accessibility of personalized information on life expectancy for people with CF (PWCF). We are looking to go further to discuss how PWCF and clinicians may want to access this information. The themes that emerge from this interview will be used to analyze the methods by which patients and clinicians wish to access this information and will help us understand the benefit to these groups.

We are making an audio recording that will be transcribed later, however the audio recording will be deleted once the transcribing is complete. We will not include your name or any identifiable information during the transcribing or in the final study report. Your participation is highly valuable, and we thank you for making yourself available for this interview, which should last an hour.

- Go through participant information sheet and the consent form, reiterating confidentiality and voluntary nature of the interview

2. Modelling in Medicine

Firstly discussing models, have you ever considered accessing information regarding life expectancy? Why or why not?

How often do you discuss life expectancy with the care team?

Life Expectancy Models

How informative do you think current available information and technology is to understanding life expectancy for CF patients?

Are you aware of the annual CF registry report which outlines median survival age estimates? Have you previously used the report and how did you interpret the median survival age estimates? Do you feel this is relevant to you and your care?

In what way do you think these models would be helpful to your ability to access care? What kinds of things would life expectancy information help you plan for?

Would these models be something best discussed in the annual general meeting with your clinicians?

Potential for an application

The aim of this section is to evaluate the comprehension, usefulness, patient preference and accessibility of data (pictorial v word) surrounding life expectancy.

- With regards to accessing personalized information on CF life expectancy what would you like to know? What do you think would be appropriate?

In this section we will be discussing the models that we have prepared to provide data on life expectancy for PWCF. We are going to be showing information about expected survival for people with CF. Because we are currently evaluating the best ways to provide this information, the data contained in these graphics and tables is imaginary. To be clear, you should not interpret this information as your expected survival. The numbers are fictitious.

Q: Do you understand / are you comfortable with this?

Q: I spoke about expected survival. The word expected can have a special statistical meaning. Do you interpret expected survival to mean the same thing as estimated survival?

As we go through the presentation, I'm going to ask you to answer some questions, and I have the questions printed here on a sheet for you as well. Please answer honestly and don't feel self-conscious if the graphic is confusing. We're asking because we want to make graphics that are easy to understand. If you find them confusing, we want to know.

Slide 1A

Expected survival is a measure of how likely it is for a person to survive up to or beyond a specific age. We will start out by looking at expected survival for all babies born with cystic fibrosis this year in the UK.

For example, for every 100 babies born with CF, we expect that 75 of them will live to 31 years of age or more. But we don't know which 75 will live to age 31 or over.

Based on this graphic, beyond what age are half of babies born with CF expected to live?

Based on this graphic, before what age will approximately half of babies born with CF pass away?

Slide 1B

This graphic also provides a summary of the estimated survival for babies born with CF in the UK this year. For example, for every 100 babies born with CF, we expect that 81 of them will live to 25 years of age or more. But we don't know which 81 will live to 25 or over.

Q: Based on this graphic, to what age are 52 babies out of 100 with CF expected to live?

Q: Do you find one of the graphics (this bar chart or the previous pictogram) easier to understand? Does one of the graphics provide information that is more useful to you?

Q: Is there a particular age or age range for which you are most interested in having survival information?

Q: If this information were real, would you want to see that at some age (here, 65 years old) only 10 out of 100 people with CF are expected to live beyond that age?

This graphic and the previous one showed estimated survival from birth but after a person has reached adulthood, estimated survival tends to be much longer because they've already survived to 18 or more years.

In the next set of graphics, we're going to show information that is more tailored to an individual. By incorporating the person's age, sex and genotype into the model for predicting survival, the information becomes specific to the group of people with CF who have those same age, sex and genotype characteristics.

Q: Are you comfortable with proceeding to the next set of graphics?

Slide 2A

In this graphic, we have entered the characteristics of a hypothetical person with CF who is female, F508 heterozygous and aged 22. I'll give you a chance to look at the graphic and read the explanation. Let me know when you're ready to proceed.

Q: Do you know your CF genotype according to the above definition?

Q: If a tool like this was available on the CF Trust web site and you could enter your age, genotype and gender, and then see estimated survival information, would you use it?

Q: Would you want to enter other ages or genotype categorisations to see how the estimates change?

Q: We've labelled graphic 1B in terms of age and graphic 2A in terms of years. Do you prefer one of these labelling schemes to the other?

Q: Based on this graphic, can you tell how many 22-year old **males** who are F508 homozygous out of 100 are expected to survive 10 or more years?

Slide 2B

This graphic shows similar information but now we've changed one of the selected characteristics (genotype category) and the graphic type. I'll give you a minute to look at the graphic. Let me know when you're ready to proceed. In this graphic, we've highlighted one place along the line where 82 out of 100 people with CF and the selected characteristics are expected to live to age 32 or beyond.

Q: Based on this graphic, can you see to what age or beyond 50 people out of 100 with CF and the selected characteristics are expected to live?

Q: If this was a real application and you could enter your characteristics, would you want this information?

Would you prefer to do this with your clinical team or on your own?

Q: Pretend for a minute that this information was real and it pertained to your characteristics. Is this information useful to you? If so, in what way?

Q: So far, we've shown you three types of graphics with estimated survival information. This one is a line graph, the previous one was a bar chart and the first one was a pictogram. Which graphic type do you prefer?

Q: Do you find any of the graphic types difficult to understand?

Slide 2C (V1/V2)

This is the same survival curve again with the addition of the labels: Bottom 15%, Middle 50% and Top 15% / Worst-Case Scenario, Most-Likely Scenario and Best-Case Scenario. Let me give you a minute to look at the graphic.

Q: What do you think the label is trying to show?

[Answer: It means that 15% or 15 out of 100 people will live to be 62 years of age or older]

Q: Do these labels make the information on the graphic more clear? Why?

In the final graphic, we're going to show information that is even more personalised. In this graphic, we've shown information for the group of people who are female, F508 homozygous and currently age 22. The next graphic will further narrow down the group of people considered by lung function, age at diagnosis, presence of infections over the past year and number of days in the hospital.

Q: Are you comfortable with proceeding to the next set of graphics?

Q: If this were an actual conversation with your healthcare team, would you want to have this more tailored information? Why or why not?

Q: If yes, would you like to access this information on your own from an app on the CF Trust web site or in a report?

Slide 3B

In this graphic, we have entered the characteristics of a hypothetical person with CF who is female, F508 homozygous, aged 22, diagnosed as a newborn, does not have CFRD, did not have staph or pseudomonas infections over the past year, had a recent FEV1 reading of between 40 and 50% and spent 7-14 days in the hospital last year receiving antibiotics.

Q: If you were presented with this, would you be able to fill out all of the information asked for? If no, which pieces do you not know?

Q: Above the section requesting information about the person is a survival curve. Do you feel comfortable interpreting this survival curve?

Q: If the characteristics on this graphic and the last one were yours and the information was real, which survival curve would you consider more reliable? More relevant to you?

Q: Does the information on this screen feel too personalised?

Q: Do you want to know how your estimated survival compares to that of other people?

Q: We have a shown a link that says "Show Uncertainty". Can you make a guess at what this may mean? [Answer: Let's talk about it using the next graphic.]

Slide 3C

This graphic is the same graphic as the previous one but with the addition of some extra grey dashed lines. Let me give you a minute to look at the graphic and read the explanatory text. Tell me when you're ready to proceed.

Q: Why do you think the dashed grey lines have been added to this graphic?
[After they try to answer, then we explain]

The idea of the dashed grey lines is to represent other possible survival curves. The turquoise line is the predicted survival curve based on our statistical model. From this line, we read information like 82 out of 100 people with CF and these selected characteristics are expected to live to age 37 or beyond. Based on the uncertainty that exists in any statistical model, there are other possible survival curves. Although the turquoise line is our best estimate of the survival

curve, it is possible that one of the dashed grey lines will be the actual survival curve for this group of people.

Q: Does seeing a graphic with uncertainty scenarios help you understand that estimated survival predictions are estimates and not certainties?

Q: Does knowing about the uncertainty in the estimated survival affect your attitude about the reliability of the information?

Q: I would now like to ask you about the presentation of information on this slide. Do you feel the wording and content of the slide is relevant and appropriate for you?

Q: How do you feel about the wording "Show/Hide Uncertainty." Or would you prefer a different label perhaps "Show other possible scenarios."

Slide 4

This information is from the 2017 CF Trust Registry Report. It displays information about the median predicted survival age. Let me give you a minute to look at it.

Q: Have you seen this information before?

Q: How do you interpret the current median predicted survival age of 47 years?

Q: Has our discussion today changed the way you interpret this information?

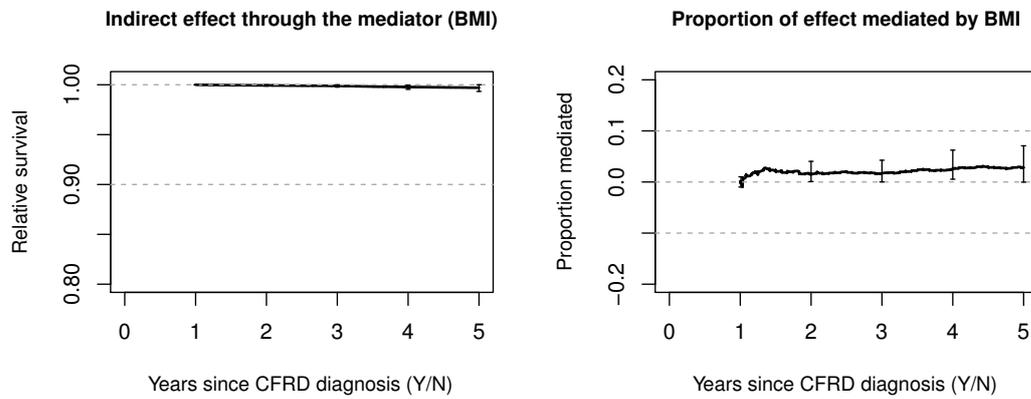
- Would the provision of this information have an impact on a PWCFs emotional response? Do you think it could have an impact on the "preservation of hope"?

Closing questions

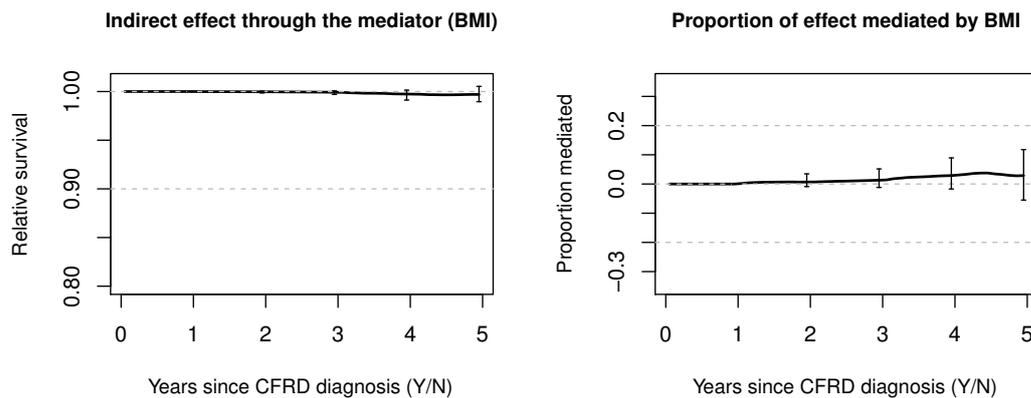
- Is there anything else you think is important to talk about in the context of this study that we haven't talked about yet?
- Thank participant for being interviewed

Appendix E

Mediation Analysis: BMI



(a) Method of Aalen effect estimates with BMI as the candidate mediator.



(b) Method of Vansteelandt effect estimates with BMI as the candidate mediator.

Figure E.1: Estimates of indirect effect of CFRD on mortality via BMI for the two mediation analysis methods. Time 0 on the x-axis represents the year when diagnosis of CFRD / no CFRD was recorded at the annual review. 95% bootstrap confidence intervals are shown at each visit time.

Appendix F

Mediation Simulation Parameters Used

Table F.1: Parameter values used to generate simulated data for the baseline scenario. To scale the hazards so that the majority of survival times were < 4 , the generated hazard was multiplied by a scaling factor, ξ .

Mediator Model		Hazard Model	
Param	Value	Param	Value
$\mu_M =$	$\begin{bmatrix} 2.9 \\ 0.0 \end{bmatrix}$	$\alpha_0 =$	0.75
$\Sigma_M =$	$\begin{bmatrix} 0.250 & -0.015 \\ -0.015 & 0.010 \end{bmatrix}$	$\alpha_{M_{[t]}}^s =$	(0.3, 0.3, 0.3, 0.3)
$\beta_{Z_0}^s =$	0.5	$\alpha_{Z_0}^s =$	0.35
$\beta_{A_t}^s =$	(0, 2, 2, 2)*	$\alpha_A^s =$	0.35**
		$\xi =$	0.35
* or (0, 0, 0, 0) for NoIE scenario		** or 0 for NoDE scenario	

Table F.2: Parameters used to generate simulated data. For each scenario, only parameters that are different from the baseline scenario reported in Table 8.1 are shown.

