The Application of Computational Statistics, Data Science and Machine Learning in the Assessment of Comorbidity and Cancer Survival

Kieran Zucker

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

University of Leeds

Leeds Institute of Medical Research

Leeds Institute for Data Analytics

January 2021

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

The delivery of this thesis has been made possible through the time, effort and support of a wide range of people. I would like to give particular thanks to some of these individuals including my supervisory team of Geoff Hall, Adam Glaser and Paul Baxter, for the huge levels of time that they have provided throughout the duration of my period of PhD study. Their advice, expertise and feedback has been invaluable throughout this process, helping to constantly push my work to be of the highest possible standards.

Ciarán McInerney has shown himself to be the consummate colleague, providing advice and support on many of the practical and academic questions I have had during the production of this thesis and the work within it. His input has helped shape not only the direction of this thesis itself, but also my wider thinking within my academic work. It has been a pleasure to work alongside him on this and other projects.

I would like to thank Colin Johnston for his advice and input and the provision of pre-processed geographical data that was used as part of the analysis in the third chapter of this thesis.

I would like to thank all of the patients whose data was used in the delivery of this research. Without NHS patients such as these, research like this would not be possible.

The delivery of this research was made possible through the kind support of Macmillan Cancer Support funding and a generous donation to the University of Leeds from Geoff Oatley

Finally I would like to thank my family and friends for their support and encouragement throughout the production of this thesis and the work contained within it.

Abstract

An ageing population has resulted in an increasing number of people living with multiple long term health conditions, prompting interest in how comorbidity impacts on the survival outcomes of patients with cancer diagnoses. To date, no analyses have been undertaken using large population data that applies a consistent methodological approach to multiple cancers and comorbidities.

Using retrospective data from patients with a known cancer diagnosis from the Leeds Cancer Centre, a range of descriptive, exploratory, inferential and predictive approaches have been applied to assess for bias in the data, as well as describe, infer and predict survival outcomes in up to 24 cancers and 40 chronic health conditions.

Analysis results highlighted multiple potential sources of bias within both the hospital and cancer dataset, with differences in demographics, missing data and meaningful inaccuracies found within the clinical coding data for comorbidity.

Results from Kaplan Meier and Cox modelling identified that comorbidity was most commonly associated with worse survival in cancer patients, however effects were highly variable and several comorbidities were found to be associated with improved survival. Cause-specific Cox models suggested that in many cases the hazard differences seen were due to a direct association with cancer cause-specific hazard. The application of random survival forests was demonstrated to provide superior predictions when compared to traditional methods. Further, they demonstrated multiple non-linear relationships between predicted survival and predictors such as age, stage and grade.

The results identify multiple potential flaws within previous comorbidity research using observational data in both oncology and other medical fields. Despite some potential sources of bias, the results presented represent one of the most comprehensive analyses of comorbidity in cancer using a consistent methodological approach and further highlight the utility of machine learning methods in this form of analysis.

Contents

Acknowledgements 3
Abstract 4
List of Figures 10
List of Tables 13
Abbreviations 14
Chapter 1: Introduction and Project Overview 15
1.1 - Clinical Context of Cancer 15
1.2 - Comorbidity 15
1.3 - Routinely Collected Data (RCD) and Randomised Clinical Trials (RCTs) 17
1.4 - Big Data, Statistical Computing, Computational Statistics, Data Science and Machine Learning. 19
1.5 - Machine Learning in Healthcare 20
1.6 - Novelty of the Research 21
1.7 - Aims 21
1.8 - Thesis Structure 21
Chapter 2 – Methodology 22
2.1 - Introduction 22
2.1.1 - Understanding Types of Analysis. 22
2.2 - Raw Data and Dataset Creation 24
2.2.1 - Clinical Coding 24
2.2.2 - Comorbidity Scores 25
2.2.3 - Identification of the Cohort 26
2.2.4 - Details of Cancer Diagnosis 27
2.2.5 - Comorbidity Data 27
2.2.6 - Identification of the Leeds Teaching Hospitals Blood Catchment Area 28
2.2.7 - De-identification of Data 28
2.2.8 - Cohorts Developed 29
2.2.9 - Data Pre-processing 29
2.2.10 – Index of Multiple Deprivation (IMD) 32
2.3 Descriptive and Exploratory Analyses 33
2.3.1 - Data Provenance 33
2.3.2 - Missingness 34
2.3.4 - Correlation / Collinearity 36
2.3.5 - Bias Confounding and Statistical Significance. 36
2.3.5 - Accuracy 38

2.3.6 - Excess Zeros 38
2.4 - Survival Analysis 39
2.4.1 - Kaplan Meier Estimates (KM) 39
2.4.2 - Cox Proportional Hazards 40
2.4.3 - Competing Risks Analysis 41
2.4.4 - Forest Methods 43
2.4.5 - Predictive Accuracy 44
2.5 - Prognosis Research Strategy Group (PROGRESS) 45
2.6 - Software and Tools Used 46
2.7 - Ethics 46
Chapter 3 – Descriptive and Exploratory Analysis of the Research Dataset 47
3.0 - Introduction 47
3.0.1 - Context 47
3.0.2 - Data Provenance: The History and Origins of PPM and the Leeds Dataset 47
3.0.3 - Aims and Objectives 48
3.1 - Methods 48
3.1.1 - Identification of the PPM cohort and sub-cohorts: 48
3.1.2 - Demographics and Basic Descriptors 49
3.1.3 - Missingness 49
3.1.4 - Patient Geography 49
3.1.5 - Accuracy of Clinical Coding 50
3.1.6 - Prevalence of Comorbidity 52
3.1.6 - Age and Comorbidity: 52
3.1.7 - Timing of comorbidity analysis 52
3.1.8 - Collinearity of Variables 53
3.2 - Results 53
3.2.1 - Demographics 53
3.2.2 - Missingness 59
3.2.3 - Patient Geography 61
3.2.4 - Accuracy of Clinical Coding 62
3.2.5 - Prevalence of Comorbidity 66
3.2.6 - Timing 69
3.2.7 – Collinearity 74
2.2.8 - Interaction between Age and Comorbidity 76
3.3 - Discussion 79
3.3.1 - Demographics: 79

3.3.2 - Missingness 80
3.3.3 - Patient Geography 81
3.3.4 - Accuracy of Clinical Coding 82
3.3.5 - Prevalence of Comorbidity 84
3.3.6 - Age and Comorbidity 85
3.3.7 – Timing of Comorbidity Diagnosis 86
3.3.8 - Collinearity 87
3.4 - Summary 88
Chapter 4 – Survival Outcomes in Comorbid Cancer Patients Using a Univariate Kaplan Meier Approach 89
4.0 - Introduction 89
4.0.1 - Aims and Objectives 89
4.1 – Methods 90
4.1.1 - Survival Methods, Whole Population Survival and Summary Statistics 90
4.1.2 - Stratified Survival of Pre-defined Cancers and Comorbidities 90
4.1.3 - Identifying Comorbidities and Cancer Sites of Focus 91
4.1.4 - Assessing the Relationship between Comorbidity Impact and Whole Population Survival 91
4.2 – Results 91
4.2.1 - Total Population Survival and Summary Statistics 91
4.2.2 - Stratified Survival 93
4.2.3 - Identifying additional Comorbidities and Sites of Interest 101
4.2.4 - Relationship between stratified Survival and Overall Survival 104
4.3 - Discussion: 105
4.3.1 - Total Population Survival and Summary Statistics 105
4.3.2 - Stratified Survival 106
4.2.3 - Identifying Additional Comorbidities and Sites of Interest 109
4.2.4 - Relationship between Stratified Survival and Overall Survival 109
4.2.5 - Limitations: 110
4.4 - Summary 113
Chapter 5: Multivariable Approach to Analysing the Relationship between Comorbidity and Cancer Survival Outcomes 114
5.0 - Introduction 114
5.0.1 – Aims and Objectives 114
5.1 - Methods 115
5.1.1 - Building Survival Models 115

5.2.2 - Analysis of Model Assumptions 115
5.2.3 - Extraction of Summary Statistics 115
5.3 - Results 116
5.3.1 - Survival Models 116
5.3.2 - Cox Model Diagnostics 117
5.3.2 - Extraction of Summary Statistics 121
5.4 - Discussion 134
5.4.1 - Assessment of Model Assumptions 134
5.4.2 - Identifying Potential Bias Using Directed Acyclic Graphs (DAGs) 135
5.4.2 - Associations between Comorbidity and Survival 139
5.4.3 - Precision and Clinical Relevance: 146
5.4.4 - Variable Inclusion 147
5.5 - Summary 148
Chapter 6 – Inferential Analysis of the Association between Cause-Specific Cancer Risk and Comorbidity in Cancer Patients 149
6.0 - Introduction 149
6.0.1 - Aims and Objectives 149
6.1 - Methods 150
6.1.1 - Data Pre-processing 150
6.1.2 - Building of Models, Testing Assumptions and Quantification of Associations 150
6.1.3 - Comparison to All-Cause Results 150
6.2 - Results 151
6.2.1 - Data Pre-processing and Exploration 151
6.2.2 - Model Assumptions 153
6.2.3 – Cancer Cause-Specific Hazards 157
6.2.4 - Comparison with All-Cause Results 168
6.3 - Discussion 171
6.3.1 - Data Pre-processing 171
6.3.2 - Model Assumptions 171
6.3.3 - Cause of Death as a Source of Bias 172
6.3.4 - Interpretation of Cause-Specific Analysis 173
6.3.5 - Comparison to Previous Research 174
6.3.6 - Cause Specific Hazard Estimates 175
6.4 - Summary 178
Chapter 7 – The Application of Random Survival Forests to Assess the Relationship between Baseline Characteristics and Cancer Survival Using a Predictive Framework 179

7.0 - Introduction:..... 179
7.0.1 – Aims and Objectives..... 179
7.1 - Methods..... 180
7.1.1 - Testing and Training Sets..... 180
7.1.2 - Hyperparameter Tuning..... 180
7.1.3 - Model Building..... 180
7.1.4 - Internal Error Assessment..... 180
7.1.5 - Model Accuracy..... 180
7.1.6 - Assessing Variable Importance..... 181
7.1.7 - Partial Dependence Plots..... 181
7.2 - Results..... 181
7.2.1 - Hyperparameter Tuning and OOB Error..... 181
7.2.2 - Predictive Accuracy..... 183
7.2.3 - VIMP..... 185
7.2.4 - Pairwise Permutation VIMP..... 190
7.2.5 - Partial Plots..... 194
7.2.6 - Survival Curves..... 203
7.3 - Discussion..... 210
7.3.1 - Optimisation:..... 210
7.3.2 - Predictive Accuracy..... 210
7.3.3 - Single Feature VIMP and Tree Depth..... 211
7.3.4 - Pairwise VIMP..... 212
7.3.5 - Partial Plots..... 214
7.3.6 - Associations between Comorbidity and Predicted Survival..... 217
7.4 - Summary..... 218
Chapter 8 – Conclusions and Further Work..... 219
8.0 - Introduction..... 219
8.1 - Accuracy, Reliability and Representativeness of Hospital Data..... 219
8.2 - The Impact of Comorbidity on Cancer Patient Survival..... 220
8.3 – Information for Predicting Cancer Outcomes..... 222
8.4 - Traditional Statistical Methods versus Machine learning Methods..... 223
8.5 – Study Limitations..... 223
8.6 - Further Work..... 224
8.7 - Summary..... 228
Appendix..... 229
Bibliography..... 244