Type	Description	Parameters different from baseline
Baseline		
	Increasing effect of mediator on hazard over time	$\alpha_{M_{[t_j]}}^s = (0.1, 0.2, 0.3, 0.4)$
	Immediate effect of mediator on hazard	$\alpha_{M_{[t_j]}}^s = (0.0, 0.75, 0.0, 0.0)$
	Delayed effect of mediator on hazard	$\alpha_{M_{[t_j]}}^s = (0.0, 0.0, 0.6, 0.6)$
Model misspecification: unmeasured confounding		
All		$\alpha_0 = 0.5, \mu_M = \begin{bmatrix} 3.4 \\ 0.0 \end{bmatrix}$
Exposure-outcome		$p_{U_0 A=1} = 0.2, p_{U_0 A=0} = 0.8$ $\beta_{U_0}^s = 0.0$ $\alpha_{U_0}^s = 0.5$ or $\alpha_{U_0}^s = -0.5$
Mediator-outcome		$p_{U_0 A=1} = 0.5, p_{U_0 A=0} = 0.5$ $\beta_{U_0}^s = -0.5$ $\alpha_{U_0}^s = 0.5$ or $\alpha_{U_0}^s = -0.5$
Exposure-mediator		$p_{U_0 A=1} = 0.2, p_{U_0 A=0} = 0.8$ $\beta_{U_0}^s = 0.75$ or $\beta_{U_0}^s = -0.75$ $\beta_{Z_0}^s = 0.25$ $\alpha_{U_0}^s = 0.0$
Model misspecification: multiplicative hazards for event times		
Exponential, fewer events		$\alpha_0 = 0.1, \alpha_A^s = 0.4, \alpha_{Z_0}^s = 0.1$ $\alpha_{M_{[t_j]}}^s = (0.4, 0.4, 0.4, 0.4)$
Exponential, more events		$\alpha_0 = 0.25, \alpha_A^s = 0.5, \alpha_{Z_0}^s = 0.25$ $\alpha_{M_{[t_j]}}^s = (0.5, 0.5, 0.5, 0.5)$
Weibull (shape = 2)		$\alpha_0 = 0.1, \alpha_A^s = 0.5, \alpha_{Z_0}^s = 0.2$ $\alpha_{M_{[t_j]}}^s = (0.3, 0.3, 0.3, 0.3)$ $\mu_M = \begin{bmatrix} 2.5 \\ 0.0 \end{bmatrix}$
All		$\xi = 0.4$
Data availability: M_0 affects exposure		
All		$p_{Z_0 A=1} = p_{Z_0 A=0} = 0.5$ $\beta_{Z_0}^s = 0.0$ $\alpha_0 = 0.35$ $\alpha_A^s = 0.5, \alpha_{Z_0}^s = 0.0$ $\alpha_{M_0}^s = 0.1$

Data availability: time-varying covariates

All	$p_{Z_0 A=1} = p_{Z_0 A=0} = 0.5$ $\beta_{Z_0}^s = 0.25$ $\beta_{L_k}^s = (0.05, 0.05, 0.05, 0.05)$ $\psi_{Z_0} = 0, \psi_{M_{\lfloor t \rfloor}} = (0, 0, 0)$ $\psi_{A_t} = (0, 0, 0, 0)$ $\alpha_0 = 0.5, \xi = 0.45$ $\alpha_A^s = 0.2, \alpha_{Z_0}^s = 0.1$ $\alpha_{M_{\lfloor t \rfloor}}^s = (0.1, 0.1, 0.1, 0.1)$ $\alpha_{L_{\lfloor t \rfloor}}^s = (0.01, 0.01, 0.01, 0.01)$
L1	$\mu_L = \begin{bmatrix} 50 \\ 0.5 \end{bmatrix}, \Sigma_L = \begin{bmatrix} 100 & -4 \\ -4 & 4 \end{bmatrix}$
L2	$\mu_L = \begin{bmatrix} 50 \\ 5 \end{bmatrix}, \Sigma_L = \begin{bmatrix} 100 & -4 \\ -4 & 4 \end{bmatrix}$
L3	$\mu_L = \begin{bmatrix} 50 \\ 15 \end{bmatrix}, \Sigma_L = \begin{bmatrix} 100 & -4 \\ -4 & 4 \end{bmatrix}$
L4	$\mu_L = \begin{bmatrix} 50 \\ 2.5 \end{bmatrix}, \Sigma_L = \begin{bmatrix} 100 & -10 \\ -10 & 25 \end{bmatrix}$
L5	$\mu_L = \begin{bmatrix} 50 \\ 2.5 \end{bmatrix}, \Sigma_L = \begin{bmatrix} 100 & -10 \\ -10 & 100 \end{bmatrix}$
L6	$\mu_L = \begin{bmatrix} 50 \\ 0.5 \end{bmatrix}, \Sigma_L = \begin{bmatrix} 100 & -4 \\ -4 & 4 \end{bmatrix}$ $\psi_{A_k} = (0, 15, 15, 15)$

Data availability: infrequent mediator measurements

F1	$\alpha_{M_{\lfloor t \rfloor}}^s = 0.3$ for $t = 0, 0.25, \dots, 3.75$ $\beta_{A_t}^s = 2.0$ for $t = 0.25, \dots, 3.75$
F2	$\alpha_{M_{\lfloor t \rfloor}}^s = 0.3$ for $t = 0, 0.25, \dots, 3.75$ $\beta_{A_t}^s = -2.0$ for $t = 0.25, \dots, 3.75$

Appendix G

Mediation Simulation Results

Table G.1: Simulation results for the baseline scenario. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.21	0.93	0.97	0.96	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
50%	1.66	0.80	0.92	0.87	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
80%	2.45	0.62	0.84	0.74	0.00	0.00	0.00	0.01	0.01	0.01
					(1%)	(1%)	(0%)	(1%)	(1%)	(1%)
<i>No Direct Effect</i>										
20%	1.24	0.95	1.00	0.95	-0.00	-0.00	-0.00	-0.00	-0.00	0.00
					(-0%)	(-0%)	(-0%)	(-0%)	(-0%)	(0%)
50%	1.73	0.86	1.00	0.86	-0.01	-0.00	-0.00	0.00	-0.00	0.00
					(-1%)	(-0%)	(-1%)	(0%)	(-0%)	(0%)
80%	2.58	0.72	1.00	0.72	-0.01	-0.00	-0.01	-0.00	0.00	0.01
					(-2%)	(-0%)	(-1%)	(-0%)	(0%)	(1%)
<i>No Indirect Effect</i>										
20%	1.26	0.97	0.97	1.00	-0.00	0.00	-0.00	0.00	0.00	-0.00
					(-0%)	(0%)	(-0%)	(0%)	(0%)	(-0%)
50%	1.78	0.91	0.91	1.00	-0.01	0.00	-0.01	0.00	0.00	0.00
					(-1%)	(0%)	(-1%)	(0%)	(0%)	(0%)
80%	2.66	0.81	0.82	1.00	-0.01	-0.00	-0.02	0.00	0.00	0.00
					(-2%)	(-0%)	(-2%)	(0%)	(0%)	(0%)

Table G.2: Simulation results for scenario U1 with uncontrolled exposure-outcome confounding where the effect of the unmeasured confounder on the hazard is positive. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.007 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.21	0.93	0.97	0.96	0.02	0.02	0.00	0.02	0.02	0.00
					(2%)	(2%)	(0%)	(2%)	(2%)	(0%)
50%	1.63	0.81	0.93	0.87	0.06	0.07	0.00	0.06	0.07	0.00
					(7%)	(7%)	(0%)	(7%)	(7%)	(0%)
80%	2.42	0.62	0.84	0.74	0.10	0.14	0.00	0.10	0.15	0.00
					(16%)	(17%)	(-0%)	(16%)	(18%)	(0%)
<i>No Direct Effect</i>										
20%	1.23	0.95	1.00	0.95	0.02	0.02	0.00	0.02	0.02	0.00
					(2%)	(2%)	(0%)	(2%)	(2%)	(0%)
50%	1.70	0.86	1.00	0.86	0.07	0.08	0.00	0.07	0.08	0.00
					(8%)	(8%)	(0%)	(8%)	(8%)	(-0%)
80%	2.53	0.72	1.00	0.72	0.13	0.17	0.00	0.13	0.18	0.01
					(18%)	(17%)	(1%)	(18%)	(18%)	(1%)
<i>No Indirect Effect</i>										
20%	1.25	0.97	0.97	1.00	0.02	0.02	0.00	0.02	0.02	0.00
					(3%)	(3%)	(0%)	(3%)	(3%)	(0%)
50%	1.75	0.91	0.91	1.00	0.07	0.07	0.00	0.07	0.07	0.00
					(8%)	(8%)	(0%)	(8%)	(8%)	(0%)
80%	2.62	0.82	0.82	1.00	0.15	0.15	0.00	0.15	0.15	0.00
					(19%)	(19%)	(-0%)	(19%)	(19%)	(-0%)

Table G.3: Simulation results for scenario U2 with uncontrolled exposure-outcome confounding where the effect of the unmeasured confounder on the hazard is negative. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.006 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.24	0.92	0.97	0.95	-0.02	-0.02	0.00	-0.02	-0.02	0.00
					(-2%)	(-2%)	(0%)	(-2%)	(-2%)	(0%)
50%	1.73	0.78	0.91	0.86	-0.06	-0.06	0.00	-0.05	-0.06	0.00
					(-7%)	(-7%)	(-0%)	(-7%)	(-7%)	(-0%)
80%	2.59	0.59	0.82	0.72	-0.09	-0.12	0.00	-0.09	-0.11	0.00
					(-15%)	(-15%)	(-0%)	(-14%)	(-14%)	(1%)
<i>No Direct Effect</i>										
20%	1.27	0.95	1.00	0.95	-0.03	-0.03	0.00	-0.03	-0.03	0.00
					(-3%)	(-3%)	(0%)	(-3%)	(-3%)	(0%)
50%	1.81	0.84	1.00	0.84	-0.07	-0.08	0.00	-0.07	-0.08	0.00
					(-8%)	(-8%)	(0%)	(-8%)	(-8%)	(0%)
80%	2.71	0.70	1.00	0.70	-0.11	-0.16	0.00	-0.11	-0.16	0.01
					(-16%)	(-16%)	(1%)	(-16%)	(-16%)	(2%)
<i>No Indirect Effect</i>										
20%	1.29	0.97	0.97	1.00	-0.03	-0.03	0.00	-0.03	-0.03	0.00
					(-3%)	(-3%)	(0%)	(-3%)	(-3%)	(0%)
50%	1.87	0.90	0.90	1.00	-0.08	-0.08	0.00	-0.08	-0.08	0.00
					(-8%)	(-8%)	(0%)	(-8%)	(-8%)	(0%)
80%	2.80	0.80	0.80	1.00	-0.13	-0.13	0.00	-0.13	-0.13	0.00
					(-17%)	(-17%)	(-0%)	(-16%)	(-16%)	(-0%)

Table G.4: Simulation results for scenario U3 with uncontrolled mediator-outcome confounding where the effect of the unmeasured confounder on the hazard is positive. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.21	0.93	0.97	0.96	0.00	-0.01	0.01	0.00	-0.01	0.01
					(0%)	(-1%)	(1%)	(0%)	(-1%)	(1%)
50%	1.65	0.81	0.92	0.87	0.00	-0.02	0.02	0.00	-0.02	0.02
					(0%)	(-2%)	(2%)	(0%)	(-2%)	(2%)
80%	2.44	0.62	0.84	0.74	0.00	-0.03	0.04	0.00	-0.04	0.05
					(0%)	(-4%)	(5%)	(0%)	(-4%)	(7%)
<i>No Direct Effect</i>										
20%	1.23	0.95	1.00	0.95	0.00	-0.01	0.01	0.00	-0.01	0.01
					(0%)	(-1%)	(1%)	(0%)	(-1%)	(1%)
50%	1.71	0.86	1.00	0.86	0.00	-0.02	0.02	0.00	-0.02	0.02
					(0%)	(-2%)	(3%)	(0%)	(-2%)	(3%)
80%	2.55	0.72	1.00	0.72	0.00	-0.05	0.04	0.00	-0.06	0.06
					(0%)	(-5%)	(6%)	(0%)	(-6%)	(8%)
<i>No Indirect Effect</i>										
20%	1.25	0.97	0.97	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
50%	1.77	0.91	0.91	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
80%	2.64	0.82	0.82	1.00	0.00	0.00	0.00	0.00	0.01	0.00
					(0%)	(1%)	(-0%)	(1%)	(1%)	(-0%)

Table G.5: Simulation results for scenario U4 with uncontrolled mediator-outcome confounding where the effect of the unmeasured confounder on the hazard is negative. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.24	0.92	0.97	0.95	0.00	0.01	-0.01	0.00	0.01	-0.01
					(0%)	(1%)	(-1%)	(0%)	(1%)	(-1%)
50%	1.75	0.78	0.91	0.85	0.00	0.03	-0.02	0.00	0.03	-0.02
					(0%)	(3%)	(-3%)	(0%)	(3%)	(-2%)
80%	2.61	0.59	0.82	0.71	0.00	0.05	-0.04	0.00	0.06	-0.04
					(0%)	(6%)	(-5%)	(0%)	(8%)	(-6%)
<i>No Direct Effect</i>										
20%	1.27	0.94	1.00	0.94	0.00	0.01	-0.01	0.00	0.01	-0.01
					(0%)	(1%)	(-1%)	(0%)	(1%)	(-1%)
50%	1.82	0.84	1.00	0.84	0.00	0.03	-0.02	0.00	0.03	-0.02
					(0%)	(3%)	(-2%)	(0%)	(3%)	(-2%)
80%	2.74	0.69	1.00	0.69	0.00	0.06	-0.03	0.00	0.07	-0.04
					(0%)	(6%)	(-5%)	(1%)	(7%)	(-5%)
<i>No Indirect Effect</i>										
20%	1.30	0.96	0.96	1.00	-0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(-0%)	(0%)	(0%)	(-0%)	(0%)
50%	1.89	0.90	0.90	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
80%	2.83	0.80	0.80	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(-0%)	(0%)	(0%)	(-0%)

Table G.6: Simulation results for scenario U5 with uncontrolled exposure-mediator confounding where the effect of the unmeasured confounder on the mediator is positive. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.22	0.93	0.97	0.95	0.01	-0.00	0.01	0.01	-0.00	0.01
					(1%)	(-0%)	(1%)	(1%)	(-0%)	(1%)
50%	1.68	0.80	0.92	0.87	0.03	0.00	0.03	0.03	0.00	0.03
					(3%)	(0%)	(3%)	(3%)	(0%)	(3%)
80%	2.50	0.61	0.83	0.73	0.04	0.00	0.05	0.05	0.01	0.05
					(7%)	(0%)	(7%)	(8%)	(1%)	(8%)
<i>No Direct Effect</i>										
20%	1.25	0.95	1.00	0.95	0.01	0.00	0.01	0.01	0.00	0.01
					(1%)	(0%)	(1%)	(1%)	(0%)	(1%)
50%	1.75	0.85	1.00	0.85	0.03	0.00	0.03	0.03	0.00	0.03
					(4%)	(0%)	(4%)	(4%)	(0%)	(4%)
80%	2.62	0.71	1.00	0.71	0.06	0.00	0.06	0.06	0.00	0.06
					(8%)	(0%)	(8%)	(8%)	(0%)	(9%)
<i>No Indirect Effect</i>										
20%	1.27	0.97	0.97	1.00	0.01	-0.00	0.01	0.01	-0.00	0.01
					(1%)	(-0%)	(1%)	(1%)	(-0%)	(1%)
50%	1.81	0.91	0.91	1.00	0.04	0.00	0.04	0.04	0.00	0.04
					(4%)	(0%)	(4%)	(4%)	(0%)	(4%)
80%	2.71	0.81	0.81	1.00	0.07	0.00	0.08	0.07	0.00	0.09
					(9%)	(0%)	(8%)	(9%)	(0%)	(9%)