List of Figures

Figure 1: Data Pre-processing Flow Diagram ... 31
Figure 2: Graphical Representation of Diabetes Mellitus Data Definitions and Subgroups ... 51
Figure 3: Consort Diagram for PPM Population ... 54
Figure 4: National, Regional and Local Population Pyramid Comparison ... 55
Figure 5: Case Registration Numbers in PPM ... 56
Figure 6: Age and Gender Distribution of the PPM All Cancer Cohort ... 58
Figure 7: Annual Missing Cancer Stage and Grade data in the All Cancer Cohort of the PPM Dataset ... 60
Figure 8: Boundary Effects ... 61
Figure 9: Fidelity of DM Identification ... 63
Figure 10: Difference in First Diabetic Diagnosis Indication from Clinical Records Comparing Clinical Coding and Abnormal HbA1c ... 64
Figure 11: Survival Trajectories for Each Diabetic Data Definition and All Patients in the All Cancer Cohort ... 65
Figure 12: Impact of Data Definitions on the Cox Derived Hazard Ratios for the Impact of Diabetes ... 66
Figure 13: Prevalence of Comorbidity in the All Cancer Cohort ... 68
Figure 14: Distribution of Comorbidity Diagnosis Event Relative to Cancer Diagnosis ... 70
Figure 15: Distribution of Comorbidity Diagnosis Events Relative to Cancer Diagnosis Daily in the Year of Cancer Diagnosis ... 71
Figure 16: Distribution of Comorbidity Diagnosis Events Relative to Cancer Diagnosis by Comorbidity ... 72
Figure 17: Site Specific Distribution of Comorbidity Diagnosis Events Relative to Cancer Diagnosis Daily in the Year of Cancer Diagnosis ... 73
Figure 18: Statistical Significant of Pairwise Spearman's Correlation ... 74
Figure 19: Spearman's Correlation Coefficients for Pairwise Comparisons ... 75
Figure 20: Number of Comorbidities per Patient in the All Cancer Cohort ... 77
Figure 21: Relationship Between Median Age at Diagnosis and and the Percentage of Patients with One or More Comorbidities ... 78
Figure 22: Overall Median Survival for Site Specific Cohorts ... 91
Figure 23: Overall Survival for Testicular Site Specific Cohort ... 92
Figure 24: Breast Cancer Survival Stratified by History of Arrhythmia ... 93
Figure 25: Impact of Comorbidity in Breast Cancer ... 94
Figure 26: Impact of Comorbidity in Colorectal Cancer ... 95
Figure 27 : Impact of Comorbidity in Lung Cancer ... 96
Figure 28: Impact of Comorbidity in Prostate Cancer ... 97
Figure 29: Association of Diabetes Mellitus to Median Survival in Site Specific Cohorts ... 98
Figure 30: Association of Stroke to Median Survival in Site Specific Cohorts ... 99
Figure 31: Association of Myocardial Infarction (MI) to Median Survival in Site Specific Cohorts 100
Figure 32: Association of Congestive Cardiac Failure (CCF) to Median Survival in Site Specific Cohorts ... 102
Figure 33: Association of Chronic Obstructive Pulmonary Disease (COPD) to Median Survival in Site Specific Cohorts ... 103
Figure 34: Correlation between Median Survival and Percentage Change in Median Survival in Comorbid Group ... 104

Figure 35: Correlation between Median Survival and Percentage Change in Median Survival in Diabetic Group..... 105
Figure 36: Association between Diabetes and Survival in Breast Cancer 116
Figure 37: Shoenfeld Residual Plot for Covariates in the Stroke in Breast Cancer Cox Model..... 117
Figure 38: Assessment of Linear Assumptions of Age in Breast Cancer Site Specific Cohort. 119
Figure 39: Effect of Influential Outliers. 120
Figure 40: Cox Derived Hazard Ratios for Comorbidity in Breast Cancer..... 122
Figure 41: Cox Derived Hazard Ratios for Comorbidity in Lung Cancer 123
Figure 42 : Cox Derived Hazard Ratios for Comorbidity in Prostate Cancer..... 124
Figure 43: Cox Derived Hazard Ratios for Comorbidity in Colorectal Cancer..... 125
Figure 44: Cox Derived Hazard Ratios for Diabetes Mellitus in All Cancer and Site Specific Cancer Cohorts 127
Figure 45: Cox Derived Hazard Ratios for MI in All Cancer and Site Specific Cancer Cohorts 128
Figure 46: Cox Derived Hazard Ratios for Stroke in All Cancer and Site Specific Cancer Cohorts .. 129
Figure 47: Cox Derived Hazard Ratios for CCF in All Cancer and Site Specific Cancer Cohorts..... 130
Figure 48: Cox Derived Hazard Ratios for COPD in All Cancer and Site Specific Cancer Cohorts.... 131
Figure 49: Diabetes and Death Directed Acyclic Graph (DAG)..... 136
Figure 50: Diabetes and Death Mediated via Cancer Directed Acyclic Graph (DAG) 136
Figure 51: : Total Causal Effect of Diabetes and Death Mediated via Cancer Directed Acyclic Graph (DAG). 136
Figure 52: Detailed Diabetes and Death Mediated via Cancer Directed Acyclic Graph (DAG) 137
Figure 53: Detailed Total Causal Effect of Diabetes and Death Mediated via Cancer Directed Acyclic Graph (DAG) 137
Figure 54: DAG of Open and Closed Causal Pathways due to the Study Design 138
Figure 55: DAG of Open and Closed Causal Pathways Due to Study Design After Adjustment. 139
Figure 56: Missing Cause of Death Data in All Cancer Cohort..... 152
Figure 57: Shoenfeld Residual Plot for Covariates in the CCF in Thyroid Cancer Cause Specific Cox Model 153
Figure 58: Assessment of Linear Assumptions of Deprivation Quintile in Thyroid Cancer Site Specific Cohort for Cancer Cause-Specific Analysis 155
Figure 59: Effect of Influential Outliers on Breast and Prostate Cancer-Cause Specific Hazard. 156
Figure 60: Effect of Influential Outliers Colorectal Cancer-Cause Specific Hazard 157
Figure 61: **Cox Derived Cancer Cause-Specific Hazard Ratios for Comorbidity in Breast Cancer** ... 158
Figure 62: Cox Derived Cancer Cause-Specific Hazard Ratios for Comorbidity in Colorectal Cancer 159
Figure 63: Cox Derived Cancer Cause-Specific Hazard Ratios for Comorbidity Lung Cancer 160
Figure 64: Cox Derived Cancer Cause-Specific Hazard Ratios for Comorbidity Prostate Cancer. ... 161
Figure 65: Cox Derived Cancer Cause-Specific Hazard Ratios for CCF in All Cancer and Site Specific Cancer Cohorts 163
Figure 66: Cox Derived Cancer Cause-Specific Hazard Ratios for COPD in All Cancer and Site Specific Cancer Cohorts 164
Figure 67: Cox Derived Cancer Cause-Specific Hazard Ratios for Diabetes Mellitus in All Cancer and Site Specific Cancer Cohorts..... 165
Figure 68: Cox Derived Cancer Cause-Specific Hazard Ratios for MI in All Cancer and Site Specific Cancer Cohorts 166
Figure 69: Cox Derived Cancer Cause-Specific Hazard Ratios for Stroke in All Cancer and Site Specific Cancer Cohorts..... 167
Figure 70: Effect of forest size on Out of Bag (OOB) Error..... 182

Figure 71: Time Dependent Accuracy with Brier Scores 184
Figure 72: Breast Cancer Importance Scores 186
Figure 73: Colorectal Cancer Importance Scores 187
Figure 74: Lung Cancer Importance Scores 188
Figure 75: Prostate Cancer Importance Scores 189
Figure 76: Relationship Between Age and Predicted Survival in Breast Cancer at 1, 5 and 10 Years 195
Figure 77: Relationship Between Age and Predicted Survival in Colorectal Cancer at 1, 5 and 10 Years 196
Figure 78: Relationship Between Age and Predicted Survival in Lung Cancer at 1, 5 and 10 Years 197
Figure 79: Relationship between Age and Predicted Survival in Prostate Cancer at 1, 5 and 10 Years 198
Figure 80: Partial Plots for Cancer Stage at Presentation 200
Figure 81: Partial Plots for Cancer Grade at Presentation 202
Figure 82: RSF Derived Stratified Survival Curves for Patients with Prior CCF 204
Figure 83: RSF Derived Stratified Survival Curves for Patients with Prior MI 205
Figure 84: RSF Derived Stratified Survival Curves for Patients with Prior COPD 206
Figure 85: RSF Derived Stratified Survival Curves for Patients with Prior Stroke 207
Figure 86: RSF Derived Stratified Survival Curves for Patients with Prior DM 208
Figure 87: RSF Derived Stratified Survival Curve Varicosities in Colorectal Cancer 209
Figure 88: RSF Derived Stratified Survival Curve Obesity in Lung Cancer 209
Figure 89: Correspondence with Dr Hemant Ishwaran 238

List of Tables

Table 1: Summary of Data Analysis Types ... 23
Table 2: Summary of ACE-27 ... 26
Table 3: Types of Missingness ... 35
Table 4: All Cancer Cohort and Site Specific Cohorts Demographics Summary ... 57
Table 5: Percentage of Complete Data for Site Specific Cohorts ... 59
Table 6: Comparison of Comorbidity Prevalence in Local, Regional and National Data ... 66
Table 7: Comparison of Published Survival Estimates to Calculated Survival Estimates ... 92
Table 8: Ten Most Impactful Comorbidities ... 101
Table 9: Summary of Linear Assumptions Assessment ... 118
Table 10: High Precision Comorbidity Hazard Ratio Results ... 133
Table 11: Summary of Completeness of Cause of Death Data by Cohort ... 151
Table 12: Summary of Cause-Specific Linear Assumptions Assessment ... 154
Table 13: High Precision Comorbidity Cancer Cause-Specific Hazard Ratio Results ... 168
Table 14: Comorbidity Cancer Cause-Specific Hazard Ratio Results Consistent with All Cause Hazard Results ... 170
Table 15: High Precision Comorbidity Cancer Cause-Specific Hazard Ratio Results with Results Consistent with All-Cause Hazard Ratio ... 170
Table 16: Optimised Hyperparameters for Random Survival Forests ... 181
Table 17: Comparison of KM, Cox and RSF Integrated Brier Scores ... 183
Table 18: Top 5 Pairwise VIMP Results in Each Cancer Site Specific Cohort ... 191
Table 19: Top 5 Feature Combinations in Each Cancer Site Specific Cohort Where Pairwise VIMP Exceeds Additive VIMP ... 192
Table 20: Top 5 Feature Combinations in Each Cancer Site Specific Cohort where additive VIMP exceeds pairwise VIMP ... 193
Table 21: Histological Groupings Applied in Each Cancer Site ... 230
Table 22: Comorbidity Code Definitions ... 231
Table 23: Complete List of R Packages and Versions Used Grouped by Package Utility ... 235
Table 24: Cancer Site Specific Cohort ICD-10 Code Definitions ... 236
Table 25: Model Specification for Cox Models ... 237
Table 26: Random Survival Model Specification ... 237
Table 27: Sensitivity Analysis of All-Cause Mortality Cox Models ... 240
Table 28: Sensitivity Analysis of Cancer Cause-Specific Mortality Cox Models ... 242

Abbreviations

CAD	Coronary Artery Disease
CCF	Congestive Cardiac Failure
CI	Confidence Interval
CKD	Chronic Kidney Disease
COPD	Chronic Obstructive Pulmonary Disease
Cox PH	Cox Proportional Hazards
CUP	Cancer of Unknown Primary
CVA	Cerebrovascular Accident
CVD	Chronic Venous Disease
DM	Diabetes Mellitus
EPR	Electronic Patient Record
FIGO	International Federation of Gynaecology and Obstetrics
HBA1c	Haemoglobin A1c
HPTN	Hypertension
IBD	Inflammatory Bowel Disease
ICD-10	International Classification of Diseases Version 10
LIME	Locally Interpretable Model Agnostic Explanations
MAR	Missing at Random
MCAR	Missing Completely at Random
MND	Motor Neurone Disease
n	Number
NCRAS	National Cancer Registration and Analytics Service
ONS	Office of National Statistics
OOB	Out of Bag
PPM	Patient Pathway Manager
PVD	Peripheral Vascular Disease
QOF	Quality Outcomes Framework
RCD	Routinely Collected Data
RCT	Randomised Controlled Trial
RSF	Random Survival Forest
TIA	Transient Ischaemic Attack
TNM	Tumour, Node, Metastasis
UK	United Kingdom
VEGF	Vascular Endothelial Growth Factor
VIMP	Variable Importance

Chapter 1: Introduction and Project Overview

1.1 - Clinical Context of Cancer

Cancer is the second most common cause of death globally and accounts for 28% of deaths[1] in the United Kingdom (UK). UK government estimates project that one in two people will develop cancer in their lifetime[2] with 3 million people living with or beyond cancer by 2030.[3] The high rate of incidence within the population has implications not just for patients, but also for health providers. The cost of the average cancer therapeutic has quadrupled over the past decade, with the cost of cancer care accounting for 4.3% of the total UK health budget in 2013[4] and 3% of hospital care spending in 2016.[5]

Over the past four decades the management of cancer has evolved significantly, with the proportion of patients surviving to ten years or more after their cancer diagnosis doubling to over 50%.[1] These improvements have been driven by advances not only in therapeutic approaches, but also by altered health behaviours, improved diagnostics, screening programmes and greater public awareness.[6,7]

Despite these improvements, large variations in the outcomes for cancer are seen across Europe and globally. Previous research has suggested that this variation may be due to a wide range of factors including different clinical practice, timing of presentation, access to secondary care, cancer waiting times and access to treatments.[7–10] This variation in geographical outcomes has also resulted in a number of scientific arguments over whether the differences seen are truly present, or due to how data is collected.[11–13] Additional debate occurs as to whether these differences in outcomes are in fact due to appropriate clinical decision making by taking a more holistic approach to care delivery. The best outcome for a patient may not be determined solely by how long they live, but also taking quality of life into account. This moves the debate away from just an issue of quantum.[14–16]

Cancer represents a condition of particular importance to governments, health care providers and the general population due to its impact on individuals, health systems, public policy and public finances.[17] As a result, understanding what drives the differences in outcomes seen both within populations and between populations, is an area of heavy focus for researchers and policy makers alike. Much of this research is focussed on health interventions in the form of treatments and screening. Beyond this, researchers have also attempted to identify groups who are at a higher risk of both developing cancer and having worse outcomes.[18–23] This research has often focussed on genetics, cancer type, extent of disease at presentation and the patient's general health at the time of diagnosis. Despite this interest, a number of methodological constraints make research of this nature challenging and the results of such studies often conflict with one another.[24] The presence and absence of comorbidity has been a particular focus for cancer outcomes research and will form the main focus of this thesis. This chapter will summarise the types of study undertaken in the area of comorbidity in cancer, cover some of the key research findings of existing research and highlight gaps in the current knowledge base that require further investigation. The key research aims will be provided, aimed at covering some of these current gaps in scientific knowledge.

1.2 - Comorbidity

Within this introduction and throughout the thesis comorbidity is referred to several times. This terminology can be found with relative abundance within the scientific literature, however despite this, the terms are applied inconsistently within different domains, such that its use in epidemiology[25], clinical practice[26], health policy[27] and management may differ. This issue is further

compounded by the use of related terms such as multi-morbidity, frailty, disease burden and burden of ill health.[28–30] These terms are sometimes used interchangeably with each other and comorbidity, but in some literature are used to describe different concepts altogether. This lack of a consistency in the use of terminology can produce further uncertainty when attempting to assess the current state of research relating to these concepts. Previous research looking at this issue[31] identified that although each definition was based on an individualised perspective, the researchers describe four distinct conceptualisations of comorbidity: 1) the nature of the health condition, 2) the relative importance of the co-occurring health conditions, 3) the chronology of presentation of the conditions, 4) expanded conceptualisations.

The nature of the health condition concept is focussed on being able to attribute a specific and discrete categorisation to it. If the individual illness, diagnosis or condition is poorly defined or nebulous, then there is a risk of overlap between one or more clinical concepts or diagnoses. A key example is the co-occurrence of anxiety and depression where without a strict definition of the nature of the condition, it is impossible to state whether this is one condition on a spectrum, or two co-occurring conditions and thus comorbidities.[32]

The importance concept is based on the use of an index condition. Any other condition that occurs within the window of time that the individual is affected by the index diagnosis is considered a comorbidity. This creates uncertainty, as depending on the clinical or research question, the index condition may differ. A patient with congestive heart failure and lung cancer may be regarded as having cancer as a comorbidity by their cardiologist and heart failure as a comorbidity by their oncologist.

The chronology of conditions also features heavily within the literature and may be applied inconsistently. This perspective encapsulates overlap and sequence, some conditions may occur with a time lag but still be interdependent, others might occur synchronously but with differences in the initial point of diagnosis. This is further complicated by the distinction between comorbidity and late effects, where a related condition that occurs after the first, but is in some way induced by the first, may be referred to as a late effect. As with comorbidity, no consensus definition for what constitutes a late effect currently exists within the literature.[33–37]

The expanded concepts of comorbidity include those with scoring systems which attempt to combine multiple aspects of conditions into a single mechanism of assessment. Examples of this include the Charlson index[38] and Elixhauser score.[39] Frailty may also be included within this conceptualisation, where the measure attempts to capture the biological effects or phenotypic expression of ill health, how this manifests, and may impair function. An example of this includes the electronic Frailty Index[40] and Hospital Frailty Index.[41]

With more people globally living longer, there are a growing number of people living with multiple health conditions.[42] It is therefore unsurprising that across a range of medical domains there is a corresponding growth in research into how patients with multiple health conditions differ in their outcome, to those with no other underlying health conditions. The lack of consistency in how the concept of comorbidity is applied adds to the existing uncertainty of results, as many studies are not comparable.

As discussed above, the increasing incidence of cancer and improved outcomes combine to create a growing population living with and beyond cancer. This has resulted in the generation of a significant body of literature focussing on the impact of comorbidities on cancer outcomes.[24,43] In addition to the issues described above, much of the previous research is limited by other factors. These include

a limited number of comorbid conditions assessed, analyses restricted to only the most common cancers, small population studies, large population studies but with limited breadth of information and a failure to adequately consider methodological limitations of the approaches applied. As a result, although a number of studies in this area have been undertaken before, little is known about many of the important aspects of the interplay between baseline health status and cancer outcomes. Key areas where knowledge is lacking include how accurate records are of health conditions in hospital datasets, whether hospitals offer a representative dataset on which to base conclusions, the impact of health conditions on less common cancers, the impact of less common health conditions on cancer outcomes, whether individual pre-existing health conditions impact on cancer outcome and which, if any, health problems are useful in predicting outcomes for cancer patients.

1.3 - Routinely Collected Data (RCD) and Randomised Clinical Trials (RCTs)

As clinical practice has increasingly focussed on improving the evidence base for care delivery and medical interventions, a consensus hierarchy of evidence quality has been established within the scientific community. The pinnacle of this is the meta-analysis or systematic review which combines multiple studies. The next tier in the evidence hierarchy is the randomised controlled trial (RCT). [44] The focus of RCTs as the most valuable form of singular study is based on a number of key advantages of randomisation. If appropriately sized, confounding is minimised by ensuring that confounding factors are equally distributed between the groups under study.[45] This can then be combined with further measures such as blinding to reduce bias from both clinician and patient expectation. Despite these key advantages there are a number of significant limitations.

High quality randomised controlled trials require significant administration and oversight. This results in significant costs in their delivery.[46] The costs of RCTs are growing year on year such that large scale RCTs are now commonly the purview of only the largest grant providers and commercial sponsors.[46] These high costs can also result in a number of other study design decisions that can limit the utility of the study results. Firstly, follow up times are commonly short to minimise costs. This often means that any advantages seen when comparing groups are only those that can be demonstrated over a period of a few months or years. The longer the study, the larger the costs involved. In some cases this may mean that long term differences, both beneficial and detrimental, may never be identified.[47] Secondly, in order to obtain clear cut findings limitations may be placed on those recruited into the trial. This includes limits on extremes of age and comorbid health conditions. This, when combined with difficulties in recruiting certain groups such as ethnic minorities and individuals from lower socioeconomic groups, may result in a study population that bares very little resemblance to the true population in which one wishes to utilise the intervention under investigation.[48,49] Within the oncology setting this is particularly true with only 1% of oncology patients participating in clinical studies[50] and RCTs more generally having been shown to exclude on average 77.1% of patients.[51–53] This results in a scenario where information on often the fittest patients, with lowest burdens of oncological and non-oncological diseases, are used to determine the clinical interventions for the majority of less well patients with more advanced disease.[47] As a result, the data may generalise well within the groups included in the study, however as the population is unrepresentative, differences in the impact of a given intervention or effect in other subgroups of the population are not assessed and may well diverge from the original study population.[54]

Additionally, RCTs become problematic when dealing with rare conditions or rare outcomes. This is due to either having issues recruiting sufficient numbers, or having to have studies with a large enough cohort to capture sufficient numbers of the rare outcomes.[55] This in many cases renders an

RCT unfeasible. In other cases, RCTs are an inappropriate study design choice due to the characteristic of interest not being an intervention or an assignable characteristic. The issue of comorbidity and its impact on long term health outcomes is one such example. It is not possible to randomly assign someone a long term health condition and thus an RCT would not be conducted. In many studies this issue is overcome through the use of subgroup analysis. In such cases the results from the RCT are split into the groups of interest and compared. This division into subgroups is a form of stratification but that is occurring outside of, and after the randomisation process and thus the benefits of controlling for confounding are bypassed.[56] This approach is, in effect, an observational study within an RCT cohort and dataset. Thus although much of the data collected is highly accurate, the population is fundamentally different from that of the wider general population. This therefore becomes an unrepresentative observational study but with excellent data quality.

An alternative approach is to conduct whole population observational studies. This has the advantage of including all patients within a given geography, making the cohort inherently more representative as it includes everyone. It does not however control for any potential confounding caused by systematic differences between the groups of interest, or ensure that the population studied is the same as wider populations, therefore also placing limits on the external utility of this approach to wider groups.