Table G.7: Simulation results for scenario U6 with uncontrolled exposure-mediator confounding where the effect of the unmeasured confounder on the mediator is negative. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.23	0.93	0.97	0.95	-0.01	0.00	-0.01	-0.01	0.00	-0.01
					(-1%)	(0%)	(-1%)	(-1%)	(0%)	(-1%)
50%	1.70	0.79	0.92	0.86	-0.03	0.00	-0.03	-0.02	0.00	-0.03
					(-3%)	(0%)	(-3%)	(-3%)	(0%)	(-3%)
80%	2.54	0.60	0.83	0.73	-0.04	0.00	-0.05	-0.04	0.01	-0.04
					(-7%)	(0%)	(-7%)	(-7%)	(1%)	(-6%)
<i>No Direct Effect</i>										
20%	1.26	0.95	1.00	0.95	-0.01	-0.00	-0.01	-0.01	-0.00	-0.01
					(-1%)	(-0%)	(-1%)	(-1%)	(-0%)	(-1%)
50%	1.78	0.85	1.00	0.85	-0.03	0.00	-0.03	-0.03	0.00	-0.03
					(-3%)	(0%)	(-3%)	(-3%)	(0%)	(-3%)
80%	2.66	0.70	1.00	0.70	-0.05	0.00	-0.05	-0.05	0.00	-0.04
					(-7%)	(0%)	(-7%)	(-7%)	(0%)	(-6%)
<i>No Indirect Effect</i>										
20%	1.28	0.97	0.97	1.00	-0.01	-0.00	-0.01	-0.01	-0.00	-0.01
					(-1%)	(-0%)	(-1%)	(-1%)	(-0%)	(-1%)
50%	1.84	0.90	0.90	1.00	-0.03	0.00	-0.04	-0.03	0.00	-0.04
					(-4%)	(0%)	(-4%)	(-4%)	(0%)	(-4%)
80%	2.75	0.81	0.81	1.00	-0.06	0.00	-0.08	-0.06	0.01	-0.08
					(-8%)	(0%)	(-8%)	(-7%)	(1%)	(-8%)

Table G.8: Simulation results for scenario W1 with event times generated according to an exponential model with 43-49% of individuals having an event. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.43	0.96	0.99	0.97	-0.00	-0.00	0.00	-0.00	0.00	-0.00
					(-0%)	(-0%)	(0%)	(-0%)	(0%)	(-0%)
50%	2.20	0.88	0.96	0.91	-0.00	-0.00	0.00	0.00	0.01	-0.00
					(-0%)	(-0%)	(0%)	(0%)	(1%)	(-0%)
80%	3.16	0.80	0.94	0.85	0.00	-0.01	0.01	0.00	0.01	-0.01
					(0%)	(-1%)	(1%)	(0%)	(1%)	(-1%)
<i>No Direct Effect</i>										
20%	1.45	0.97	1.00	0.97	0.00	0.00	-0.00	0.00	0.00	-0.00
					(0%)	(0%)	(-0%)	(0%)	(0%)	(-0%)
50%	2.24	0.92	1.00	0.92	-0.00	0.00	-0.00	-0.00	0.00	-0.00
					(-0%)	(0%)	(-0%)	(-0%)	(0%)	(-0%)
80%	3.20	0.87	1.00	0.87	-0.00	0.00	-0.00	-0.00	0.01	-0.01
					(-0%)	(0%)	(-0%)	(-0%)	(1%)	(-1%)
<i>No Indirect Effect</i>										
20%	1.48	0.99	0.99	1.00	0.00	-0.00	0.00	0.00	-0.00	0.00
					(0%)	(-0%)	(0%)	(0%)	(-0%)	(0%)
50%	2.28	0.96	0.96	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
80%	3.24	0.93	0.94	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)

Table G.9: Simulation results for scenario W2 with event times generated according to an exponential model with 79-83% of individuals having an event. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{cox}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.22	0.91	0.98	0.94	0.00	-0.01	0.00	0.00	0.00	0.00
					(0%)	(-1%)	(1%)	(0%)	(0%)	(0%)
50%	1.67	0.76	0.93	0.82	-0.01	-0.02	0.01	0.00	0.00	0.00
					(-1%)	(-2%)	(2%)	(0%)	(0%)	(0%)
80%	2.47	0.56	0.85	0.66	-0.01	-0.05	0.03	0.00	0.01	0.00
					(-2%)	(-6%)	(4%)	(0%)	(1%)	(-1%)
<i>No Direct Effect</i>										
20%	1.25	0.94	1.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(-0%)	(0%)	(0%)
50%	1.77	0.83	1.00	0.83	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
80%	2.63	0.68	1.00	0.68	-0.01	0.00	0.00	0.00	0.01	-0.01
					(-1%)	(0%)	(-1%)	(0%)	(1%)	(-1%)
<i>No Indirect Effect</i>										
20%	1.29	0.97	0.97	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
50%	1.87	0.91	0.91	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
80%	2.78	0.83	0.83	1.00	0.00	-0.01	0.01	0.00	0.00	0.00
					(0%)	(-1%)	(1%)	(0%)	(0%)	(0%)

Table G.10: Simulation results for scenario W3 with event times generated according to a Weibull model with shape parameter = 2 and 89-91% of individuals having an event. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{cox}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.39	0.93	0.97	0.96	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
50%	1.94	0.80	0.91	0.87	0.00	-0.01	0.01	0.00	0.01	0.00
					(0%)	(-1%)	(1%)	(0%)	(1%)	(0%)
80%	2.72	0.61	0.82	0.74	0.00	0.00	0.00	0.00	0.04	-0.03
					(0%)	(0%)	(0%)	(1%)	(5%)	(-4%)
<i>No Direct Effect</i>										
20%	1.45	0.96	1.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
50%	2.06	0.88	1.00	0.88	0.00	0.01	-0.01	0.00	0.01	-0.01
					(0%)	(1%)	(-1%)	(0%)	(1%)	(-1%)
80%	2.86	0.76	1.00	0.76	0.00	0.03	-0.02	0.00	0.05	-0.03
					(0%)	(3%)	(-3%)	(0%)	(5%)	(-5%)
<i>No Indirect Effect</i>										
20%	1.46	0.96	0.96	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
50%	2.08	0.89	0.90	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
80%	2.88	0.79	0.79	1.00	0.00	0.00	0.01	0.00	0.00	0.00
					(0%)	(0%)	(1%)	(1%)	(0%)	(0%)

Table G.11: Simulation results for scenario B1 where the baseline mediator measurement affects exposure and the hazard prior to the first visit time. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects, with adjustment for M_0</i>										
20%	1.26	0.91	0.96	0.95	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
50%	1.78	0.74	0.87	0.85	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
80%	2.66	0.53	0.75	0.71	0.00	0.01	-0.00	0.00	0.01	-0.00
					(0%)	(1%)	(-0%)	(1%)	(2%)	(-0%)
<i>Both Direct and Indirect Effects, without adjustment for M_0</i>										
20%	1.26	0.91	0.96	0.95	-0.00	0.00	-0.01	-0.00	0.00	-0.00
					(-1%)	(0%)	(-1%)	(-0%)	(0%)	(-0%)
50%	1.78	0.74	0.87	0.85	-0.01	0.00	-0.01	-0.01	0.00	-0.01
					(-1%)	(0%)	(-2%)	(-1%)	(0%)	(-1%)
80%	2.66	0.53	0.75	0.71	-0.02	0.01	-0.03	-0.02	0.01	-0.02
					(-3%)	(1%)	(-4%)	(-3%)	(2%)	(-3%)

Table G.12: Simulation results for scenario B2 where the baseline mediator measurement affects exposure and the survival time in all waves. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects, with adjustment for M_0</i>										
20%	1.24	0.91	0.96	0.95	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
50%	1.73	0.76	0.88	0.86	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
80%	2.58	0.55	0.76	0.72	0.00	0.01	-0.00	0.00	0.02	0.00
					(0%)	(1%)	(-0%)	(1%)	(2%)	(0%)
<i>Both Direct and Indirect Effects, without adjustment for M_0</i>										
20%	1.24	0.91	0.96	0.95	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
					(-1%)	(-1%)	(-1%)	(-1%)	(-1%)	(-1%)
50%	1.73	0.76	0.88	0.86	-0.03	-0.01	-0.02	-0.03	-0.01	-0.02
					(-4%)	(-2%)	(-2%)	(-4%)	(-2%)	(-2%)
80%	2.58	0.55	0.76	0.72	-0.05	-0.03	-0.04	-0.04	-0.01	-0.04
					(-9%)	(-3%)	(-5%)	(-8%)	(-2%)	(-5%)

Table G.13: Simulation results for scenario L1 with a time-varying covariate with mean random slope = 0.5 that is not affected by the exposure. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.26	0.96	0.98	0.98	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(0%)	(-0%)	(-0%)	(0%)	(-0%)
50%	1.78	0.87	0.93	0.93	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(-0%)	(0%)	(1%)	(-0%)
80%	2.67	0.74	0.86	0.86	0.00	0.01	0.00	0.00	0.00	0.01
					(-0%)	(1%)	(-0%)	(0%)	(1%)	(1%)
<i>No Direct Effect</i>										
20%	1.28	0.98	1.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(-0%)	(0%)	(0%)	(-0%)
50%	1.85	0.93	1.00	0.93	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(-0%)	(0%)	(0%)	(0%)
80%	2.76	0.85	1.00	0.85	0.00	0.01	0.00	0.00	0.01	0.01
					(0%)	(1%)	(-0%)	(1%)	(1%)	(1%)
<i>No Indirect Effect</i>										
20%	1.28	0.97	0.97	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(-0%)	(0%)	(-0%)	(-0%)	(0%)
50%	1.84	0.93	0.93	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(-0%)	(0%)	(0%)	(-0%)
80%	2.76	0.85	0.85	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(0%)	(-0%)	(0%)	(0%)	(-0%)

Table G.14: Simulation results for scenario L2 with a time-varying covariate with mean random slope = 5.0 that is not affected by the exposure. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.25	0.96	0.98	0.98	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(-0%)	(-0%)	(-0%)	(0%)	(-0%)
50%	1.77	0.87	0.93	0.93	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(0%)	(-0%)	(-0%)	(-0%)	(0%)
80%	2.64	0.75	0.86	0.87	0.00	0.01	-0.01	0.00	0.01	0.01
					(-0%)	(1%)	(-1%)	(0%)	(1%)	(1%)
<i>No Direct Effect</i>										
20%	1.28	0.98	1.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(-0%)	(0%)	(-0%)	(-0%)	(0%)
50%	1.84	0.93	1.00	0.93	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(0%)	(-0%)	(-0%)	(0%)	(0%)
80%	2.74	0.86	1.00	0.85	0.00	0.01	0.00	0.00	0.00	0.01
					(0%)	(1%)	(-0%)	(0%)	(0%)	(1%)
<i>No Indirect Effect</i>										
20%	1.28	0.97	0.98	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
50%	1.84	0.93	0.93	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(-0%)	(0%)	(0%)	(-0%)
80%	2.74	0.86	0.85	1.00	0.00	0.01	0.00	0.01	0.01	0.00
					(0%)	(1%)	(-0%)	(1%)	(1%)	(-0%)

Table G.15: Simulation results for scenario L3 with a time-varying covariate with mean random slope = 15.0 that is not affected by the exposure. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.24	0.96	0.98	0.98	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(-0%)	(-0%)	(-0%)	(-0%)	(-0%)
50%	1.74	0.87	0.93	0.94	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(0%)	(-0%)	(0%)	(0%)	(0%)
80%	2.57	0.76	0.87	0.87	0.00	0.01	0.00	0.00	0.01	0.00
					(-0%)	(1%)	(-1%)	(-0%)	(1%)	(1%)
<i>No Direct Effect</i>										
20%	1.27	0.98	1.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(-0%)	(-0%)	(-0%)	(-0%)	(0%)
50%	1.81	0.93	1.00	0.93	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(-0%)	(0%)	(0%)	(0%)
80%	2.67	0.86	1.00	0.86	0.00	0.01	0.00	0.00	0.01	0.01
					(0%)	(1%)	(-0%)	(0%)	(1%)	(1%)
<i>No Indirect Effect</i>										
20%	1.27	0.98	0.98	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
50%	1.81	0.93	0.93	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(-0%)	(0%)	(0%)	(-0%)
80%	2.67	0.86	0.86	1.00	0.00	0.01	0.00	0.01	0.01	0.00
					(1%)	(1%)	(-0%)	(1%)	(1%)	(-0%)

Table G.16: Simulation results for scenario L4 with a time-varying covariate that is not affected by the exposure and has a moderate random slope variance. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.26	0.96	0.98	0.98	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(0%)	(-0%)	(-0%)	(0%)	(-0%)
50%	1.78	0.87	0.93	0.93	-0.00	0.01	-0.01	0.00	0.00	0.00
					(-0%)	(1%)	(-1%)	(-0%)	(-0%)	(0%)
80%	2.65	0.74	0.86	0.86	0.00	0.03	-0.02	0.00	0.01	0.00
					(-0%)	(3%)	(-3%)	(0%)	(1%)	(0%)
<i>No Direct Effect</i>										
20%	1.28	0.98	1.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(0%)	(-0%)	(-0%)	(-0%)	(0%)
50%	1.84	0.93	1.00	0.93	0.00	0.01	-0.01	0.00	0.00	0.00
					(-0%)	(1%)	(-1%)	(0%)	(0%)	(0%)
80%	2.75	0.85	1.00	0.85	0.00	0.03	-0.02	0.00	0.00	0.01
					(-0%)	(3%)	(-3%)	(0%)	(0%)	(1%)
<i>No Indirect Effect</i>										
20%	1.28	0.97	0.98	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
50%	1.84	0.93	0.93	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(-0%)	(0%)	(0%)	(-0%)
80%	2.75	0.85	0.85	1.00	0.00	0.01	0.00	0.00	0.01	0.00
					(0%)	(1%)	(-0%)	(1%)	(1%)	(-0%)

Table G.17: Simulation results for scenario L5 with a time-varying covariate that is not affected by the exposure and has a large random slope variance. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.26	0.96	0.98	0.98	-0.00	0.01	-0.01	0.00	0.00	0.00
					(-0%)	(1%)	(-1%)	(-0%)	(0%)	(0%)
50%	1.77	0.87	0.93	0.93	0.00	0.02	-0.02	0.00	0.00	0.00
					(-0%)	(2%)	(-2%)	(-0%)	(0%)	(0%)
80%	2.64	0.75	0.86	0.86	-0.00	0.07	-0.07	0.00	0.01	0.00
					(-0%)	(9%)	(-8%)	(0%)	(1%)	(0%)
<i>No Direct Effect</i>										
20%	1.28	0.98	1.00	0.98	0.00	0.01	-0.01	0.00	0.00	0.00
					(-0%)	(1%)	(-1%)	(-0%)	(-0%)	(0%)
50%	1.83	0.93	1.00	0.93	0.00	0.02	-0.02	0.00	0.00	0.00
					(0%)	(2%)	(-2%)	(0%)	(0%)	(0%)
80%	2.73	0.86	1.00	0.85	0.00	0.09	-0.07	0.00	0.00	0.01
					(0%)	(9%)	(-8%)	(0%)	(-0%)	(1%)
<i>No Indirect Effect</i>										
20%	1.28	0.98	0.98	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
50%	1.83	0.93	0.93	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(-0%)	(0%)	(0%)	(-0%)
80%	2.73	0.86	0.85	1.00	0.00	0.00	0.00	0.00	0.01	0.00
					(0%)	(1%)	(-0%)	(0%)	(1%)	(-0%)