Although observational studies can be conducted in a prospective manner, it is also possible to conduct them in a retrospective manner. With the rapid growth in the use of electronic healthcare records, increasing volumes of clinical data are being captured and recorded year on year[57]. This information has become an increasing focus of study, with retrospective routinely collected data forming the basis of many observational studies. This approach results in a number of theoretical advantages including reduced costs, a more representative population, longer follow up data, and the potential to study rare diseases or outcomes.[58]

Despite this, many within the research community have cautioned against the assumption that these theoretical advantages are borne out in reality.[59–63] The costs of individual studies may be lower, but when considering the substantial costs involved in the adoption and maintenance of systems to collect and capture routinely collected data (RCD), it may be viewed that the costs have simply moved from the research community to healthcare providers instead. Although the population may be more representative than in a highly selected RCT, in many instances RCD is still not representative of whole populations. This occurs as the source from which data is obtained may not include all individuals of interest. This is seen when analysing hospital data, which will show a bias towards patients with the greatest levels of ill health or more severe presentations of conditions. If a reliance is placed on the use of structured data then accuracy and completeness of this[64], which may vary between hospitals, may render the data less representative.

It is common for different health providers to collect their own data. Where patients move between different providers, then the RCD can in turn become fragmented, such that an individual data source may not hold all the important and relevant data for an individual.[65,66] Although these issues could be overcome by combining data from multiple sources, this linkage process presents a number of legal and ethical issues which can be challenging, time consuming and costly to overcome.[67–70]

Although the large volumes of data available present a significant opportunity for novel insight to be generated, it also poses a risk. The large cohort sizes result in extremely small p values and therefore dramatically increase the risks of false discovery, with highly significant false-positive and false negative discoveries being identified.[61] As alluded to above, studies without randomisation are at risk of being impacted by bias, which can further exacerbate the issues of false discovery. As

differences in confounding and bias can exist between groups with identified relationships in observational data, accounting for what associations are due to the exposure of interest, and what is due to other factors can be extremely challenging.

These limitations are often used to suggest that the solution to RCT related issues is in fact to improve RCTs to make them larger, cheaper and more representative, rather than relying on RCD to fill gaps in current RCT knowledge.[62,63] In some instances a hybrid approach to trials is being adopted to include both traditional RCT data collection and the integration of RCD. This may be in using RCD to assist with participant identification and improving recruitment, the use of historic RCD for patients recruited to a trial to increase the volumes of data collected within the trial or using RCD to assess long term outcomes after formal in trial follow up has ended.[63,71,72] In many circumstances this may well be a solution, however in the context of an exposure which cannot be assigned, this approach cannot be taken. Instead researchers must make use of RCD whilst understanding and accepting its benefits and limitations, in order to attempt to answer important clinical questions that RCTs are unable to provide the answers to.

Within the context of cancer research RCTs have over the past several decades formed the backbone of research evidence.[73] This has been fuelled, in part, by a heavy focus on comparisons of interventions, which lend themselves particularly well to this form of study design. As greater focus been placed on how representative RCTs are in the cancer population and the rapidly growing costs of oncological drugs and drug trials[74], the oncology research community has increasingly utilised RCD as an adjunct to traditional RCTs.[75,76] The RCD approach is often made simpler within the context of cancer research as many countries collect cancer data nationally resulting in large well curated RCD that can be used for analysis.[77] RCD based research is therefore becoming a growing part of the cancer research information space, as either a standalone resource or in combination with RCT evidence.

1.4 - Big Data, Statistical Computing, Computational Statistics, Data Science and Machine Learning.

As the adoption of electronic healthcare records has increased over the past decade, the volume of health data available has increased exponentially over time.[78] When this is combined with other potential sources of data relating to health, such as wearable technology and consumer data, the volumes of information have become so large that manual calculation of many statistical methods would be at best impractical, if not impossible. The term “big data” has often been used to describe these large data resources, but despite the frequency with which this term is used within both the scientific and more general media, there is no consistency in the definition of the term.[79–82] A number of formal and informal definitions exist which makes determining what constitutes big data challenging and creates potential confusion when comparing studies.[83]

Even when focussing on the methodological domains, there is still inconsistency with the terminology used. Statistical computing may be broadly thought of as the use of computers to undertake statistical analysis.[84] Computational statistics is a term that was previously used interchangeably with statistical computing, however is now used to provide a slightly different meaning. Its use now tends to focus on computer methods that are reliant on approaches that are themselves computationally difficult not just by virtue of the size of the data. This includes the simulation of a distribution, methods that require multiple analyses, resampling, data partitioning and random number generation.[84] It is however also used to describe work that attempts to discover new information about the structure of a dataset, something that is also termed data mining.[85]

Further confusion is introduced through the use of the term data science. IBM for example defines data science as combining "the scientific method, math and statistics, specialized programming, advanced analytics, AI, and even storytelling to uncover and explain the … insights buried in data".[86] When attempting to draw distinctions between these different terms, the scientific literature is of little help. Many blogs and websites attempt to differentiate them based on the nature of the problem to be solved, scale of data, prediction versus explanation, however even here there is no consistency.[87,88] The research in these areas are made more difficult to compare as the terminology used to describe the same information is different within statistical computing and data science. "Variables" are termed "features", "observations" are termed "instances" and "dummy variables" are termed "one hot encoding".

The lack of consistency in both describing and defining the data and domain is likely to introduce ambiguity and thus focussing on the type of analysis and type of method is a more consistent approach. Here the purpose of the analysis can be outlined, such as descriptive, predictive, causal or inferential analysis.[89] The method itself can also be classified such as traditional statistical methods and machine learning. Here traditional statistical methods describes methods that are well established and initially implemented for the purpose of trying to infer information through the fitting of a project specific probability distribution.[90] Machine learning methods are those that use adaptive algorithms to identify patterns in complex data where the methods were commonly developed with a predictive focus.[91]

1.5 - Machine Learning in Healthcare

Over the past decade machine learning technology has been applied in a broad range of healthcare settings.[91,92] Two areas where the technology has been applied extensively are clinical imaging[93] and genetics.[94] Many of these projects have garnered significant media coverage such as tools to detect melanoma from photographs[95] and the automated analysis of imaging for retinal disease.[96] Despite the wide coverage of these in the media, in most cases developments in medical machine learning have yet to be translated into real world use of these technologies.[97]

Although for some this might be a point of frustration, there are also examples of where the technology has performed differently to the way it was intended. A high profile example of this was the Chex-Net project[98] which aimed to create an algorithm that could detect pneumonia on a chest radiograph. The results were published in a high impact journal, however later scrutiny showed that the entire project was based on a flawed dataset and thus the algorithm's accuracy in real world use would be much worse than stated, to the extent that if deployed, it would have potentially caused harm.[99,100]

Within the area of oncology, machine learning applications have mainly been focussed on screening, diagnosis, prognostication, staging, treatment selection and drug discovery.[101–103] Within the area of prognostication the majority of studies have been applied within very narrow confines. These include examples of the assessment of single test modalities, using single types of data such as genetics, or survival prediction at one fixed time point in a single cancer.[101,103] As such, there has been limited use of machine learning methods in the prediction of long term cancer outcomes across multiple cancers using a large dataset. Previous research has suggested potential accuracy benefits of these technologies over traditional methods, however, despite this, they have yet to be explored on a large scale.[104–106]

1.6 - Novelty of the Research

The work presented in this thesis strives to overcome some of these issues around previous research whilst addressing current gaps in understanding about baseline characteristics and in particular, prior health conditions and cancer outcomes. It will apply a consistent definition of comorbidity across multiple conditions, both common and less common, in a range of cancers that are common and less common. A consistent approach will be applied to the description of analyses undertaken, focussing on the type of analysis and type of method, such that the analyses will be consistent across cancers and comorbidities to allow for easy direct comparisons, which have been challenging in the past based on published data. The analyses are additionally focussed on the current gaps in knowledge highlighted in the above sections.

The work will assess the utility of a large English regional dataset of oncology RCD for reliable observational studies. The data will be used to describe, infer and predict the survival outcomes of cancer patients in the presence and absence of common long term health conditions and compare the accuracy of traditional statistical methods to machine learning approaches in cancer survival prediction. The seven aims for the research are set out below.

1.7 - Aims

1. To describe the characteristics of the research dataset and examine how these may introduce, bias, confounding and error into subsequent analysis and interpretation.
2. Quantify differences in survival outcomes within the Leeds Cancer Centre dataset compared to national survival data.
3. Describe the survival outcomes of cancer patients with and without comorbidity.
4. Identify which cancers and comorbidities have the most consistent survival effects.
5. Quantify the association between comorbidity and all-cause mortality in cancer patients using a multivariable approach.
6. Quantify the association between comorbidity and cancer cause-specific mortality in cancer patients using a multivariable approach.
7. Apply a predictive framework using machine learning to assess the utility of baseline characteristics including comorbidity in the prediction of survival in cancer patients.

1.8 - Thesis Structure

Following this chapter the thesis will continue with details focussing on the methods applied and their related theory. This will provide context to the approaches used, but will leave the fine details of the exact analyses undertaken to be described in each of the chapters of analysis. Subsequent chapters will highlight one or more of the seven aims on which it will focus and describe a number of specific objectives through which the aim will be delivered. Each of these chapters will then present background information, a detailed account of how the analysis was undertaken, results of the analyses followed by a discussion section explaining and expanding on results and placing them in the wider context of known and potential future research. The analyses are intentionally ordered such that the complexity of them increases as the reader progresses through the thesis. Chapter 3 will apply predominantly descriptive and exploratory analyses to address aims 1-2. Chapter 4 will apply the Kaplan Meier[107] method to meet aims 3-4. Chapters 5 and 6 will apply the Cox proportional hazards[108] model to address aims 5-6. Chapter 7 will apply random survival forests[109] to the data to address aim 7. The thesis will conclude with a final chapter which will summarise the overall findings of the work and address the primary research aims highlighted above.

Chapter 2 – Methodology

2.1 - Introduction

This chapter outlines some of the key methodological concepts that were considered and applied in the delivery of the subsequent research. Further, a detailed summary of the dataset generation process and tools used in the delivery of the research are provided. The fine detail of methods applied to deliver the analyses are covered in the methods section included within chapters 3-7.

2.1.1 - Understanding Types of Analysis.

A number of controversial publications have called into question the validity of the majority of scientific research.[110–113] A key issue that has been highlighted is the application of the correct interpretation of findings, relative to the type of analysis being undertaken.[114] In many instances conclusions are drawn that are beyond the scope of the type of analysis conducted. Data analysis can broadly be broken down into six key types which are descriptive, exploratory, inferential, predictive, causal and mechanistic. **Table 1** provides further detail on each. These distinctions are important as the depth and scale of conclusions that can be drawn differ substantially between them. When the application of a causal conclusion to an inferential study occurs, particularly alongside other inappropriate statistical practices commonly reported such as data tampering[115], interpreting findings based on expectation, ignoring missing data and violation of assumptions,[116,117] there is the potential to provoke dangerous changes to medical practice which are not evidenced. An example is that the rate of Parkinson's has been shown to be lower in smokers than non-smokers.[118] It would be inappropriate to suggest causation and thus patients with a high risk for Parkinson's should be encouraged to smoke.

The analyses undertaken within this thesis include; descriptive, exploratory, inferential and predictive but intentionally avoid attempts at causal inference due to the potential pitfalls of this approach in observational data.[119–121] A more detailed discussion on this can be found in chapter 5.

The following sections of this methodology chapter attempt to address key aspects of the analysis undertaken following the order of the analysis types above. Initially the development of the research dataset and key concepts underpinning this are described, along with the subsequent descriptive, exploratory, inferential and predictive analyses undertaken.

Type	Characteristics	Examples
Raw Data / No Data Analysis	• No summarisation approaches applied	Data table of patient observations from EHR
Descriptive	• Data summary • No interpretation applied	Demographics table
Exploratory	• Data summary applied • Interpretation applied • No attempt to quantify if discoveries will likely hold in a new sample or population	Is there a relationship or evidence of correlation between age and stage at cancer diagnosis?
Inferential	• Data summary applied • Interpretation applied • Attempt to quantify if discoveries will likely hold in a new sample or population • Not trying to predict measures for individuals • Not trying to assess how changing the average of one measurement will affect another	What is the estimated association between prior stroke and breast cancer survival after accounting for age, gender and deprivation?
Predictive	• Data summary applied • Interpretation applied • Trying to predict measures for individuals or populations • Attempts to quantify if accuracy will likely hold in a new sample or population • Not trying to assess how changing the average of one measurement will affect another	Can demographics be used to predict an individual patient's survival probability at one year after breast cancer diagnosis?
Causal	• Data summary applied • Interpretation applied • Attempts to quantify if discoveries will likely hold in a new sample or population • Attempts to assess how changing the average of one measurement will affect another on average • Trying to predict measures for individuals or populations	On average does altering the dose of a pain killer reduce pain symptoms?
Mechanistic	• Data summary applied • Interpretation applied • Attempts to quantify if discoveries will likely hold in a new sample or population • Trying to assess how changing the average of one measurement will affect another in a deterministic way.	Does reducing the gauge of material used in arterial stents reduce iatrogenic vascular resistance?

Table 1: Summary of Data Analysis Types *- Derived from publication by Leek et al[114], the table provides a list of the seven analysis types, a summary of their characteristics and an illustrative example of each.*

2.2 - Raw Data and Dataset Creation

Within this section the concepts underpinning the data extraction and cohort identification methods applied are described. This includes information on clinical coding, comorbidity scores, cause of death data and boundary effects. Further description is provided on how these were considered and applied in the development of our dataset. A detailed description is provided for the development of the dataset on which all subsequent analyses were conducted.

2.2.1 - Clinical Coding

The process of clinical coding involves the application of a recognised medical coded ontology to a clinical record. A number of different systems exist which may be designed for a particular setting or specialty or may be setting agnostic. Common examples include SNOMED-CT[122], Read Codes[123,124] and ICD-10.[125] In all cases the aim is to try and have a standardised way of recording clinical information. SNOMED-CT and Read codes are more extensive than ICD-10 codes as the coding is more granular and also includes clinical events, procedures and investigations codes where ICD-10 is just a disease classification, although other elements such as morphology of cancer are also included within this.

The Leeds Teaching Hospitals electronic patient record (EPR) from which our research data is derived, is based on ICD-10 diagnostic coding. This coding in done according to national guidelines which form the basis of the Hospital Episode Statistics[126] database aiming to capture hospital activity data nationally. Amongst a number of data items, this database includes information on new and established clinical diagnoses. In contrast to primary care where coding is often done during routine clinical care delivery, hospital clinical coding is only completed after an admission event. Here, upon discharge, a primary code is given as the main reason for admission. Secondary codes are then applied which are for any other condition the patient is known to have and also any other reason for the admission. In each case the coding is applied based on a review of the clinical notes and completed by hospital clinical coders who are specially trained hospital administrative staff.

This coding database was used as the basis for identification of non-cancer diagnoses in our dataset. Clinical concept definitions for each condition of interest was developed by a clinician using ICD-10 codes. The code definitions table can be found in the appendix. For a detailed description of the data processing done for this, please refer to section 2.2.8 below.

It is important to note that the primary purpose of hospital clinical coding is financial rather than clinical. Until recently the amount that the hospital was paid was dependent on the combination of conditions for that patient[127]. Thus the use of this coding for clinical description is a fundamentally different use for this data than the purpose for which it was collected.

Cancer diagnosis events were identified using ICD-10 codes, however this data is captured more widely due to the requirements to provide data to the National Cancer Registration and Analytics Service (NCRAS).[128] The identification of cancer patients was therefore based on this separate coded dataset.

Histology data recorded alongside cancer diagnostic coding is done using the ICD-10 morphological coding. This is very granular however also allows for the use of non-specific codes. This results in many subdivisions of common histology. This could create issues when using them as categorical variables in analysis as it would create small numbers and low precision. As such for each cancer, histology data was reviewed and grouped based to the common subdivisions guided in part by TNM Version 7.[129]

2.2.2 - Comorbidity Scores

With an ageing population globally particularly in more economically developed countries a growing proportion of patients are being managed with multiple health conditions.[42,130] The interplay between conditions is of huge importance to clinicians, not only on a population level in influencing treatment guidelines, but also in assisting with treatment decisions on an individual level.[18,30,131–133] Understanding and representing these complex interactions and their effects on key outcomes is often simplified through the use of comorbidity scores and frailty scores. These two approaches are similar but subtly different with comorbidity scores usually being based on specific diseases and frailty score often being based on clinical concepts in addition. Key examples are the Charlson score[38], Elixhauser score[39], Hospital Frailty Index[41], electronic frailty index[40], and ACE-27 score.[134,135] These scores often provide weighting for different conditions and also may incorporate some form of severity scoring in addition. These scores have been used in multiple clinical contexts to assess if they have correlation with clinical outcomes of interest or in some cases are used as the basis of an outcome prediction.[136,137]

It is important to note that in most cases these scores are derived from administrative data rather than data curated specifically for this purpose. Furthermore, in most cases the score is based on expert opinion or based on mapping to concepts. They are then subsequently applied to an outcome of interest, to assess if it has some predictive value as opposed to developing a score specifically optimised for its intended use case. This has the advantage of making them easy to calculate and interpret, but does not ensure that optimal predictive power is obtained for all outcomes of interest. Some scores are more complex than others such that they take longer to estimate and that they can only be scored by a clinical assessment contemporaneously, as opposed to retrospectively. This may enhance utility and data accuracy of the scoring, whilst limiting its practicality and implementation.

Once such detailed scoring system is the ACE-27 score. This is comprised of multiple disease domains with a severity grading for each. This score has been extensively assessed in the adult population in particular in cancer patients.[138–142] The nature of the severity score does however prevent robust retrospective calculation of patient scores. In order to select our comorbidities of interest, the ACE-27 score was used as a framework to identify conditions to assess. All conditions named in the score were included for analysis. In some instances conditions grouped in ACE-27 were divided to provide a more granular analysis, particularly were conditions in a grouping were considered heterogeneous. Two of the domains highlighted by the ACE-27 scoring system were not included, namely mental health and substance abuse. These were not included due to issues of data availability. As metal health care records and substance abuse records are stored separately they were not available for analysis and therefore not included.

ACE-27 Domain	ACE-27 Named Subdivision	Conditions Included For Analysis
Cardiovascular	MI Angina/CAD CCF Arrhythmias Hypertension Venous Disease Peripheral Arterial Disease	MI Angina/CAD CCF Arrhythmias Hypertension Chronic Venous Insufficiency Thromboembolic Disease Varicosities Peripheral Arterial Disease
Respiratory Disease	NA	Asthma Chronic Obstructive Pulmonary Disease Restrictive Lung Disease Other Respiratory Conditions
GI Systems	Hepatic Stomach/Intestine Pancreas	Liver Dysfunction Malabsorption Inflammatory Bowel Disease Pancreatitis Peptic Ulcer Disease
Renal System	End Stage Renal Disease	Renal Dysfunction
Endocrine System	Diabetes Mellitus	Diabetes Mellitus (all) Type 1 Diabetes Mellitus Type 2 Diabetes Mellitus Other Diabetes Mellitus
Neurological	Stroke Dementia Paralysis Neuromuscular	Stroke Dementia Demyelination Motor Neurone Disease Parkinson's Transient Ischaemic Attack Other Neuromuscular Disorders
Psychiatric	NA	Not included
Rheumatological	NA	Ankylosing Spondylitis Gout Psoriatic Arthritis Rheumatoid Arthritis Other Rheumatologically Conditions
Immunological	AIDS	AIDS/HIV
Malignancy	Solid Tumour Leukaemia and Myeloma Lymphoma	As per site specific cohorts
Substance Abuse	Alcohol Illicit Drugs	Not assessed
Body Weight	Obesity	Obesity

Table 2: Summary of ACE-27 – *Summary of the domains and subdivisions that make up the ACE-27 scoring system. A final column of the derived conditions / comorbidities used for analysis within this study is included in the third column.*

2.2.3 - Identification of the Cohort

All patients included within the study were identified from the Leeds Cancer Centre's electronic health records. This system, called Patient Pathway Manager (PPM)[143], is an electronic clinical record keeping system that is underpinned by a SQL based database which captures and stores data from clinical systems and allows direct data entry.

SQL extracts were used to identify all patients with a cancer diagnosis made from 2018 and before based on an ICD-10 code starting with "C". Patients were only included where they had a legitimate care relationship with Leeds Teaching Hospitals. Any patients without an NHS number were excluded as this risked the incorrect joining of data from across the PPM data sources.

2.2.4 - Details of Cancer Diagnosis

Data relating a patient's cancer diagnosis was derived from the diagnosis tables of PPM which contains all of the key cancer diagnostic information held by Leeds Cancer Centre. This diagnostic information includes data items such as demographics, grade, stage, morphology, genetics, molecular testing, patient details, path to diagnosis and originating hospital, which is also referred to as an originating unit. Patients with multiple cancer diagnoses were included only once in each site specific cohort. A single patient could appear multiple times if they had developed multiple cancers. Where a patient had developed multiple instances of a given cancer then only the first record of diagnosis was included within the site specific group. Inclusion of the same patient multiple times would risk over representing their other baseline characteristics within a given cohort. Having them represented across multiple cohorts would not be affected by this, as each cohort was analysed as a standalone dataset and not combined.

The cancer sites chosen were based on the published list of common cancers produced by Cancer Research UK.[1] Where a cancer was provided as a named site in this list it was considered for inclusion as a site specific cohort. Cancers not mentioned were included as "other" cancer in the all cancer cohort. This produced a total of 24 specified cancer sites and site specific cohorts.

Each cancer cohort was selected based on the ICD-10 label applied to that instance of cancer within the record. A three character ICD-10 match was used for this. This level of ICD-10 coding was selected as this depth of coding is highly accurate, whereas there is less consistency in greater depths of ICD-10 coding. The data defections of each cancer site can be found in the appendix.

Tumour staging was available in a number of formats including TNM[129], WHO 1-4 and FIGO.[144] This range of data was kept to allow for more or less granular groupings depending on the analysis undertaken. A less granular classification is of utility in trying to reduce the number of dummy variables required within the dataset to allow for regression analysis and also ensures larger numbers of patients within each stage group of a given cancer site. More granular data may however have utility in enhancing predictions. There is a residual issue that staging systems have changed over time. No current method is available for standardising these without a manual review of all cases. This would be impractical given the scale of the dataset and was therefore left in its original form but represents a potential limitation of the approach.

Full histology data was also extracted however as with stage data this level of granularity would have led to difficulties with regression analyses. All morphological descriptions for each cancer site were reviewed and aggregated into clinically relevant morphological groupings by an oncologist. These groupings are outlined in **Table 21** in the appendix.

Survival times were calculated based on date of diagnosis and date of death, where a patient was not known to have died the censor date applied was the date of extraction.