Table G.18: Simulation results for scenario L6 with a time-varying covariate that is affected by the exposure. The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.005 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.27	0.94	0.98	0.96	0.00	-0.02	0.02	0.00	0.00	0.00
					(-0%)	(-2%)	(2%)	(-0%)	(-0%)	(0%)
50%	1.81	0.82	0.93	0.88	0.00	-0.05	0.05	0.00	0.00	0.00
					(-0%)	(-5%)	(6%)	(-0%)	(0%)	(0%)
80%	2.71	0.66	0.86	0.76	0.00	-0.09	0.09	0.00	0.01	0.01
					(-0%)	(-11%)	(12%)	(0%)	(1%)	(1%)
<i>No Direct Effect</i>										
20%	1.32	0.97	1.00	0.97	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(-0%)	(0%)	(-0%)	(-0%)	(0%)
50%	1.93	0.92	1.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(-0%)	(0%)	(0%)	(-0%)	(0%)
80%	2.88	0.84	1.00	0.84	0.00	0.00	0.00	0.00	0.00	0.01
					(-0%)	(-0%)	(0%)	(0%)	(0%)	(1%)
<i>No Indirect Effect</i>										
20%	1.29	0.95	0.97	0.98	0.00	-0.02	0.02	0.00	0.00	0.00
					(-0%)	(-2%)	(2%)	(0%)	(0%)	(0%)
50%	1.88	0.87	0.92	0.94	0.00	-0.05	0.06	0.00	0.00	0.00
					(0%)	(-6%)	(6%)	(0%)	(0%)	(-0%)
80%	2.80	0.75	0.85	0.89	0.00	-0.10	0.11	0.00	0.00	0.00
					(0%)	(-11%)	(13%)	(0%)	(1%)	(0%)

Table G.19: Simulation results for scenario F1 when event times were generated based on 4 mediator measurements per year but only annual mediator measurements are available in the analysis and A positively affects M . The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.006 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events Time	Truth			Aalen			Vansteelandt ^{add}			
	TE	DE	IE	TE	DE	IE	TE	DE	IE	
<i>Both Direct and Indirect Effects</i>										
20%	1.21	0.93	0.97	0.96	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(-0%)	(-0%)	(-0%)	(0%)	(0%)
50%	1.66	0.80	0.92	0.87	0.00	-0.06	0.06	0.00	-0.06	0.06
					(0%)	(-7%)	(7%)	(0%)	(-6%)	(7%)
80%	2.45	0.62	0.84	0.74	0.00	-0.12	0.13	0.00	-0.11	0.13
					(0%)	(-15%)	(18%)	(0%)	(-14%)	(18%)
<i>No Direct Effect</i>										
20%	1.24	0.95	1.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(-0%)	(0%)	(0%)	(-0%)
50%	1.72	0.86	1.00	0.86	0.00	-0.07	0.07	0.00	-0.07	0.07
					(0%)	(-7%)	(8%)	(0%)	(-7%)	(8%)
80%	2.57	0.72	1.00	0.72	0.00	-0.16	0.15	0.00	-0.15	0.14
					(0%)	(-16%)	(20%)	(1%)	(-15%)	(20%)
<i>No Indirect Effect</i>										
20%	1.26	0.97	0.97	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(-0%)	(0%)	(0%)	(-0%)	(0%)
50%	1.78	0.91	0.91	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
80%	2.66	0.82	0.82	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(-0%)	(0%)	(-0%)	(0%)	(0%)	(-0%)

Table G.20: Simulation results for scenario F2 when event times were generated based on 4 mediator measurements per year but but only annual mediator measurements are available in the analysis and A negatively affects M . The estimated absolute bias is the difference between the estimated effect and the true effect. Percent bias is shown beneath the absolute bias in parentheses. Results are shown at times corresponding to the 20th, 50th (median) and 80th percentile of event occurrence in each sub-scenario. The Monte Carlo Standard Error was less than 0.010 for the estimate of absolute bias for both methods at all time-points in all three sub-scenarios.

Events	Time	Truth			Aalen			Vansteelandt ^{add}		
		TE	DE	IE	TE	DE	IE	TE	DE	IE
<i>Both Direct and Indirect Effects</i>										
20%	1.32	1.03	0.96	1.07	0.00	0.01	-0.01	0.00	0.01	-0.01
					(0%)	(1%)	(-1%)	(0%)	(1%)	(-1%)
50%	1.93	1.08	0.89	1.21	0.00	0.11	-0.13	0.00	0.11	-0.13
					(0%)	(13%)	(-11%)	(0%)	(13%)	(-11%)
80%	2.88	1.18	0.79	1.48	0.01	0.22	-0.31	0.00	0.20	-0.29
					(0%)	(27%)	(-21%)	(0%)	(25%)	(-19%)
<i>No Direct Effect</i>										
20%	1.36	1.08	1.00	1.08	0.00	0.02	-0.02	0.00	0.02	-0.02
					(0%)	(2%)	(-2%)	(0%)	(2%)	(-2%)
50%	2.04	1.24	1.00	1.24	0.00	0.14	-0.15	0.00	0.14	-0.15
					(0%)	(14%)	(-12%)	(0%)	(14%)	(-12%)
80%	3.01	1.52	1.00	1.53	0.00	0.31	-0.35	-0.00	0.28	-0.33
					(0%)	(31%)	(-23%)	(-0%)	(28%)	(-21%)
<i>No Indirect Effect</i>										
20%	1.26	0.97	0.97	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
50%	1.78	0.91	0.91	1.00	0.00	0.00	0.00	0.00	0.00	0.00
					(0%)	(0%)	(0%)	(0%)	(0%)	(0%)
80%	2.66	0.82	0.82	1.00	0.00	0.00	0.00	0.00	0.01	0.00
					(0%)	(1%)	(-0%)	(1%)	(1%)	(-0%)

Appendix H

Statement of Joint Work

General

My PhD supervisors provided important discussion and guidance on all of the work described in this thesis.

Dynamic Prediction of Survival

Part II of this thesis, “Dynamic Prediction of Survival”, contains work published in the paper: Tanner, K.T., Sharples, L.D., Daniel, R.M. and Keogh, R.H. (2020), Dynamic survival prediction combining landmarking with a machine learning ensemble: Methodology and empirical comparison. *J R Stat Soc Series A*, <https://doi.org/10.1111/rssa.12611>. I developed the methods with guidance from Professor Ruth Keogh, Professor Linda Sharples, and Dr Rhian Daniel and conducted all analyses. I drafted the manuscript, which was then finalised with input from the other authors. The final version of the manuscript benefited from the input from two reviewers.

Communication of Survival Predictions

Part III of this thesis, “Communication of Survival Predictions”, includes a description of some work performed jointly with Fahad Malik, a qualitative researcher and Ruth Keogh, my PhD supervisor. Specifically, Dr Malik prepared the ethics submission to obtain permission to conduct interviews with clinicians and people with CF. This included creation of a consent form and participant information sheet. I performed the literature review of preferred graphical formats and labels, ability to understand graphics by a broad audience and best practices in survival communication. Based on that review, I designed a paper-based sample presentation of life expectancy information and accompanying script to be used in the semi-structured interviews with clinicians and people with cystic fibrosis (CF). Dr Malik prepared the overall interview scripts including information about our study, current access to information, understanding how the interviewee felt about

accessing life expectancy information and incorporation of the paper-based model. Of the 7 people with CF interviewed, Dr Malik interviewed 6 of them; Ruth Keogh and I interviewed one. Of the health care professionals interviewed, Dr Malik interviewed 7, Ruth Keogh interviewed 3 and I interviewed 3. Dr Malik transcribed all of the recorded interviews. I used thematic analysis to summarise key themes expressed in the interviews and selected representative quotes. Using the results of the interviews, I updated the paper-based life expectancy presentation and built an R Shiny online tool for presentation of life expectancy conditional on age and gender. Throughout this process, Ruth Keogh provided supervision and helpful ideas and comments.

Bibliography

- O. O. Aalen, M. J. Stensrud, V. Didelez, R. Daniel, K. Røysland, and S. Strohmaier. Time-dependent mediators in survival analysis: Modeling direct and indirect effects with the additive hazards model. *Biometrical Journal*, 62:532–549, 2020. doi: 10.1002/bimj.201800263.
- S. D. Aaron, A. L. Stephenson, D. W. Cameron, and G. A. Whitmore. A statistical model to predict one-year risk of death in patients with cystic fibrosis. *Journal of Clinical Epidemiology*, 68:1336–1345, 2015. doi: 10.1016/j.jclinepi.2014.12.010.
- A. I. Adler, B. S. F. Shine, P. Chamnan, C. S. Haworth, and D. Bilton. Genetic Determinants and Epidemiology of Cystic Fibrosis – Related Diabetes: Results from a British cohort of children and adults. *Diabetes Care*, 31(9):1789–1794, 2008. doi: 10.2337/dc08-0466.
- A. M. Alaa and M. van der Schaar. Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning. *Scientific Reports*, 8(11242), 2018. doi: 10.1038/s41598-018-29523-2.
- P. D. Allison. Discrete-Time Methods for the Analysis of Event Histories. *Sociological Methodology*, 13:61–98, 1982. URL <http://www.jstor.org/stable/270718?origin=crossref>.
- D. G. Altman and P. Royston. What do we mean by validating a prognostic model? *Statistics in Medicine*, 19:453–473, 2000.
- D. G. Altman, Y. Vergouwe, P. Royston, and K. G. M. Moons. Prognosis and prognostic research: validating a prognostic model. *BMJ*, 338(b605), 2009. doi: 10.1136/bmj.b605.
- D. F. Alwin and R. M. Hauser. The Decomposition of Effects in Path Analysis. *American Sociological Review*, 40(1):37–47, 1975.
- J. S. Ancker, Y. Senathrajah, R. Kukafka, and J. B. Starren. Design Features of Graphs in Health Risk Communication : A Systematic Review. *Journal*

- of the American Medical Informatics Association, 13(6):608–619, 2006. doi: 10.1197/jamia.M2115.
- P. K. Andersen and M. Pohar Perme. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19:71–99, 2010. doi: 10.1177/0962280209105020.
- J. R. Anderson, K. C. Cain, and R. D. Gelber. Analysis of Survival by Tumor Response. *Journal of Clinical Oncology*, 1:710–719, 1983.
- E. R. Andrinopoulou, D. Rizopoulos, M. L. Geleijnse, E. Lesaffre, A. J. J. C. Bogers, and J. J. M. Takkenberg. Dynamic prediction of outcome for patients with severe aortic stenosis: application of joint models for longitudinal and time-to-event data. *BMC Cardiovascular Disorders*, 15(28), 2015. doi: 10.1186/s12872-015-0035-z.
- P. Aram, L. Trela-Larsen, A. Sayers, A. F. Hills, A. W. Blom, E. V. McCloskey, V. Kadiramanathan, and J. M. Wilkinson. Estimating an Individual’s Probability of Revision Surgery After Knee Replacement: A Comparison of Modeling Approaches Using a National Dataset. *American Journal of Epidemiology*, 187(10):2252–2262, 2018. doi: 10.1093/aje/kwy121.
- P. Aurora, A. Wade, P. Whitmore, and B. Whitehead. A model for predicting life expectancy of children with cystic fibrosis. *European Respiratory Journal*, 16(6):1056–1060, 2000. doi: 10.1034/j.1399-3003.2000.16f06.x.
- R. M. Baron and D. A. Kenny. The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182, 1986.
- R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005. doi: 10.1002/sim.2059.
- L. Bessonova, N. Volkova, M. Higgins, L. Bengtsson, S. Tian, C. Simard, M. W. Konstan, G. S. Sawicki, A. Sewall, S. Nyangoma, A. Elbert, B. C. Marshall, and D. Bilton. Data from the US and UK cystic fibrosis registries support disease modification by CFTR modulation with ivacaftor. *Thorax*, 73(8):731–740, 2018. doi: 10.1136/thoraxjnl-2017-210394.
- I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Statistics Surveys*, 5:44–71, 2011. doi: 10.1214/09-SS047.

- A. W. Bowman. Graphics for uncertainty. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 182(2):403–418, 2019. doi: 10.1111/rssa.12379.
- M. P. Boyle and K. De Boeck. A new era in the treatment of cystic fibrosis: Correction of the underlying CFTR defect. *The Lancet Respiratory Medicine*, 1(2):158–163, 2013. doi: 10.1016/S2213-2600(12)70057-7.
- V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa.
- L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001a. doi: 10.1023/A:1010933404324.
- L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001b. doi: 10.1214/ss/1009213726.
- A. L. Brennan, K. M. Gyi, D. M. Wood, J. Johnson, R. Holliman, D. L. Baines, B. J. Philips, D. M. Geddes, M. E. Hodson, and E. H. Baker. Airway glucose concentrations and effect on growth of respiratory pathogens in cystic fibrosis. *Journal of Cystic Fibrosis*, 6(2):101–109, 2007. doi: 10.1016/j.jcf.2006.03.009.
- N. Bridges, R. Rowe, and R. I. Holt. Unique challenges of cystic fibrosis-related diabetes. *Diabetic Medicine*, 35(9):1181–1188, 2018. doi: 10.1111/dme.13652.
- L. J. Brighton and K. Bristowe. Communication in palliative care: Talking about the end of life, before the end of life. *Postgraduate Medical Journal*, 92(1090):466–470, 2016. doi: 10.1136/postgradmedj-2015-133368.
- S. Brilleman. simjm: Simulate Joint Longitudinal and Survival Data, 2019. URL <https://github.com/sambrilleman/simjm>.
- S. L. Brilleman, M. J. Crowther, M. Moreno-Betancur, J. Buros Novik, and R. Wolfe. Joint longitudinal and time-to-event models via Stan., 2018. URL https://github.com/stan-dev/stancon_talks/.
- R. Brookmeyer and M. H. Gail. Biases in Prevalent Cohorts. *Biometrics*, 43(4):739–749, 1987. doi: 10.2307/2531529.
- G. Brown. *Diversity in Neural Network Ensembles*. Doctor of philosophy, The University of Birmingham, 2004.
- G. Brown, J. Wyatt, R. Harris, and X. Yao. Diversity creation methods: A survey and categorisation. *Journal of Information Fusion*, 6(1), 2005. doi: 10.1016/j.inffus.2004.04.004.