2.2.5 - Comorbidity Data

All conditions described within **Table 2** were turned into individual comorbidity labels. For each of these a coding definition was developed via a review of the ICD-10 coding system to identify all the relevant codes for that condition. These data definitions may be found **Table 22** within the appendix.

Hospital clinical coding data was used as the primary source of comorbidity identification. As described above, this data is created for each patient admitted to the hospital. This data therefore captures the presence of a given condition on a given admission date. All clinical coding data for each patient within the research dataset was algorithmically assessed to identify the first instance that a given ICD-10 code appeared within their clinical record. Those patients with it noted on or

before the date of cancer diagnosis were treated as a prior diagnosis and those where the first recorded data was after the date of cancer diagnosis had the condition treated as a late effect.

Data enhancement was also undertaken where further data sources were available. In the case of diabetes Haemoglobin A1c (HbA1c)[145] data was used to identify patients with diabetes who lacked relevant clinical codes as well as any earlier evidence of diabetes in those with a clinical code. All HbA1c results for patients in the dataset were assessed algorithmically to identify instances of results at or above 48 mmol/mol.[146] Patients with pre-diabetic HbA1c results were also identified using a 42 mmol/mol threshold[147] Where these abnormal results were found, the first date of abnormal results were identified and used to update the comorbidity data within the study dataset. Where both coding and blood data was found, the earliest date of diabetic diagnostic data from either source was used as the date of diabetic diagnosis and whether this preceded the cancer diagnosis or not.

Obesity data was also enhanced using height and weight data held within PPM. All height and weight records within the trust were combined. BMI[148] was calculated for each entry and then labelled as either obese, overweight or other. No category of underweight was included as this does not form part of the ACE-27 scoring system. Implausible BMI records with those above 100 were excluded from the dataset. The date of obesity from height and weight data was then combined with clinical coding such that the earliest date was used as the diagnostic date of obesity.

2.2.6 - Identification of the Leeds Teaching Hospitals Blood Catchment Area

As the main laboratory for undertaking blood analyses for the Leeds metropolitan area, the majority of test results for patients should be available from the hospital EPR. There are however patients who are referred from outside of this geographical area for management of their oncological diagnosis in Leeds. Some areas may also sit on the boundary with other hospital service areas, with some GP practices in these boundary areas sending their blood samples to a different centre for processing. This creates what is termed a boundary effect in which patients at or beyond the boundary of LTHT's catchment area have a lower probability of complete and accurate data. It is therefore important to be able to distinguish those patients who are inside and those who are outside of the area for which the majority or all of a patient's blood test results will be available in LTHT datasets.

In order to identify these patients, aggregated data was created to identify the GP practices that had sent a volume of 10,000 blood tests or more to LTHT. Those meeting this threshold were regarded and being within the catchment area. Patients registered to these practice were therefore identified as being within the LTHT blood catchment area.

2.2.7 - De-identification of Data

All data was extracted such that is was de-identified, removing names, NHS numbers, full postcodes and dates of birth. Information required based on dates of birth were calculated apriori, such as age at diagnosis with dates of birth being retained in month and year format only. Postcode derived data such as deprivation scores were calculated and added at source prior to analysis, such that when used for research postcode sector information and IMD quintiles[149] were available.

2.2.8 - Cohorts Developed

Two categories of cohort were developed the all cancer cohort and the site specific cohorts. The all cancer cohort includes each known cancer patient only once within the dataset. This was achieved by limiting patients with multiple cancer diagnoses to only the first cancer diagnosis.

The second class of dataset was the site specific cohorts. These aim to include all patients with a cancer diagnosis in a particular cancer site. This was achieved by limiting each cohort to a specific or combination of specific cancers as defined by a 3 digit ICD-10 code. Each patient was represented only once such that patients with multiple diagnoses of the same cancer were included based on their first diagnosis. As this process was done on a site by site basis, patients may be present multiple times across the different site specific datasets but only once in each.

2.2.9 - Data Pre-processing

Below is a description of the processing that was conducted in order to generate the research datasets for the study. **Figure 1** demonstrates the broad steps visually alongside a more detailed description provided below.

1) **Cancer Diagnoses**
    a) All definitive primary cancer diagnoses prior to 2019 were identified and extracted.
    b) Extended data for patients extracted including performance status, numbered staging data and deprivation measures merged with standard diagnostic extract.
    c) Survival time calculated from date of diagnosis and date of death or date of extraction.
    d) Patient age calculated and converted into ten year age band.
    e) ICD-10 Codes used to label cancer diagnosis by cancer site.
    f) Staging data aggregated from numbered subdivisions into just numbers e.g. 2a to 2 and standardised to Arabic numerals from roman numerals.
    g) Grade data converted into NA, Ungradable, low, intermediate or high grade (except prostate).
    h) Prostate cancer diagnoses converted to standardise recording of Gleason grading.
    i) Histology grouping definitions used to create aggregated histology label.
2) **Cause of Death**
    a) All cause of death data for patients extracted.
    b) Partial string match used to identify all deaths with a “C” code ICD-10 code as 1a, 1b or 1c.
    c) Deaths labelled as “cancer” or “non-cancer” death.
    d) Cause of death data joined to cancer diagnosis data.
    e) New cause-specific death status field created.
    f) Non-cancer cause deaths changed from 1 (deceased) to 0 (censored).
3) **Height and Weight Data**
    a) Height and weight data from all sources in PPM extracted.
    b) BMI calculated.
    c) Implausible values removed.
    d) Measurements categorised as overweight or obese recorded as binary 0/1 for each.
    e) First date for overweight recordings and obese recordings identified for each patient.
    f) Weight data joined onto cancer diagnosis table.

4) Combined Diabetic Diagnostic Information

a) All admission events with a diabetic clinical code were extracted.
b) The earliest admission event for each patient were kept and the others removed.
c) All HbA1c data for patients were extracted.
d) All lab results were converted to numerical data type with standardised units of mmol/mol.
e) Where results were "<20" and ">130" they were converted to 20 and 130 respectively.
f) HbA1c results were categorised as Normal (<42), pre-diabetic (>= 42 and <48) or Diabetic (>=48).
g) The earliest date for each patient for each category was identified.
h) Summary statistics for per patient HbA1c results including number of tests, mean, min, max and variance.
i) Results for clinical coding and HbA1c were joined.
j) Earliest date indicator from either bloods or clinical coding calculated.

5) Other Comorbidity Data

a) All clinical coding data for patients extracted.
b) Table created for each comorbidity of interest containing admission events with a clinical code matching those from that condition's code definition.
c) The earliest admission per patient was identified with the date stored.
d) Repeated patient entries removed so that each comorbidity contained patients only once with a 1 indicating the presence of the condition and the date it was first recorded.
e) Height and weight data was combined with obesity data.
f) Earliest date for evidence of obesity from either coding or height and weight data was derived.

6) **Combining data**

a) Cancer diagnosis information was joined with the comorbidity tables.
b) Where after the join comorbidity information was NA due to them having no evidence of comorbidity these were converted to 0.
c) Each comorbidity was assessed against the cancer diagnosis date to identify if the condition predated or post-dated the cancer diagnosis. Where the comorbidity occurred before the cancer diagnosis, the comorbidity retained its label of 1 for the comorbidity. It was assigned a 0 for the comorbidity as a late effect. Where the comorbidity was diagnosed after the cancer diagnosis, the comorbidity label was changed to 0 and the late effect comorbidity label was recorded as a 1.

7) Creating Cohorts

a) All diagnoses relating to a site were extracted into a new table. The first occurrence of each patient was retained with later diagnoses for that patient removed.
b) Within the original table the first diagnosis of each patient was retained with subsequent diagnoses removed.

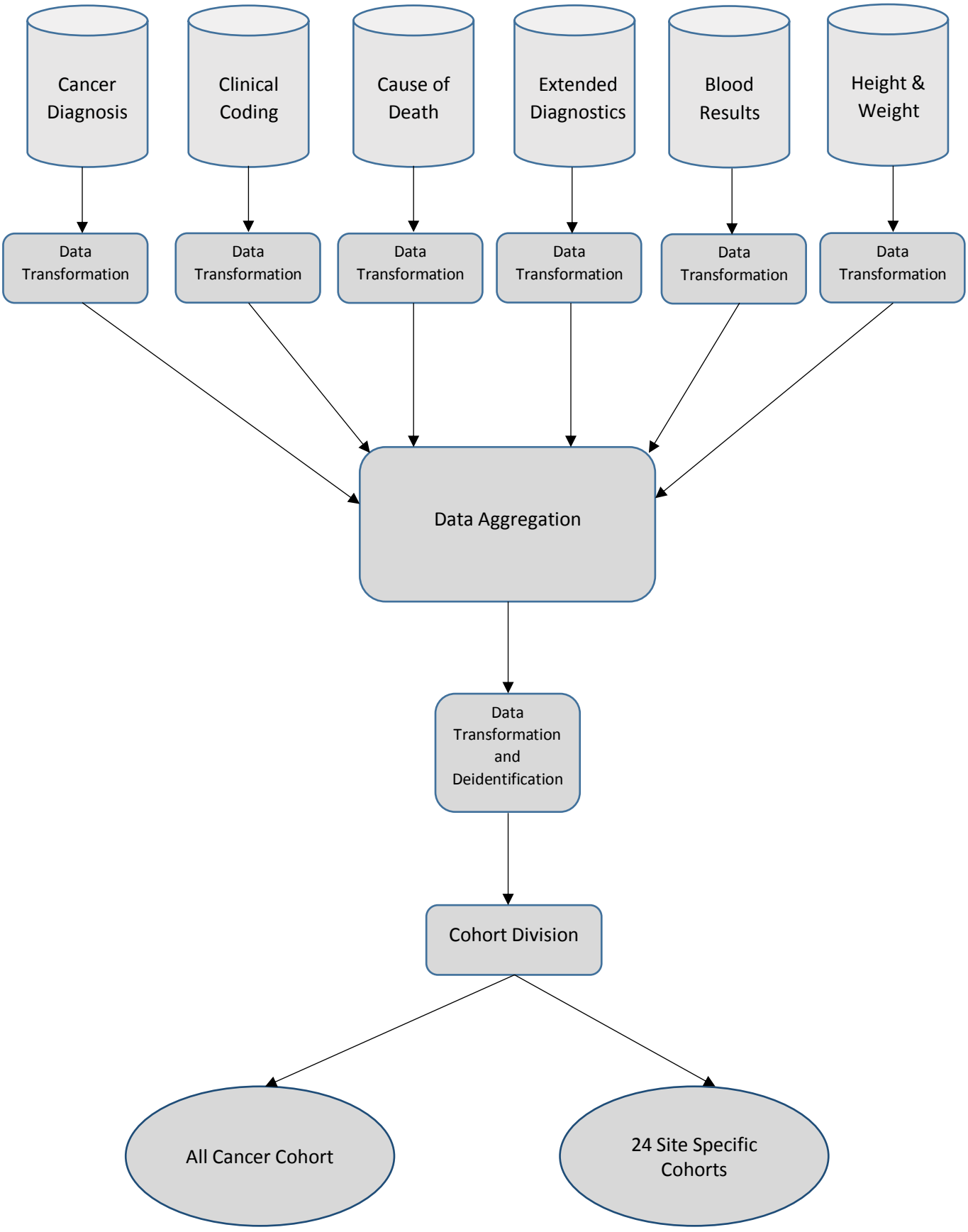


Figure 1: Data Pre-processing Flow Diagram *– Visual summary of the creation of the research datasets. This includes the data sources, data transformations, aggregation, de-identification and divisions of the cohorts that were undertaken in the creation of the All Cancer Cohort and the Site Specific Cohorts*

2.2.10 – Index of Multiple Deprivation (IMD)

Within this thesis the English IMD is used as the measure of deprivation for the purposes of analysis. The IMD framework is made up of seven domains which include:

1. Income
2. Employment
3. Crime
4. Living environment
5. Barriers to housing and services
6. Education, skills and training
7. Health Deprivation and disability

Within these several sub-domains have been created with a total of 37 indicators across the scoring rubric. Two of particular note are those relating specifically to income deprivation affecting children and income deprivation affecting older people. IMD is calculated on Lower Layer Super Output Areas (LSOA) which are then ranked and place in deciles or quintiles.[150] As a result, IMD is a tool for deprivation discrimination between geographical areas and therefore cannot be used to define poverty as there are no absolute values for this.

Despite its widespread use a number of criticisms and limitations of IMD in general have been highlighted. Firstly, as the most granular area is LSOA this can include a wide range of people as this covers an area of approximately 1000-2000 people. Previous studies have shown that this creates areas within which deprivation levels can vary dramatically with deprived areas showing particularly high levels of variation.[151] Others have criticised how the raw data used to calculate the scores for each domain are not available and thus assessing the usefulness and appropriateness of the score is limited.[152] The calculation of the score also relies on the use of the same metrics in multiple domains such as receiving state benefits for ill health being found within both the income and health domains. This double counting introduces a bias in the scoring based on certain data items that underlie the index. Others have criticised the relative effectiveness of the score in rural versus urban areas although rurality and deprivation were not shown to have correlation when explicitly tested in Scotland.[151]

Within the Leeds PPM dataset English IMD 2015 data is available for patients as IMD quintile. This is based on their most recently registered postcode and therefore introduces several further potential pitfalls. Firstly, if patients move outside of England they will not have an IMD score as within the UK IMD is calculated within each devolved nation and due to it being a relative index, comparisons between nations of the UK are not possible. No UK wide IMD currently exists. A further limitation is that if patients are diagnosed and have died prior to IMD 2015's introduction it is likely that their relative and contemporaneous score may be different, however as with the devolved nation data, as IMD is a relative scale, comparisons between versions are not a valid approach. Finally the data within the PPM dataset does not include information on the breakdown of the domains that make up the ranking for IMD which provides less granular information. As IMD quintile was the only metric available within the dataset to measure deprivation it was used however the above stated limitations should be considered when interpreting the results of later analyses

2.3 Descriptive and Exploratory Analyses

In order to assess and understand both simple and complex analyses conducted on a research dataset it is important to first understand the basic characteristics of the raw and pre-processed data. Factors such as data provenance, accuracy, missingness, correlation and collinearity can all have a significant impact on the choice of approach in data handling, analysis, as well as introducing bias and impacting external validity. Here each of these is considered in turn and how in each case attempts were made to assess them.

2.3.1 - Data Provenance

Data provenance describes the origins of data items and the process by which it arrived in a given dataset. When dealing with printed information documents are in essence unable to be altered without being replaced with a new edition. In electronic data this is not the case and aspects may be altered and transformed in many ways without leaving any obvious trace of how, when or why this process took place. Previous research has focussed on three aspects 1) correctness, 2) completeness and 3) relevancy[153]. These concepts are based on the assumption that data capture occurs via an automated process which in healthcare delivery is not always the case. As a result, the context of data capture is crucial in understanding its provenance and conceptually it can be helpful to separate "where", "why" and "who" provenance into separate groups as each can have a significant impact on the data obtained and how this may be impactful[154].

"Where" provenance, the source of data, may initially seem trivial to assess and describe, however it can easily become complex and challenging in the real world of healthcare delivery. When a collection of data comes from multiple systems a description of the original source may be simple to identify. If however information flows into a dataset through a multistep process, it is important to identify any data transformations that may have taken place. Changes from original source data such as rounding, aggregation and changes in data types can alter the data to an extent that impacts on the analysis outcomes obtained. A real world non-healthcare example of this was the loss of 50% of the value of the Vancouver stock exchange over 23 months due to rounding error.[155] In 1983 after the launch of the stock exchange, computer software rounded stock values down using floor() functions rather than traditional rounding using round(). This resulted in small losses of stock value daily, which eventually accumulated deflating the market value to half of what it should have been. If this sort of rounding error was to take place in the context of health economics for example, small values, over many patients, over many years could result in large scale differences. If different data sources apply different rounding then they may no longer be suitable for combining or comparing.

"Where" data provenance issues may also arise when the same information is captured in multiple locations. An example of this is a patient's height where it could be recorded on different occasions or in different systems. Outside of extremes of age, height is usually a fixed measurement, however in the real world height recording may vary. If a system includes just a single value, it is important to know where the value comes from, is it an average, a maximum value, the most recent, etc. Understanding these small changes are important for knowing how to interpret a given data item but also knowing whether or not different measures are comparable. Even where a single value is available, software or units of measurement can change over time, meaning that values at different time points are no longer directly comparable without the application of considered adjustments and alterations. A practical example of this is changes in TNM staging over time in oncology.

"Why" data provenance issues consider the motivation used for the initial collection of data. This can have a profound impact on the data collected and its potential accuracy for the task of interest. Where data is being harnessed in analysis for a secondary use, it is important to consider if the data

is fit for purpose or if the analysis or conclusions need to be altered in some way to fit the task of interest. An example of this is in the context of primary care clinical coding. One reason for collection of this data is in the delivery of care, however a significant amount of coding is conducted for the purposes of financial remuneration[156]. Those conditions that attract a payment may be more likely to be recorded than those that do not. As such, the accuracy and completeness of diagnostic recording may be heavily influenced by the purpose of the data being input. These differences may have meaningful effects if patients are misclassified in the source data and may also result in a different scale of impact when considering different conditions.

"Who" provenance identifies the individual, system or process that input the information. This may impact on accuracy and validity in a number of ways. A computer from a laboratory automatically storing test results is likely to be more accurate than an individual manually typing in a result from a printed report. Knowing this may alter the way that researchers choose to handle outlier results. The accuracy of clinical data input by clinicians versus administrative staff may differ depending on medical complexity, such that information input by one group may be considered more likely to be accurate than another. Clinical coding from a discharge summary letter completed by a trained administrator may well differ from the same information recorded by a healthcare professional reviewing clinical notes and discussing things with a patient. These distinctions are of importance as if a tool is designed to be used obtaining data in the latter setting, then an analysis of historic data collected in the former, may yield different or inaccurate results.

In order to overcome these issues expertise was sought from members of the hospital informatics department. Data items included in the analysis were discussed to understand where the information was obtained from, identify any transformations that may have been applied and how data was originally captured. Additionally where possible front end systems were reviewed to better understand potential sources of error and bias. An example of this was the decision not to include HbA1c results from the diabetes management system as these were typed in by clinical staff, as compared to the laboratory results, which were automatically generated from the laboratory machines. Although this did not remove the possibility of testing error, it did remove the risk of human typing errors.

2.3.2 - Missingness

A key attribute of relevance to any dataset is how complete it is. This can be based on both the completeness of discrete cases, i.e. relevant individuals are missed entirely or completeness within cases, where data items for a given patient are absent or unknown. In practice, identifying missing discrete cases is not possible using a single dataset as it would require information from other data sources that may have accurately captured a patient, such as a cancer registry, to identify missing individuals.

Completeness of a dataset within known samples can be quantified by looking at each data item and assessing the percentage in which the data is found and recorded. Where data items are missing a number of approaches may be used. Firstly, these cases with missing data can be excluded in what is called a complete case analysis.[157] This approach has a number of pitfalls including reducing the sample size, reducing precision of confidence intervals and reducing statistical power. Additionally, depending on the pattern of missingness, it may also introduce bias. An alternative approach is to impute the value that is missing. This can be considered as a second best approach, where the single best approach is to have no missing data at all. A number of methods can be used to impute data, however first one must consider the mechanism and pattern of missing data. These are detailed in **Table 3**[158]. Where a variable is deemed to be "missing completely at random" no specific modelling

of the relationship between other variables and an imputed value is needed. Where the variable is "missing at random", some method to model the relationship between other variables and the imputed value is needed[159–161]. Where the value is "not missing at random", imputation is inappropriate. In most cases the distinction between missing at random and not missing at random is based on domain knowledge and expertise, as opposed to something that can be assessed numerically.

An additional issue may arise due to the volume of missing data. Where large portions of data are missing it may be difficult to create reliable imputations as there is insufficient complete data on which to base the imputation process. In order to assess suitability for imputation a two-step assessment was used to determine the appropriate approach:

1. Volume: If more than 20% of data is missing imputation is inappropriate.[162]
2. Pattern: Combine analytical approaches and domain expertise to determine the suspected pattern of missingness.

If the first test was failed by a data item then the second test was not investigated further.

Type	**Characteristics**	**Example**
Missing Completely at Random (MCAR)	1. Probability of a variable being missing is unrelated to the value of that variable. 2. Probability of a variable being missing is unrelated to the value of other variables within the dataset. 3. Probability of a variable being missing may be related to the value of other variables not within the dataset.	To assess the key determinants of patient satisfaction. The MCAR assumption would be violated if when assessed younger patients were less likely to complete their feedback questionnaire.
Missing at Random (MAR)	1. Probability of a variable being missing is unrelated to the value of that variable. 2. Probability of a variable being missing is related to the value of other variables within the dataset.	To assess the key determinants of patient satisfaction. The MAR assumption would be met if when assessed, younger patients were less likely to complete their feedback questionnaire, however the probability of returning the questionnaire is not related to a patient's level of satisfaction.
Not Missing at Random	1. Probability of a variable being missing is related to the value of that variable.	Patients who are dissatisfied are more likely to return their questionnaire than those who are satisfied.

Table 3: Types of Missingness *– Summary of each of the three types of missingness with a corresponding description of its characteristics and an example based on a patient feedback questionnaire.*

2.3.4 - Correlation / Collinearity

Collinearity usually refers to the non-independence of predictor variables in a regression analysis. This presents problems as it increases the variance of the regression parameters which can lead to inaccurate assessment and identification of relevant predictors in a statistical model.[163] In effect, if one variable is a true cause of an outcome of interest and another is strongly associated with that predictor, then the effect may be shared between the two, deflating the effect size of the true cause and inflating that of the correlated variable.

As such it is important to understand whether this may be an issue within a dataset such that the necessary consideration is given to both the analysis methods applied and interpretation of the output results. In order to conduct this assessment our dataset was subjected to pairwise correlation estimates with a Spearman's rank order correlation test.[164] This was chosen due to the data items predominantly being ordinal or binary, which may be considered as an extreme form of ordinal. In each case the analysis focussed on both the statistical significance and scale of correlation.

2.3.5 - Bias Confounding and Statistical Significance.

Within any analysis it is important to consider the potential for bias and confounding and how these impact on the analysis design and interpretation. There are three main forms of bias 1) selection bias, 2) information bias and 3) confounding.[165]

Selection bias occurs where the method used to select individuals into a study incorporates systematic differences between groups, where these differences impact on the outcome of interest[166]. If for example a study compares outcomes from two different hospitals, then the characteristics of the local population may lead to selection bias. If one hospital is located in a more affluent area, this may introduce selection bias due to the socioeconomic characterises of the population that attend that hospital. Selection bias can take many forms and is a particular risk in observational studies where the allocation to a group within a study is often based on the exposure of interest. This is in contrast to RCTs where the exposure of interest is determined at random after selection.