- J. Burkell. What are the chances? Evaluating risk and benefit information in consumer health materials. *Journal of the Medical Library Association*, 92(2): 200–8, 2004.
- J. B. Buse, S. C. Bain, J. F. Mann, M. A. Nauck, S. E. Nissen, S. Pocock, N. R. Poulter, R. E. Pratley, M. Linder, T. M. Fries, D. D. Ørsted, and B. Zinman. Cardiovascular risk reduction with liraglutide: An exploratory mediation analysis of the leader trial. *Diabetes Care*, 43(7):1546–1552, 2020. doi: 10.2337/dc19-2251.
- R. Buzzetti, G. Alicandro, L. Minicucci, S. Notarnicola, M. L. Furnari, G. Giordano, V. Lucidi, E. Montemitto, V. Raia, G. Magazzú, G. Vieni, S. Quattrucci, A. Ferrazza, R. Gagliardini, N. Cirilli, D. Salvatore, and C. Colombo. Validation of a predictive survival model in Italian patients with cystic fibrosis. *Journal of Cystic Fibrosis*, 11(1):24–29, 2012. doi: 10.1016/j.jcf.2011.08.007.
- Cardiothoracic Advisory Group on behalf of NHSBT. POLICY POL231 / 3 Lung Candidate Selection Criteria. Technical report, NHSBT, 2018. URL https://nhsbt.dbe.blob.core.windows.net/umbraco-assets-corp/10834/pol231_3-lung-selection-policy.pdf.
- B. R. Celli, C. Cote, J. Marin, C. Casanova, M. Montes de Oca, R. A. Mendez, V. P. Plata, and H. J. Cabral. The Body-Mass Index, Airflow Obstruction, Dyspnea, and Exercise Capacity Index in Chronic Obstructive Pulmonary Disease. *New England Journal of Medicine*, 350(10):1005–12, 2004. doi: 10.1056/NEJMoa021322.
- P. Chamnan, B. S. F. Shine, C. S. Haworth, D. Bilton, and A. I. Adler. Diabetes as a determinant of mortality in cystic fibrosis. *Diabetes Care*, 33(2):311–316, 2010. doi: 10.2337/dc09-1215.
- W. Chang, J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges. shiny: Web Application Framework for R, 2021. URL <https://cran.r-project.org/package=shiny>.
- E. Chapman, A. Landy, A. Lyon, C. S. Haworth, and D. Bilton. End of life care for adult cystic fibrosis patients: Facilitating a good enough death. *Journal of Cystic Fibrosis*, 4(4):249–257, 2005. doi: 10.1016/j.jcf.2005.07.001.
- A. Chatterjee and S. N. Lahiri. Bootstrapping Lasso Estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011. doi: 10.1198/jasa.2011.tm10159.

- T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li. *xgboost: Extreme Gradient Boosting*, 2018. URL <https://cran.r-project.org/package=xgboost>.
- Y. Y. Chi and J. G. Ibrahim. Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62:432–445, 2006. doi: 10.1111/j.1541-0420.2005.00448.x.
- B. Choodari-Oskooei, P. Royston, and M. K. Parmar. A simulation study of predictive ability measures in a survival model I: Explained variation measures. *Statistics in Medicine*, 31:2627–2643, 2012. doi: 10.1002/sim.4242.
- E. Christodoulou, M. Jie, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 2019. doi: 10.1016/j.jclinepi.2019.02.004.
- M. G. Clarke, K. P. Kennedy, and R. P. MacDonagh. Discussing life expectancy with surgical patients: Do patients want to know and how should this information be delivered? *BMC Medical Informatics and Decision Making*, 8:24, 2008. doi: 10.1186/1472-6947-8-24.
- A. K. Clift, C. A. Coupland, R. H. Keogh, K. Diaz-Ordaz, E. Williamson, E. M. Harrison, A. Hayward, H. Hemingway, P. Horby, N. Mehta, J. Benger, K. Khunti, D. J. Spiegelhalter, A. Sheikh, J. Valabhji, R. A. Lyons, J. Robson, M. G. Semple, F. Kee, P. Johnson, S. Jebb, T. Williams, and J. Hippisley-Cox. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *BMJ*, 371, 2020. doi: 10.1136/bmj.m3731.
- J. M. Collaco, K. S. Raraigh, L. J. Appel, and G. R. Cutting. Respiratory pathogens mediate the association between lung function and temperature in cystic fibrosis. *Journal of Cystic Fibrosis*, 15(6):794–801, 2016. doi: 10.1016/j.jcf.2016.05.012.
- G. Collins. Clinical prediction models: a field in crisis? In *LSHTM Centre for Statistics and Methodology Seminar*, 2021.
- A. Cooper. *The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity*. Sams Publishing, second edition, 2004. ISBN 0-672-32614-0.

- C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3): 273–297, 1995. doi: 10.1007/BF00994018.
- D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. doi: 10.1007/978-1-4612-4380-9_37.
- M. J. Crowther and P. C. Lambert. Simulating biologically plausible complex survival data. *Statistics in Medicine*, 32:4118–4134, 2013. doi: 10.1002/sim.5823.
- Cystic Fibrosis Foundation. 2017 Patient Registry Annual Data Report. Technical report, Cystic Fibrosis Foundation, Bethesda, MD, 2018. URL <https://www.cff.org/Research/Researcher-Resources/Patient-Registry/2017-Patient-Registry-Annual-Data-Report.pdf>.
- Cystic Fibrosis Foundation. 2019 Patient Registry Annual Data Report. Technical report, Cystic Fibrosis Foundation, Bethesda, MD, 2020. URL <https://www.cff.org/Research/Researcher-Resources/Patient-Registry/2019-Patient-Registry-Annual-Data-Report.pdf>.
- Cystic Fibrosis Trust. Cystic Fibrosis-Related Diabetes Factsheet May 2017. Technical report, London, 2017. URL <https://www.cysticfibrosis.org.uk/sites/default/files/2020-11/CFRDfactsheetSep2017.pdf>.
- Cystic Fibrosis Trust. Cystic fibrosis FAQs, 2021. URL <https://www.cysticfibrosis.org.uk/what-is-cystic-fibrosis/faqs#Howcommoniscysticfibrosis?>
- R. B. D’Agostino, M.-L. Lee, A. Belanger, L. Cupples, K. Anderson, and W. B. Kannel. Relation Of Pooled Logistic Regression To Time Dependent Cox Regression Analysis: The Framingham Heart Study. *Statistics in Medicine*, 9: 1501–1515, 1990.
- C. Davis, A. McNair, A. Brigic, M. Clarke, S. Brookes, M. Thomas, and J. Blazeby. Optimising methods for communicating survival data to patients undergoing cancer surgery. *European Journal of Cancer*, 46(18):3192–3199, dec 2010. doi: 10.1016/j.ejca.2010.07.030. URL <http://dx.doi.org/10.1016/j.ejca.2010.07.030https://linkinghub.elsevier.com/retrieve/pii/S0959804910007082>.
- L. C. de Wreede, M. Fiocco, and H. Putter. The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine*, 99(3):261–274, 2010. doi: 10.1016/j.cmpb.2010.01.001.

- L. C. de Wreede, M. Fiocco, and H. Putter. mstate : An R Package for the Analysis of Competing Risks and Multi-State Models. *Journal of Statistical Software*, 38(7):1–30, 2011. doi: 10.18637/jss.v038.i07.
- T. P. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E. W. Steyerberg, and K. G. Moons. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology*, 68:279–289, 2015. doi: 10.1016/j.jclinepi.2014.06.018.
- V. Didelez. Defining causal mediation with a longitudinal mediator and a survival outcome. *Lifetime Data Analysis*, 2018. doi: 10.1007/s10985-018-9449-0.
- V. Didelez, A. Philip Dawid, and S. Geneletti. Direct and indirect effects of sequential treatments. *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence, UAI 2006*, (1993):138–146, 2006.
- J. A. Dodge, P. A. Lewis, M. Stanton, and J. Wilsher. Cystic fibrosis mortality and survival in the UK: 1947-2003. *European Respiratory Journal*, 29(3):522–526, 2007. doi: 10.1183/09031936.00099506.
- F. Downen, K. Sidhu, E. Broadbent, and H. Pilmore. Communicating projected survival with treatments for chronic kidney disease: Patient comprehension and perspectives on visual aids. *BMC Medical Informatics and Decision Making*, 17(1):1–8, 2017. doi: 10.1186/s12911-017-0536-z.
- A. Edwards, R. Thomas, R. Williams, A. L. Ellner, P. Brown, and G. Elwyn. Presenting risk information to people with diabetes: Evaluating effects and preferences for different formats by a web-based randomised controlled trial. *Patient Education and Counseling*, 63(3 SPEC. ISS.):336–349, 2006. doi: 10.1016/j.pec.2005.12.016.
- D. Edwards, P. Cullinan, D. Taylor-Robinson, D. K. Schluter, S. B. Carr, and C.-E. group. Exploring the relationship between CFRD treatment and lung function. *Journal of Cystic Fibrosis*, 18S1:WS15–3, 2019.
- B. Efron. How Biased Is the Apparent Error Rate of a Prediction Rule? *J American Statistical Association*, 81(394):461–470, 1986.
- J. S. Elborn. Cystic fibrosis. *Lancet*, 388:2519–31, 2016. doi: 10.1016/S0140-6736(16)00576-6.
- J. A. Elliott and I. N. Olver. Hope, life, and death: A qualitative analysis of dying cancer patients’ talk about hope. *Death Studies*, 33(7):609–638, 2009. doi: 10.1080/07481180903011982.

- E. G. Engelhardt, A. H. Pieterse, P. K. Han, N. Van Duijn-Bakker, F. Cluitmans, E. Maartense, M. M. Bos, N. I. Weijl, C. J. Punt, P. Quarles Van Ufford-Mannesse, H. Sleeboom, J. E. Portielje, K. J. Van Der Hoeven, F. J. Woei-A-Jin, J. R. Kroep, H. C. De Haes, E. M. Smets, and A. M. Stiggelbout. Disclosing the Uncertainty Associated with Prognostic Estimates in Breast Cancer: Current Practices and Patients Perceptions of Uncertainty. *Medical Decision Making*, 37(3):179–192, 2017. doi: 10.1177/0272989X16670639.
- P. Fanen, A. Wohlhuter-Haddad, and A. Hinzpeter. Genetics of cystic fibrosis: CFTR mutation classifications toward genotype-based CF therapies. *International Journal of Biochemistry and Cell Biology*, 52:94–102, 2014. doi: 10.1016/j.biocel.2014.02.023.
- L. Ferrer, H. Putter, and C. Proust-Lima. Individual dynamic predictions using landmarking and joint modelling: Validation of estimators and robustness assessment. *Statistical Methods in Medical Research*, 0(0):1–18, 2018. doi: 10.1177/0962280218811837.
- J. M. Fortin, L. K. Hirota, B. E. Bond, A. M. O’Connor, and N. F. Col. Identifying patient preferences for communicating risk estimates: A descriptive pilot study. *BMC Medical Informatics and Decision Making*, 1:1–11, 2001. doi: 10.1186/1472-6947-1-1.
- J. Fosen, E. Ferkingstad, Ø. Borgan, and O. O. Aalen. Dynamic path analysis—a new approach to analyzing time-dependent covariates. *Lifetime Data Analysis*, 12(2):143–167, 2006. doi: 10.1007/s10985-006-9004-2.
- Y. Freund and R. E. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.
- J. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.
- J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010.
- F. Frost, P. Dyce, D. Nazareth, V. Malone, and M. J. Walshaw. Continuous glucose monitoring guided insulin therapy is associated with improved clinical outcomes in cystic fibrosis-related diabetes. *Journal of Cystic Fibrosis*, 17(6):798–803, 2018. doi: 10.1016/j.jcf.2018.05.005.

- F. Frost, P. Dyce, A. Ochota, S. Pandya, T. Clarke, M. J. Walshaw, and D. S. Nazareth. Cystic fibrosis-related diabetes: optimizing care with a multidisciplinary approach. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, 12:545–552, 2019. doi: 10.2147/dms0.s180597.
- F. Frost, M. J. Walshaw, and D. Nazareth. Cystic fibrosis-related diabetes: an update. *QJM: An International Journal of Medicine*, pages 1–4, 2020. doi: 10.1093/qjmed/hcaa256.
- I. R. Fulcher, E. J. Tchetgen Tchetgen, and P. L. Williams. Mediation Analysis for Censored Survival Data Under an Accelerated Failure Time Model. *Epidemiology*, 28(5):660–666, 2017. doi: 10.1097/EDE.0000000000000687.
- A. Gelman and Y.-S. Su. *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. 2018. URL <https://cran.r-project.org/package=arm>.
- P. M. George, W. Banya, N. Pareek, D. Bilton, P. Cullinan, M. E. Hodson, and N. J. Simmonds. Improved survival at low lung function in cystic fibrosis: Cohort study from 1990 to 2007. *BMJ*, 342(d1008), 2011. doi: 10.1136/bmj.d1008.
- T. A. Gerds and M. Schumacher. Efron-type measures of prediction error for survival analysis. *Biometrics*, 63:1283–1287, 2007. doi: 10.1111/j.1541-0420.2007.00832.x.
- T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13):2173–2184, 2013. doi: 10.1002/sim.5681.
- R. B. Geskus. *Data Analysis with Competing Risks and Intermediate States*. Chapman and Hall / CRC Biostatistics Series, Boca Raton, 2016. ISBN 13:978-1-4665-7035-1.
- G. Gigerenzer. Simple tools for understanding risks: from innumeracy to insight. *BMJ*, 327:741–744, 2003. doi: 10.1136/bmj.327.7417.741.
- Y. Goldberg and M. R. Kosorok. Support Vector Regression for Right Censored Data. *Electronic Journal of Statistics*, 11(1):532–569, 2017. doi: 10.1214/17-EJS1231.
- X. Gong, M. Hu, and L. Zhao. Big Data Toolsets to Pharmacometrics: Application of Machine Learning for Time-to-Event Analysis. *Clinical and Translational Science*, 11(3):305–311, 2018. doi: 10.1111/cts.12541.

- E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and Comparison of Prognostic Classification Schemes for Survival Data. *Statistics in Medicine*, 18:2529–2545, sep 1999. doi: 10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5.
- P. M. Grambsch and T. M. Therneau. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika*, 81(3):515–526, 1994. doi: 10.2307/2337123.
- A. Granados, C. L. Chan, K. L. Ode, A. Moheet, A. Moran, and R. Holl. Cystic fibrosis related diabetes: Pathophysiology, screening and diagnosis. *Journal of Cystic Fibrosis*, 18:S3–S9, 2019. doi: 10.1016/j.jcf.2019.08.016.
- I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
- R. G. Hagerty, P. N. Butow, P. M. Ellis, E. A. Lobb, S. C. Pendlebury, N. Leighl, C. MacLeod, and M. H. Tattersall. Communicating with realism and hope: Incurable cancer patients’ views on the disclosure of prognosis. *Journal of Clinical Oncology*, 23(6):1278–1288, 2005. doi: 10.1200/JCO.2005.11.138.
- P. K. Han. Conceptual, Methodological, and Ethical Problems in Communicating Uncertainty in Clinical Evidence. *Med Care Res Rev.*, 70(1 0):14S–36S, 2013. doi: 10.1177/1077558712459361.
- J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.
- F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the Yield of Medical Tests. *JAMA: The Journal of the American Medical Association*, 247(18):2543–2546, 1982. doi: 10.1001/jama.1982.03320430047030.
- T. Hastie. `gam: Generalized Additive Models`, 2018. URL <https://cran.r-project.org/package=gam>.
- T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. London: Chapman and Hall, 1990. ISBN 0412343908.
- K. M. Hayllar, S. G. Williams, A. E. Wise, S. Pouria, M. Lombard, M. E. Hodson, and D. Westaby. A prognostic model for the prediction of survival in cystic fibrosis. *Thorax*, 52(4):313–317, 1997. doi: 10.1136/thx.52.4.313.
- P. J. Heagerty and Y. Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61:92–105, 2005. doi: 10.1111/j.0006-341X.2005.030814.x.

- G. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology*, 16:117, 2016. doi: 10.1186/s12874-016-0212-5.
- G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona. A comparison of joint models for longitudinal and competing risks data, with application to an epilepsy drug randomized controlled trial. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2018a. doi: 10.1111/rssa.12348.
- G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona. JoineRML: A joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC Medical Research Methodology*, 18(1):1–14, 2018b. doi: 10.1186/s12874-018-0502-1.
- J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, M. May, and P. Brindle. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *BMJ*, 335: 136–141, 2007. doi: 10.1136/bmj.39261.471806.55.
- T. K. Ho. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–282, 1995. doi: 10.1109/ICDAR.1995.598994.
- J. W. Hogan and N. M. Laird. Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in medicine*, 16(1-3):259–272, 1997.
- R. G. Holloway, R. Gramling, and A. G. Kelly. Estimating and communicating prognosis in advanced neurologic disease. *Neurology*, 80(8):764–772, 2013. doi: 10.1212/WNL.0b013e318282509c.
- A. Horsley, S. Cunningham, and J. A. Innes, editors. *Cystic Fibrosis*. Oxford University Press, Oxford Respiratory Medicine Library, Oxford, 2nd ed. edition, 2015. ISBN 978-0-19-870294-8.
- S. Innes and S. Payne. Advanced cancer patients’ prognostic information preferences: A review. *Palliative Medicine*, 23:29–39, 2009. doi: 10.1177/0269216308098799.
- H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *Annals of Applied Statistics*, 2(3):841–860, 2008. doi: 10.1214/08-AOAS169.

- H. Ishwaran, T. A. Gerds, U. B. Kogalur, R. D. Moore, S. J. Gange, and B. M. Lau. Random survival forests for competing risks. *Biostatistics*, 15(4):757–773, 2014. doi: 10.1093/biostatistics/kxu010.
- G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, New York, 2013a. ISBN 978-1-4614-7137-0. doi: 10.1007/978-1-4614-7138-7.
- G. James, D. Witten, T. Hastie, and R. Tibshirani. Statistical Learning. In *An Introduction to Statistical Learning with Applications in R*, chapter 2. Springer, New York, 2013b. ISBN 978-1-4614-7137-0. doi: 10.1007/978-1-4614-7138-7.
- G. James, D. Witten, T. Hastie, and R. Tibshirani. Resampling Methods. In *An Introduction to Statistical Learning with Applications in R*, chapter 5. Springer, New York, 2013c. ISBN 978-1-4614-7137-0. doi: 10.1007/978-1-4614-7138-7.
- N. P. Jewell and J. P. Nielsen. A framework for consistent prediction rules based on markers. *Biometrika*, 80(1):153–164, 1993. doi: 10.1093/biomet/80.1.153.
- A. P. Kalogeropoulos, J. Thankachen, J. Butler, and J. C. Fang. Diuretic and renal effects of spironolactone and heart failure hospitalizations: a TOPCAT Americas analysis. *European Journal of Heart Failure*, 22(9):1600–1610, 2020. doi: 10.1002/ejhf.1917.
- J. D. Y. Kang and J. L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539, 2007. doi: 10.1214/07-STS227.
- J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(24):1–12, 2018. doi: 10.1186/s12874-018-0482-1.
- K. Kayani, R. Mohammed, and H. Mohiaddin. Cystic fibrosis-related diabetes. *Frontiers in Endocrinology*, 9, 2018. doi: 10.3389/fendo.2018.00020.
- R. H. Keogh. GitHub page: ruthkeogh, 2018. URL https://github.com/ruthkeogh/landmark_CF.
- R. H. Keogh, D. Bilton, R. Cosgriff, D. Kavanagh, O. Rayner, and P. M. Sedgwick. Results from an online survey of adults with cystic fibrosis: Accessing and using life expectancy information. *PLoS ONE*, 14(4):1–13, 2019a. doi: 10.1371/journal.pone.0213639.

- R. H. Keogh, S. R. Seaman, J. K. Barrett, D. Taylor-Robinson, and R. Szczesniak. Dynamic Prediction of Survival in Cystic Fibrosis: A Landmarking Analysis Using UK Patient Registry Data. *Epidemiology*, 30(1):29–37, 2019b. doi: 10.1097/EDE.0000000000000920.
- R. H. Keogh, S. R. Seaman, J. M. Gran, and S. Vansteelandt. Simulating longitudinal data from marginal structural models using the additive hazard model. *Biometrical Journal*, pages 1–16, 2021. doi: 10.1002/bimj.202000040.
- E. Kerem, L. Viviani, A. Zolin, S. MacNeill, E. Hatziagorou, H. Ellemunter, P. Drevinek, V. Gulmans, U. Krivec, and H. Olesen. Factors associated with FEV1 decline in cystic fibrosis: Analysis of the ECFS patient registry. *European Respiratory Journal*, 43(1):125–133, 2014. doi: 10.1183/09031936.00166412.
- B. E. Kiely, M. H. N. Tattersall, and M. R. Stockler. Certain death in uncertain time: Informing hope by quantifying a best case scenario. *Journal of Clinical Oncology*, 28(16):2802–2804, 2010. doi: 10.1200/JCO.2009.27.3326.
- B. E. Kiely, G. McCaughan, S. Christodoulou, P. J. Beale, P. Grimison, J. Trotman, M. H. Tattersall, and M. R. Stockler. Using scenarios to explain life expectancy in advanced cancer: Attitudes of people with a cancer experience. *Supportive Care in Cancer*, 21(2):369–376, 2013. doi: 10.1007/s00520-012-1526-4.
- C. Koch, M. Rainisio, U. Madessani, H. K. Harms, M. E. Hodson, G. Mastella, S. G. McKenzie, J. Navarro, and B. Strandvik. Presence of cystic fibrosis-related diabetes mellitus is tightly linked to poor lung function in patients with cystic fibrosis: Data from the European epidemiologic registry of cystic fibrosis. *Pediatric Pulmonology*, 32(5):343–350, 2001. doi: 10.1002/ppul.1142.
- M. W. Konstan, J. S. Wagener, D. R. VanDevanter, D. J. Pasta, A. Yegin, L. Rasouliyan, and W. J. Morgan. Risk factors for rate of decline in FEV1 in adults with cystic fibrosis. *Journal of Cystic Fibrosis*, 11(5):405–411, 2012. doi: 10.1016/j.jcf.2012.03.009.
- E. B. Laber and S. A. Murphy. Adaptive confidence intervals for the test error in classification. *Journal of the American Statistical Association*, 106(495):904–913, 2011. doi: 10.1198/jasa.2010.tm10053.
- A. Lafourcade, M. His, L. Baglietto, M. C. Boutron-Ruault, L. Dossus, and V. Rondeau. Factors associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: The French E3N cohort. *BMC Cancer*, 18(1):171, 2018. doi: 10.1186/s12885-018-4076-4.

- T. Lange and J. V. Hansen. Direct and indirect effects in a survival context. *Epidemiology*, 22(4):575–581, 2011. doi: 10.1097/EDE.0b013e31821c680c.
- S. Lannig, B. Thorsteinsson, J. Nerup, and C. Koch. Influence of the development of diabetes mellitus on clinical status in patients with cystic fibrosis. *European Journal of Pediatrics*, 151(9):684–687, 1992. doi: 10.1007/BF01957574.
- L. Lapointe-Shaw, Z. Bouck, N. A. Howell, T. Lange, A. Orchanian-Cheff, P. C. Austin, N. M. Ivers, D. A. Redelmeier, and C. M. Bell. Mediation analysis with a time-to-event outcome: A review of use and reporting in healthcare research. *BMC Medical Research Methodology*, 18(1):118, 2018. doi: 10.1186/s12874-018-0578-7.
- C. Lawson and R. Hanson. *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics, 1995. doi: 10.1137/1.9781611971217.
- E. LeDell. Scalable Super Learning. In P. Buehlmann, P. Drineas, M. Kane, and M. J. van der Laan, editors, *Handbook of Big Data*, chapter 19, pages 339–357. CRC Press, Boca Raton, 2016. ISBN 13:978-1-4822-4907-1.
- C. Lee, J. Yoon, and M. van der Schaar. Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis with Competing Risks based on Longitudinal Data. *IEEE Transactions on Biomedical Engineering*, XX(X):1–12, 2019. doi: 10.1109/TBME.2019.2909027.
- C. Lehoux Dubois, V. Boudreau, F. Tremblay, A. Lavoie, Y. Berthiaume, R. Rabasa-Lhoret, and A. Coriati. Association between glucose intolerance and bacterial colonisation in an adult population with cystic fibrosis, emergence of *Stenotrophomonas maltophilia*. *Journal of Cystic Fibrosis*, 16(3):418–424, 2017. doi: 10.1016/j.jcf.2017.01.018.
- C. Lewis, S. M. Blackman, A. Nelson, E. Oberdorfer, D. Wells, J. Dunitz, W. Thomas, and A. Moran. Diabetes-related mortality in adults with cystic fibrosis: Role of genotype and sex. *American Journal of Respiratory and Critical Care Medicine*, 191(2):194–200, 2015. doi: 10.1164/rccm.201403-0576OC.
- H. Lin, C. E. McCulloch, and S. T. Mayne. Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine*, 21(16):2369–2382, 2002. doi: 10.1002/sim.1179.
- T. G. Liou, F. R. Adler, S. C. Fitzsimmons, B. C. Cahill, J. R. Hibbs, and B. C. Marshall. Predictive 5-Year Survivorship Model of Cystic Fibrosis. *American Journal of Epidemiology*, 153(4):345–352, 2001.

- T. G. Liou, C. Kartsonaki, R. H. Keogh, and F. R. Adler. Evaluation of a five-year predicted survival model for cystic fibrosis in later time periods. *Scientific Reports*, 10:6602, 2020. doi: 10.1038/s41598-020-63590-8.
- R. J. A. Little. Regression with Missing X's: A Review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992. URL <https://escholarship.org/uc/item/84j7c2w5>.
- D. P. MacKinnon, G. Warsi, and J. H. Dwyer. A simulation study of mediated effect measures. *Multivariate Behav Res*, 30(1):41–62, 1995. doi: 10.1207/s15327906mbr3001_3.
- J. D. Malley, J. Kruppa, A. Dasgupta, K. G. Malley, and A. Ziegler. Probability Machines: Consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine*, 51:74–81, 2012. doi: 10.3414/ME00-01-0052.
- B. C. Marshall, S. M. Butler, M. Stoddard, A. Moran, T. G. Liou, and W. J. Morgan. Epidemiology of cystic fibrosis-related diabetes. *Journal of Pediatrics*, 146(5):681–687, 2005. doi: 10.1016/j.jpeds.2004.12.039.
- T. Martinussen and T. H. Scheike. *Dynamic Regression models for survival data*. Springer-Verlag, NY, New York, 2006. ISBN 978-0-387-20274-7. doi: 10.1007/0-387-33960-4.
- T. Martinussen and S. Vansteelandt. On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime Data Analysis*, 19(3):279–296, 2013. doi: 10.1007/s10985-013-9242-z.
- N. Mayer-Hamblett, M. Rosenfeld, J. Emerson, C. H. Goss, and M. L. Aitken. Developing cystic fibrosis lung transplant referral criteria using predictors of 2-year mortality. *American Journal of Respiratory and Critical Care Medicine*, 166(12 I):1550–1555, 2002. doi: 10.1164/rccm.200202-087OC.
- M. Maziarz, P. Heagerty, T. Cai, and Y. Zheng. On longitudinal prediction with time-to-event outcome: Comparison of modeling options. *Biometrics*, 73:83–93, 2017. doi: 10.1111/biom.12562.
- P. McCullagh and J. A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman and Hall, 1989. ISBN 9780412317606.
- MDCalc. BODE Index for COPD Survival, 2021. URL <https://www.mdcalc.com/bode-index-copd-survival>.

- N. Meinshausen. Quantile Regression Forests. *Journal of Machine Learning Research*, 7:983–99, 2006.
- Memorial Sloan Kettering Cancer Center. Overall Survival Probability Following Surgery, 2021. URL https://www.mskcc.org/nomograms/colorectal/overall_survival_probability.
- D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, 2018. URL <https://cran.r-project.org/package=e1071>.
- C. E. Milla, J. Billings, and A. Moran. Diabetes is associated with dramatically decreased survival in female but not male subjects with cystic fibrosis. *Diabetes Care*, 28(9):2141–2144, 2005. doi: 10.2337/diacare.28.9.2141.
- U. B. Mogensen, H. Ishwaran, and T. A. Gerds. Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *Journal of Statistical Software*, 50(11), 2012. doi: 10.18637/jss.v050.i11.
- K. Mohan, H. Miller, P. Dyce, R. Grainger, R. Hughes, J. Vora, M. Ledson, and M. Walshaw. Mechanisms of glucose intolerance in cystic fibrosis. *Diabetic Medicine*, 26(6):582–588, 2009. doi: 10.1111/j.1464-5491.2009.02738.x.
- A. Moheet and A. Moran. CF-related diabetes: Containing the metabolic miscreant of cystic fibrosis. *Pediatric Pulmonology*, 52(April):S37–S43, 2017. doi: 10.1002/ppul.23762.
- A. Moran, J. Dunitz, B. Nathan, A. Saeed, B. Holme, and W. Thomas. Cystic Fibrosis–Related Diabetes: Current Trends in Prevalence, Incidence, and Mortality. *Diabetes Care*, 32(9):1626–1631, 2009. doi: 10.2337/dc09-0586.
- A. Moran, K. Pillay, D. Becker, A. Granados, S. Hameed, and C. L. Acerini. ISPAD Clinical Practice Consensus Guidelines 2018: Management of cystic fibrosis-related diabetes in children and adolescents. *Pediatric Diabetes*, 19(Suppl. 27): 64–74, 2018. doi: 10.1111/pedi.12732.
- T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019. doi: 10.1002/sim.8086.
- S. I. Mozumder, P. W. Dickman, M. J. Rutherford, and P. C. Lambert. InterPreT cancer survival: A dynamic web interactive prediction cancer survival tool for health-care professionals and cancer epidemiologists. *Cancer Epidemiology*, 56: 46–52, 2018. doi: 10.1016/j.canep.2018.07.009.

- K. M. Mullen and I. H. M. van Stokkum. nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS), 2012. URL <https://cran.r-project.org/package=nnls>.
- D. J. Muscatello, A. Searles, R. Macdonald, and L. Jorm. Communicating population health statistics through graphs: A randomised controlled trial of graph design interventions. *BMC Medicine*, 4(33), 2006. doi: 10.1186/1741-7015-4-33.
- G. Naik, H. Ahmed, and A. G. K. Edwards. Communicating risk to patients and the public. *The British Journal of General Practice*, 62(597):213–6, 2012. doi: 10.3399/bjgp12X636236.
- National Institute for Health and Care Excellence (NICE). Cystic fibrosis: diagnosis and management, 2017. URL <https://www.nice.org.uk/guidance/ng78/chapter/Recommendations#annual-and-routine-reviews>.
- NHS Oxford University Hospitals. Oxford Adult Cystic Fibrosis Centre : Annual Review, 2019. URL <https://www.ouh.nhs.uk/services/departments/specialist-medicine/respiratory-medicine/cystic-fibrosis/review.aspx>.
- L. Nkam, J. Lambert, A. Latouche, G. Bellis, P. Burgel, and M. Hocine. A 3-year prognostic score for adults with cystic fibrosis. *Journal of Cystic Fibrosis*, 16(6):702–708, 2017. doi: 10.1016/j.jcf.2017.03.004.
- M. Ochman, M. Latos, M. Urlik, T. Stącel, M. Necki, Z. Tatoj, F. Zawadzki, M. Wajda-Pokrontka, P. Przybyłowski, and M. Zembala. Cystic Fibrosis: From Qualification to Lung Transplantation, a Single Center Experience. *Annals of Transplantation*, 24:185–190, 2019. doi: 10.12659/AOT.914328.
- Office for National Statistics. Past and projected period and cohort life tables, 2018-based, UK: 1981 to 2068, 2021. URL <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/pastandprojecteddatafromtheperiodandcohortlifetables/1981to2068>.
- I. N. Olver. Evolving definitions of hope in oncology. *Current Opinion in Supportive and Palliative Care*, 6(2):236–41, 2012. doi: 10.1097/SPC.0b013e3283528d0c.
- S. M. O’Riordan and A. Moran. Cystic fibrosis-related diabetes. In A. Horsley, S. Cunningham, and J. A. Innes, editors, *Cystic Fibrosis*, chapter 8. Oxford University Press, Oxford Respiratory Medicine Library, Oxford, second edition, 2015. ISBN 978-0-19-870294-8.

- E. Paige, J. Barrett, D. Stevens, R. H. Keogh, M. J. Sweeting, I. Nazareth, I. Petersen, and A. M. Wood. Landmark Models for Optimizing the Use of Repeated Measurements of Risk Factors in Electronic Health Records to Predict Future Disease Risk. *American Journal of Epidemiology*, 187(7):1530–1538, 2018. doi: 10.1093/aje/kwy018.
- S. M. Parker, J. M. Clayton, K. Hancock, S. Walder, P. N. Butow, S. Carrick, D. Currow, D. Ghersi, P. Glare, R. Hagerty, and M. H. Tattersall. A Systematic Review of Prognostic/End-of-Life Communication with Adults in the Advanced Stages of a Life-Limiting Illness: Patient/Caregiver Preferences for the Content, Style, and Timing of Information. *Journal of Pain and Symptom Management*, 34(1):81–93, 2007. doi: 10.1016/j.jpainsymman.2006.09.035.
- E. T. Parner and P. K. Andersen. Regression analysis of censored data using pseudo-observations. *The Stata Journal*, 10(3):408–422, 2010.
- S. Parvez, K. Abdel-Kader, M. K. Song, and M. Unruh. Conveying uncertainty in prognosis to patients with ESRD. *Blood Purification*, 39(1-3):58–64, 2015. doi: 10.1159/000368954.
- J. Pearl. Direct and Indirect Effects. *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 411–420, 2001. URL <https://dl.acm.org/citation.cfm?id=2074022.2074073>.
- M. J. Pencina and R. B. D’Agostino. Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. *Statistics in Medicine*, 23:2109–2123, 2004. doi: 10.1002/sim.1802.
- M. J. Pencina, R. B. D’Agostino, and L. Song. Quantifying discrimination of Framingham risk functions with different survival C statistics. *Statistics in Medicine*, 31(15):1543–1553, 2012. doi: 10.1002/sim.4508.
- P. Philipson, I. Sousa, P. Diggle, P. Williamson, R. Kolamunnage-Dona, R. Henderson, and G. Hickey. `joineR`: Joint Modelling of Repeated Measurements and Time-to-Event Data, 2018. URL <https://github.com/graemeleehickey/joineR>.
- J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team. `nlme`: Linear and Nonlinear Mixed Effects Models, 2018. URL <https://cran.r-project.org/package=nlme>.
- D. Pitocco, L. Fuso, E. G. Conte, F. Zaccardi, C. Condoluci, G. Scavone, R. A. Incalzi, and G. Ghirlanda. The diabetic lung—a new target organ? *The Review of Diabetic Studies*, 9(1):23–35, 2012. doi: 10.1900/RDS.2012.9.23.

- E. C. Polley and M. J. van der Laan. Super Learning for Right-Censored Data. In *Targeted Learning: Causal Inference for Observational and Experimental Data*, chapter 16. Springer, New York, 2011.
- E. C. Polley, S. Rose, and M. J. van der Laan. Super Learning. In *Targeted Learning: Causal Inference for Observational and Experimental Data*, chapter 3. Springer, New York, 2011. ISBN 978-1-4419-9781-4.
- E. C. Polley, E. LeDell, C. Kennedy, and M. van der Laan. SuperLearner: Super Learner Prediction, 2018. URL <https://cran.r-project.org/package=SuperLearner>.
- M. Price, R. Cameron, and P. Butow. Communicating risk information: The influence of graphical display format on quantitative information perception—Accuracy, comprehension and preferences. *Patient Education and Counseling*, 69(1-3):121–128, 2007. doi: 10.1016/j.pec.2007.08.006.
- P. Probst and A.-L. Boulesteix. To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, 18:1–18, 2018.
- C. Proust-Lima and P. Blanche. Dynamic Predictions. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorisch, F. Ruggeri, and J. Teugels, editors, *Wiley StatsRef: Statistics Reference Online*. Wiley Online Library, 2016. doi: 10.1002/9781118445112.stat07876.
- C. Proust-Lima, M. Séne, J. M. G. Taylor, and H. Jacqmin-Gadda. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research*, 23(1):74–90, 2014. doi: 10.1177/0962280212445839.
- C. Proust-Lima, V. Philipps, and B. Liqueur. Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. *Journal of Statistical Software*, 78(2), 2017. doi: 10.18637/jss.v078.i02.
- J. Pruitt and J. Grundin. Personas : Practice and Theory. *Proceedings of the 2003 conference on Designing for user experiences*, pages 1–15, 2003. doi: 10.1145/997078.997089.
- H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in Biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430, 2007. doi: 10.1002/sim.2712.
- B. S. Quon and M. L. Aitken. Cystic Fibrosis: What to Expect Now in the Early Adult Years. *Paediatric Respiratory Reviews*, 13:206–214, 2012. doi: 10.1016/j.prrv.2012.03.005.

- R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, 2020. URL <https://www.r-project.org/>.
- M. S. Rahman, G. Ambler, B. Choodari-Oskoei, and R. Z. Omar. Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC Medical Research Methodology*, 17(60), 2017. doi: 10.1186/s12874-017-0336-2.
- T. Rakow, R. J. Wright, C. Bull, and D. J. Spiegelhalter. Simple and multistate survival curves: Can people learn to use them? *Medical Decision Making*, 32(6):792–804, 2012. doi: 10.1177/0272989X12451057.
- F. Ratjen, C. Hug, G. Marigowda, S. Tian, X. Huang, S. Stanojevic, C. E. Milla, P. D. Robinson, D. Waltz, J. C. Davies, M. Rosenfeld, T. Starner, G. Retsch-Bogart, J. Chmiel, D. Orenstein, C. Milla, R. Rubenstein, S. Walker, A. Cornell, F. Asfour, P. Black, J. Colombo, D. Froh, S. McColley, F. Ruiz, D. Quintero, A. Casey, G. Mueller, P. Flume, F. Livingston, M. Rock, B. O’Sullivan, H. Schmidt, T. Lahiri, J. McNamara, A. Chidekel, L. Sass, T. Keens, D. Schaeffer, M. Solomon, M. Chilvers, L. Lands, S. Junge, M. Griese, D. Staab, T. Pressler, S. van Koningsburggen-Rietschel, L. Naehrlich, A. Reid, I. Balfour-Lynn, D. Urquhart, T. Lee, A. Munck, I. S. Gaudelus, C. De Boeck, P. Reix, A. Malfroot, S. Bui, H. Selvadurai, P. Robinson, C. Wainwright, B. Clements, J. Hilton, and L. Hjelte. Efficacy and safety of lumacaftor and ivacaftor in patients aged 6–11 years with cystic fibrosis homozygous for F508del-CFTR: a randomised, placebo-controlled phase 3 trial. *The Lancet Respiratory Medicine*, 5(7):557–567, 2017. doi: 10.1016/S2213-2600(17)30215-1.
- L. Regard, C. Martin, G. Chassagnon, and P. R. Burgel. Acute and chronic non-pulmonary complications in adults with cystic fibrosis. *Expert Review of Respiratory Medicine*, 13(1):23–38, 2019. doi: 10.1080/17476348.2019.1552832.
- L. Richiardi, R. Bellocco, and D. Zugna. Mediation analysis in epidemiology: Methods, interpretation and bias. *International Journal of Epidemiology*, 42(5):1511–1519, 2013. doi: 10.1093/ije/dyt127.
- R. M. Ripley, A. L. Harris, and L. Tarassenko. Non-linear survival analysis using neural networks. *Statistics in Medicine*, 23:825–842, 2004. doi: 10.1002/sim.1655.
- D. Rizopoulos. JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. *Journal of Statistical Software*, 35(9), 2010.

- D. Rizopoulos. Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. *Biometrics*, 67(3):819–829, 2011. doi: 10.1111/j.1541-0420.2010.01546.x.
- D. Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R*. Chapman and Hall / CRC Biostatistics Series, Boca Raton, 2012. ISBN 9781439872864.
- D. Rizopoulos. The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data using MCMC. *Journal of Statistical Software*, 72(7), 2016. doi: 10.18637/jss.v072.i07.
- D. Rizopoulos, G. Molenberghs, and E. Lesaffre. Dynamic Predictions with Time-Dependent Covariates in Survival Analysis using Joint Modeling and Landmarking. *Biometrical Journal*, 59(6):1261–1276, 2017.
- J. M. Robins and S. Greenland. Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology*, 3(2):143–155, 1992.
- J. M. Robins and T. S. Richardson. Alternative graphical causal models and the identification of direct effects. In *Causality and psychopathology: finding the determinants of disorders and their cures*, pages 103–158. Oxford University Press, Oxford, 2011. ISBN 9780199754649. doi: DOI:10.1093/oso/9780199754649.003.0011.
- N. J. Ronan, J. S. Elborn, and B. J. Plant. Current and emerging comorbidities in cystic fibrosis. *La Presse Médicale*, 46(6, Part 2):e125–e138, 2017. doi: 10.1016/j.lpm.2017.05.011.
- A. Rotnitzky and J. M. Robins. Inverse Probability Weighted Estimation in Survival Analysis. In P. Armitage and T. Coulton, editors, *Encyclopedia of Biostatistics*. Wiley, New York, 2nd edition, 2005.
- S. M. Rowe, S. Miller, and E. J. Sorscher. Cystic fibrosis. *New England Journal of Medicine*, 352(19):1992–2001, 2005. doi: 10.1056/NEJMra043184.
- P. Royston and D. G. Altman. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13(33), 2013. doi: 10.1186/1471-2288-13-33.
- G. S. Sawicki, D. E. Sellers, and W. M. Robinson. High Treatment Burden in Adults with Cystic Fibrosis: Challenges to Disease Self-Management. *Journal of Cystic Fibrosis*, 8(2):91–96, 2009. doi: 10.1016/j.jcf.2008.09.007.

- M. Schapira, A. Nattinger, and C. McHorney. Frequency or probability? A qualitative study of risk communication formats used in health care. *Med Decis Making*, 21(6):459–467, 2001.
- D. K. Schlüter, R. Griffiths, A. Adam, A. Akbari, M. L. Heaven, S. Paranjothy, A. M. Nybo Andersen, S. B. Carr, T. Pressler, P. J. Diggle, and D. Taylor-Robinson. Impact of cystic fibrosis on birthweight: A population based study of children in Denmark and Wales. *Thorax*, 74(5):447–454, 2019. doi: 10.1136/thoraxjnl-2018-211706.
- M. R. Segal. Regression Trees for Censored Data. *Biometrics*, 44(1):35–47, 1988.
- M. R. Segal. Tree-Structured Survival Analysis. *Epidemiology*, 8(4):344–346, 1997.
- G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.
- I. Shpitser and E. T. Tchetgen. Causal inference with a graphical hierarchy of interventions. *Annals of Statistics*, 44(6):2433–2466, 2016. doi: 10.1214/15-AOS1411.
- N. J. Simmonds. Ageing in cystic fibrosis and long-term survival. *Paediatric Respiratory Reviews*, 14(SUPPL.1):6–9, 2013. doi: 10.1016/j.prrv.2013.01.007.
- J. D. Singer and J. B. Willett. It’s About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. *Journal of Educational Statistics*, 18(2):155–195, 1993. doi: 10.3102/10769986018002155.
- G. D. Smith. Epidemiology, epigenetics and the ‘Gloomy Prospect’: Embracing randomness in population health research and practice. *International Journal of Epidemiology*, 40(3):537–562, 2011. doi: 10.1093/ije/dyr117.
- D. J. Spiegelhalter. Risk and Uncertainty Communication. *Annual Review of Statistics and Its Application*, 4(1):31–60, 2017. doi: 10.1146/annurev-statistics-010814-020148.
- D. J. Spiegelhalter, M. Pearson, and I. Short. Visualizing uncertainty about the future. *Science*, 333(6048):1393–1400, 2011. doi: 10.1126/science.1191181.
- G. E. Stanford, K. Dave, and N. J. Simmonds. Pulmonary Exacerbations in Adults With Cystic Fibrosis: A Grown-up Issue in a Changing Cystic Fibrosis Landscape. *Chest*, 159(1):93–102, 2021. doi: 10.1016/j.chest.2020.09.084.
- S. Stanojevic, J. Sykes, A. L. Stephenson, S. D. Aaron, and G. A. Whitmore. Development and external validation of 1- and 2-year mortality prediction models in cystic fibrosis. *European Respiratory Journal*, 54(3), 2019. doi: 10.1183/13993003.00224-2019.

- A. J. Steele, S. C. Denaxas, A. D. Shah, H. Hemingway, and N. M. Luscombe. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS ONE*, 13(8):e0202344, 2018. doi: <https://doi.org/10.1371/journal.pone.0202344>.
- A. L. Stephenson, L. A. Mannik, S. Walsh, M. Brotherwood, R. Robert, P. B. Darling, R. Nisenbaum, J. Moerman, and S. Stanojevic. Longitudinal trends in nutritional status and the relation between lung function and BMI in cystic fibrosis: A population-based cohort study. *American Journal of Clinical Nutrition*, 97(4):872–877, 2013. doi: [10.3945/ajcn.112.051409](https://doi.org/10.3945/ajcn.112.051409).
- E. W. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation and Updating*. Springer, New York, 2009. ISBN 978-0-387-77244-8.
- E. W. Steyerberg and F. E. Harrell. Prediction models need appropriate internal, internal-external, and external validation. *Journal of Clinical Epidemiology*, 69: 245–247, 2016. doi: [10.1016/j.jclinepi.2015.04.005](https://doi.org/10.1016/j.jclinepi.2015.04.005).
- E. W. Steyerberg, F. E. Harrell, G. J. Borsboom, M. Eijkemans, Y. Vergouwe, and J. F. Habbema. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 54: 774–781, 2001. doi: [10.1016/S0895-4356\(01\)00341-9](https://doi.org/10.1016/S0895-4356(01)00341-9).
- E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. A. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, 21(1):128–138, 2010. doi: [10.1097/EDE.0b013e3181c30fb2](https://doi.org/10.1097/EDE.0b013e3181c30fb2).
- S. Strohmaier, K. Røysland, R. Hoff, O. Borgan, T. R. Pedersen, and O. O. Aalen. Dynamic path analysis - a useful tool to investigate mediation processes in clinical survival trials. *Statistics in Medicine*, 34:3866–3887, 2015. doi: [10.1002/sim.6598](https://doi.org/10.1002/sim.6598).
- K. Suresh, J. M. Taylor, D. E. Spratt, S. Daignault, and A. Tsodikov. Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical Journal*, 59(6):1277–1300, 2017. doi: [10.1002/bimj.201600235](https://doi.org/10.1002/bimj.201600235).
- R. Szczesniak, S. L. Heltshe, S. Stanojevic, and N. Mayer-Hamblett. Use of FEV1 in cystic fibrosis epidemiologic studies and clinical trials: A statistical perspective for the clinical researcher. *Journal of Cystic Fibrosis*, 16(3):318–326, 2017. doi: [10.1016/j.jcf.2017.01.002](https://doi.org/10.1016/j.jcf.2017.01.002).

- R. D. Szczesniak, C. Brokamp, W. Su, G. L. Mcphail, J. Pestian, and J. P. Clancy. Improving Detection of Rapid Cystic Fibrosis Disease Progression—Early Translation of a Predictive Algorithm Into a Point-of-Care Tool. *IEEE Journal of Translational Engineering in Health and Medicine*, 7:1–8, 2019. doi: 10.1109/JTEHM.2018.2878534.
- C. Taylor and S. Connolly. Gastrointestinal disease and nutrition. In A. Horsley, S. Cunningham, and J. A. Innes, editors, *Cystic Fibrosis*, chapter 7. Oxford University Press, Oxford Respiratory Medicine Library, Oxford, second edition, 2015. ISBN 978-0-19-870294-8.
- D. Taylor-Robinson and F. Kee. Precision public health — the Emperor’s new clothes. *International Journal of Epidemiology*, 48(1):1–6, 2019. doi: 10.1093/ije/dyy184.
- D. Taylor-Robinson, M. Whitehead, F. Diderichsen, H. V. Olesen, T. Pressler, R. L. Smyth, and P. Diggle. Understanding the natural progression in %FEV1 decline in patients with cystic fibrosis: a longitudinal study. *Thorax*, 67(10): 860–866, 2012. doi: 10.1136/thoraxjnl-2011-200953.
- D. Taylor-Robinson, O. Archangelidi, S. B. Carr, R. Cosgriff, E. Gunn, R. H. Keogh, A. MacDougall, S. Newsome, D. K. Schluter, S. Stanojevic, and D. Bilton. Data Resource Profile: The UK Cystic Fibrosis Registry. *International Journal of Epidemiology*, pages 1–7, 2017. doi: 10.1093/ije/dyx196.
- D. Taylor-Robinson, D. K. Schlüter, P. J. Diggle, and J. K. Barrett. Explaining the sex effect on survival in cystic fibrosis: A joint modeling study of UK registry data. *Epidemiology*, 31(6):872–879, 2020. doi: 10.1097/EDE.0000000000001248.
- T. Therneau and P. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York, 2000. ISBN 0-387-98784-3.
- T. M. Therneau. A Package for Survival Analysis in S, 2015. URL <https://cran.r-project.org/package=survival>.
- W. A. Thompson. On the treatment of grouped observations in life studies. *Biometrics*, 33:463–470, 1977.
- R. Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.
- R. J. Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16:385–395, 1997.

- L. J. Trevena, B. J. Zikmund-Fisher, A. Edwards, W. Gaissmaier, M. Galesic, P. K. Han, J. King, M. L. Lawson, S. K. Linder, I. Lipkus, E. Ozanne, E. Peters, D. Timmermans, and S. Woloshin. Presenting quantitative information about decision outcomes: A risk communication primer for patient decision aid developers. *BMC Medical Informatics and Decision Making*, 13(SUPPL. 2):S7, 2013. doi: 10.1186/1472-6947-13-S2-S7.
- A. A. Tsiatis and M. Davidian. Joint Modeling of Longitudinal and Time-to-Event Data: An Overview. *Statistica Sinica*, 14:809–834, 2004.
- G. Tutz and M. Schmid. *Modeling Discrete Time-to-Event Data*. Springer International Publishing, Switzerland, 2016. ISBN 978-3-319-28156-8. doi: 10.1007/978-3-319-28158-2.
- UK Cystic Fibrosis Registry. UK Cystic Fibrosis Registry Annual Data Report 2018. Technical report, Cystic Fibrosis Trust, London, 2019. URL <https://www.cysticfibrosis.org.uk/the-work-we-do/uk-cf-registry>.
- UK Cystic Fibrosis Registry. UK Cystic Fibrosis Registry Annual Data Report 2019. Technical report, Cystic Fibrosis Trust, London, 2020. URL https://www.cysticfibrosis.org.uk/sites/default/files/2020-12/2019RegistryAnnualDataReport_Sep2020.pdf.
- University of Cambridge. Predict breast cancer, 2021a. URL <https://breast.predict.nhs.uk/>.
- University of Cambridge. Predict prostate, 2021b. URL <https://prostate.predict.nhs.uk/>.
- H. Uno, T. Cai, M. J. Pencina, R. B. D’Agostino, and L. J. Wei. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Statistics in Medicine*, 30(10):1105–1117, 2011. doi: 10.1002/sim.4154.
- L. Valeri and T. J. VanderWeele. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, 18(2):137–150, 2013. doi: 10.1037/a0031034.
- V. Van Belle, K. Pelckmans, S. Van Huffel, and J. A. Suykens. Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53:107–118, 2011. doi: 10.1016/j.artmed.2011.06.006.

- A. M. van der Bles, S. van der Linden, A. L. Freeman, and D. J. Spiegelhalter. The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14):7672–7683, 2020. doi: 10.1073/pnas.1913678117.
- M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2003. URL <http://www.bepress.com/ucbbiostat/paper130>.
- M. J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, 2011. ISBN 978-1-4419-9781-4.
- M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.
- H. C. van Houwelingen. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine*, 19:3401–3415, 2000. doi: 10.1002/1097-0258(20001230)19:24<3401::aid-sim554>3.0.co;2-2.
- H. C. van Houwelingen. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 2007. doi: 10.1111/j.1467-9469.2006.00529.x.
- H. C. van Houwelingen and H. Putter. Dynamic predicting by landmarking as an alternative for multi-state modeling: An application to acute lymphoid leukemia data. *Lifetime Data Analysis*, 14:447–463, 2008. doi: 10.1007/s10985-008-9099-8.
- H. C. van Houwelingen and H. Putter. *Dynamic Prediction in Clinical Survival Analysis*. CRC Press, Boca Raton, 2012. ISBN 978-1-4398-3533-3.
- L. Van Sambeek, E. S. Cowley, D. K. Newman, and R. Kato. Sputum glucose and glycemc control in cystic fibrosis-related diabetes: A cross-sectional study. *PLoS ONE*, 10(3):1–10, 2015. doi: 10.1371/journal.pone.0119938.
- T. J. VanderWeele. Causal mediation analysis with survival data. *Epidemiology*, 22(4):582–585, 2011. doi: 10.1097/EDE.0b013e31821db37e.Causal.
- T. J. VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, New York, 2015. ISBN 978-0199325870.

- T. J. VanderWeele. Mediation Analysis: A Practitioner’s Guide. *Annual Review of Public Health*, 37:17–32, 2016. doi: 10.1146/annurev-publhealth-032315-021402.
- T. J. VanderWeele and S. Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2(4):457–468, 2009. doi: 10.4310/sii.2009.v2.n4.a7.
- T. J. VanderWeele and S. Vansteelandt. Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2(1):95–115, 2013. doi: 10.1515/em-2012-0010.
- T. J. VanderWeele, S. Vansteelandt, and J. M. Robins. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, 25(2):300–306, 2014. doi: 10.1097/EDE.0000000000000034.
- S. Vansteelandt and R. M. Daniel. Interventional Effects for Mediation Analysis with Multiple Mediators. *Epidemiology*, 28(2):258–265, 2017. doi: 10.1097/EDE.0000000000000596.
- S. Vansteelandt, M. Linder, S. Vandenberghe, J. Steen, and J. Madsen. Mediation analysis of time-to-event endpoints accounting for repeatedly measured mediators subject to time-varying confounding. *Statistics in Medicine*, 38(24):4828–4840, 2019. doi: 10.1002/sim.8336.
- W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0. URL <http://www.stats.ox.ac.uk/pub/MASS4>.
- D. M. Vock, J. Wolfson, S. Bandyopadhyay, G. Adomavicius, P. E. Johnson, G. Vazquez-Benitez, and P. J. O’Connor. Adapting machine learning techniques to censored time-to-event health record data: a general-purpose approach using inverse probability of censoring weighting. *J Biomed Inform*, 61:119–131, 2016. doi: 10.1016/j.jbi.2016.03.009.
- E. F. Vonesh, D. E. Schaubel, W. Hao, and A. J. Collins. Statistical methods for comparing mortality among ESRD patients: Examples of regional/international variations. *Kidney International*, 57(Suppl. 74), 2000. doi: 10.1046/j.1523-1755.2000.07405.x.
- S. Vosbergen, J. M. Mulder-Wiggers, J. P. Lacroix, H. M. Kemps, R. A. Kraaijenhagen, M. W. Jaspers, and N. Peek. Using personas to tailor educational messages to the preferences of coronary heart disease patients. *Journal of Biomedical Informatics*, 53:100–112, 2015. doi: 10.1016/j.jbi.2014.09.004.

- V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer US, Boston, 2005. ISBN 978-0-387-00152-4. doi: 10.1007/b106715.
- C. E. Wainwright, J. S. Elborn, B. W. Ramsey, G. Marigowda, X. Huang, M. Cipolli, C. Colombo, J. C. Davies, K. De Boeck, P. A. Flume, M. W. Konstan, S. A. McColley, K. McCoy, E. F. McKone, A. Munck, F. Ratjen, S. M. Rowe, D. Waltz, and M. P. Boyle. Lumacaftor–Ivacaftor in Patients with Cystic Fibrosis Homozygous for Phe508del CFTR. *New England Journal of Medicine*, 373(3):220–231, 2015. doi: 10.1056/NEJMoa1409547.
- V. Waters, S. Stanojevic, E. G. Atenafu, A. Lu, Y. Yau, E. Tullis, and F. Ratjen. Effect of pulmonary exacerbations on long-term lung function decline in cystic fibrosis. *European Respiratory Journal*, 40(1):61–66, 2012. doi: 10.1183/09031936.00159111.
- M. R. Weiser, M. Goenen, J. F. Chou, M. W. Kattan, and D. Schrag. Predicting survival after curative colectomy for cancer: Individualizing colon cancer staging. *Journal of Clinical Oncology*, 29(36):4796–4802, 2011. doi: 10.1200/JCO.2011.36.5080.
- P. Whiting, Al Maiwenn, L. Burgers, M. Westwood, S. Ryder, M. Hoogendoorn, N. Armstrong, A. Allen, H. Severens, and J. Kleijnen. Ivacaftor for the treatment of patients with cystic fibrosis and the G551D mutation: A systematic review and cost-effectiveness analysis. *Health Technology Assessment*, 18(18):1–106, 2014. doi: 10.3310/hta18180.
- WHO Regional Office for Europe. Body mass index-BMI, 2021. URL <https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi>.
- D. H. Wolpert. Stacked Generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.
- A. Wong, A. T. Young, A. S. Liang, R. Gonzales, V. C. Douglas, and D. Hadley. Development and Validation of an Electronic Health Record–Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment. *JAMA Network Open*, 1(4):e181018, 2018. doi: 10.1001/jamanetworkopen.2018.1018.
- M. N. Wright and A. Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 2017. doi: 10.18637/jss.v077.i01.

- K. Yamagishi. When a 12.86% Mortality is More Dangerous than 24.14%: Implications for Risk Communication. *Applied Cognitive Psychology*, 11:495–506, 1997.
- Y. Zeng and P. Breheny. The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R. *ArXiv e-prints*, 1701.05936, 2017. URL <https://arxiv.org/abs/1701.05936>.
- W. Zheng and M. J. van der Laan. Causal Mediation in a Survival Setting with Time-Dependent Mediators. *U.C. Berkeley Division of Biostatistics Working Paper Series*, (295):1–49, 2012. URL <http://biostats.bepress.com/ucbbiostat/paper295>.
- Y. Zheng and P. J. Heagerty. Partly conditional survival models for longitudinal data. *Biometrics*, 61:379–391, 2005. doi: 10.1111/j.1541-0420.2005.00323.x.
- Y. Zheng and P. J. Heagerty. Prospective accuracy for longitudinal markers. *Biometrics*, 63:332–341, 2007. doi: 10.1111/j.1541-0420.2006.00726.x.