In order to look for identifiable selection bias, analysis was undertaken comparing the locally derived cohorts to national data to identify any differences. Where differences are found to exist, then this will be referenced in the interpretation of results when discussing potential inferential conclusions.

Information bias is a systematic difference that is introduced through the collecting, processing recording or recall of data.[166] This includes misclassification error, recall bias and approaches to data missingness. As with selection bias this study has a relatively high risk of information bias as it is reliant on routinely collected clinical data which is more prone to error than formally collected trial data.

In order to assess for potential information bias an analysis of data missingness and misclassification error was conducted. Further information on missingess is found above in section 2.3.3 and an assessment of misclassification in section 2.3.5

Confounding may be considered as a mixing of effects such that the effect of an exposure on an outcome of interest is combined with the effect of other factors. This causes the relationship between the exposure and outcome to become distorted from that of the true relationship.[165] Confounding may distort effects in either direction such that true effects no longer become apparent or that an effect is falsely identified as being present. The presence of confounding can

make it particularly challenging to identify causal relationships in data. In order for a variable to be a true confounder it must have a number of attributes these include:[165]

1) They must have predictive value even in the absence of the exposure.
2) They must not be an intermediate step between exposure and outcome.
3) They must be associated with the exposure, but not a proxy for it.

In order to address confounding within our analyses domain expertise was applied to identify potential sources of confounding. Where inferential analyses were being conducted, adjustment for measured confounding was applied. Any variables that were temporally downstream of the exposure of interests was discounted as a potential confounder as it could be considered a downstream intermediate between the exposure and outcome. Where a potential confounder was believed to be present but not able to be adjusted for it has been given special comment within the discussion section and used to frame the interpretation of the results.

In many cases issues may also rise due to differing levels of confounding between two groups, such that not only is confounding impacting on estimates, but it is effecting the groups being compared differently in addition.

Within the medical literature it is common practice to apply statistical significance tests to analyses. This test is an assessment of the probability of the null hypothesis being incorrect with convention being that 95% probability or greater is a probability cut off to reject the null hypothesis. It is important to note that where an analysis fails to meet this metric it does not mean that there is truly no difference and additionally if the p value does meet this threshold, it does not mean a true difference does exist. This leads to an inevitable possibility of false discovery including type 1 and type II error, where an effect occurring by chance is attributed to a true effect.[111,167]

The appropriateness of the application of significance tests is improved when a study includes individuals within each group who are truly pulled from the same population. This for example occurs in randomised controlled trials. In observational studies this can be an issue as the individuals are commonly drawn from different populations, with significant selection bias which limits the appropriateness of statistical significance testing.[59]

An additional issue can arise with multiple testing. Although in each analysis the probability space meets the particular threshold applied, when multiple analyses are done there is an inherent false discovery rate even when results are not combined. As such, if a 5% threshold is applied and 100 analyses undertaken, then 5 significant results would be expected to occur on average by chance where the null hypothesis should in fact have been accepted. Some therefore suggest that the p value threshold should be adjusted. A number of methods for this exist including the Bonferroni correction.[168] This involves dividing the 5% threshold by the number of analyses undertaken, to create a more stringent threshold. This will reduce the chances of false discovery, however will increase the risk of falsely rejecting a true difference.

p values are also impacted by the power of a study in that large sample sizes impact on the p values seen[61]. This causes issues in studies involving large sample sizes as in such situations small differences can be highly statistically significant. Thus results that have little relevance and are of low scientific importance may in fact be seen as spuriously important.

This study is observational, involves large populations and multiple comparisons and thus is potentially prone to all of these issues. In order to address these a number of steps have been taken to improve the interpretability of results. In some cases results will be provided with p value

thresholds at the standard 95% and corrected level. This allows analysis to take into account significance such that both a strict and more lax criterion can be applied. Where significance tests are applied, analyses consider the clinical significance of the results, such that small magnitude but highly significant statistical tests can be contrasted with less statistically significant but highly clinically relevant differences[169,170]. Further, attempts were made to assess precision. Here the confidence interval can be assessed to look at not only whether the null hypothesis should be rejected, but also as to whether the results are precise enough to be reassured that the results are likely to be of value. Results that have low precision may be considered with a higher level of scepticism than results from analyses that have higher levels of precision.

In view of these potential issues a number of approaches have been applied throughout our analyses.

1) First importance in discussion has been given to try and identify potential bias and confounding as potential sources of effect on estimates.
2) Where an inferential analysis involves the comparison of two groups, no p values are provided. Comparisons between values or variables may however be presented with p values for descriptive analyses.
3) Where a p value is presented, it is done in such a way as to enable assessment for a p value threshold with and without adjustment.
4) Confidence intervals will be used as a guide of effect but will not be used to comment on statistical significance.
5) Confidence intervals will be used to assess the precision of estimates.
6) Effect size and clinical significance will be used in part to determine the importance of a result.

2.3.5 - Accuracy

In order for any dataset to yield reliable results, it is important that the information stored within the dataset is correct. In many instances, it is difficult to assess whether routinely collected data is correctly input as there is no other source to which the data can be compared. In cases where a comparison can be made, such as in the case of height being recorded multiple times, it is often impossible to know when a discrepancy occurs which is the true value. Certain aspects discussed above have already highlighted areas where some concerns have been raised within the literature already, particularly in relation to cause of death data and clinical coding. As no comparison source was available for cause of death data, our accuracy assessment focussed on clinical coding. Here, an assessment was conducted to how the recording of clinical coding compared to the comorbidity recording of blood tests in the diagnosis of diabetes mellitus. These were compared in terms of the proportion of the diabetic population identified, how this changed with geography and how they differed in terms of diagnosis date. This analysis allowed for the identification of routes to enhance future analyses and inform the interpretation of results.

2.3.6 - Excess Zeros

In exploratory analyses where the relationship between two factors includes count data with an excess of zeros, there is a need to consider the appropriated distributions used to model these relationships. A standard Poisson regression may result in poor performance due to the excess of zeros within the data. In such cases, it may be possible to use zero inflated models instead.[171,172] This involves the development of a two-step model where the first is a logistic regression to estimate the relationship between variables and likelihood of being a zero count. The second involves a model to assess the relationship with the non-zero count data. To assess whether the standard or zero

inflated model is the more appropriate, a Vuong test[173] may be applied to identify the best model. The zero inflated model allows for interpretation in terms of the risk of being a zero count and also the effect on increasing count. Where exploratory analysis demonstrates an excess of zero count data then zero inflated models will be employed and compared to standard models to ensure the most appropriate model is selected.

2.4 - Survival Analysis

Survival analysis aims to quantify the probability of survival from a given dataset. In reality, the end point need not be death and thus time-to-event analysis may be a more appropriate term to use. In theory, if all information was known about patients then a standard analysis using the total survival or event time as the basis for analysis would be feasible. In reality both within observational data and clinical trials, the time to event data is incomplete. This can arise if the event does not occur before the end of the study period, the data is not recorded or a patient drops out or is lost to follow up. In these instances the total time to event is unknown which is what the inclusion of censoring attempts to overcome. Within this section we will outline the key methods used for descriptive, inferential and predictive survival models utilised within the subsequent analyses.

2.4.1 - Kaplan Meier Estimates (KM)

In 1958 Edward Kaplan and Paul Meier[107] published a paper outlining a method for analysing data with incomplete survival information. This approach is based on assigning three variables to each individual within the study:

1) Length of time known for the individual (serial time).

2) Event status at the end of their known time (event or censored).

3) Study group to which they belong.

By assigning these variables it is possible to include all patients from time 0 until something happens for that individual either in the form of an event or them being censored.

KM analysis is based on the creation of intervals. Each interval end point is defined by the occurrence of the event of interest and the start is either time 0 or the previous occurrence of the event of interest. Importantly, censoring events are not used to define intervals. Visually intervals are the horizontal lines on a KM chart with each vertical shift denoting the start of a new interval. One key aspect of KM analysis which particularly enhances its utility is that intervals may vary in length. At each interval an interval probability is calculated based on the number of individuals alive at the start of an interval divided by the number at risk. This is where censoring influences the analysis, the patients who were censored during the previous interval are no longer considered at risk and therefore are removed from the denominator. It is important to note however that the probability presented on a KM curve is not the interval risk. If this were presented then survival could go up over time. Instead it is the cumulative probability that is presented which is derived from multiplying the interval probabilities together over time. In the event that censoring occurs at the same time as an event defining a new interval then the censoring is considered to have occurred after the event.

Where two groups are being compared then a log rank test is usually applied to assess statistical significance.[174] This is generated by calculating the chi-square for each time of an event in either group which is then summed to derive the ultimate chi-square for comparing the full curves. In practice this involves comparing the observed number of events to the expected number of deaths

were there to be no difference between the groups. If instead an assessment is wanted to compare the event rate in each group relative to one another then a hazard ratio can be generated instead.

A number of considerations should be applied when assessing KM estimates. Firstly, the shape of the curve should be considered as this denotes important information about the population used to derive the analysis. Curves with many intervals suggests a larger population. Large vertical drops after each event suggests either high levels of censoring or lower population numbers overall. High levels of censoring should be interrogated further to understand why this was, particularly in the context of an intervention, as it might suggest an issue such as the treatment being so toxic that it cannot be tolerated and patients drop out. This is what is termed "informative censoring" as the censoring is directly caused by the exposure of interest[175]. In the context of an observational study it may still be of importance as high levels of censoring in one group versus another may suggest differences in how groups are followed up or managed.

The total number at risk should also be assessed at each time point. As the number of events and censoring events accumulates over time, the at risk population falls.[176] It is also important to note that after the first censoring event occurs, the probability shown is not the true probability of survival, but rather the estimate, as the true probability is unknown due to the censoring events. This means that the higher the level of censoring the greater the levels of uncertainty. The combination of a shrinking at risk population, combined with censoring means that with increasing time along the curve, the level of precision falls and the uncertainty of estimates increases. Thus more weight should be given to the interpretation of earlier survival times with a larger population than those later on with low at risk numbers and higher levels of censoring.

Another key limitation of this approach is that it is limited to a univariate approach and thus cannot be used to adjust for potential confounding. Within the subsequent analyses this method will be used as an initial screening to identify differential survival based on the presence or absence of comorbidities of interest. Differences between the curves will be assessed using log rank tests however clinical significance will be assessed using differences in median survival times. This can therefore be considered a descriptive or exploratory analysis of survival outcomes, as without adjustment for potential confounding it is not possible to infer what results beyond our cohort may show.

2.4.2 - Cox Proportional Hazards

The Cox proportional hazards[108] approach to time to event analysis is the most widely used method in the medical literature. It has the advantage over Kaplan Meier estimates in that it allows for a multivariable approach. This enables adjustment for confounding to be done within the model building process. The inclusion of additional variables, often termed covariates, allows for these to be conditioned on and adjusted for. Where these are felt to be potential confounders, their influence is in effect removed from that of the variable of interest. In theory, this leaves just the true effect plus any residual confounding. A more detailed discussion on confounding and adjustment can be found in chapter 5. As with KM estimates Cox models allow for the handling of censored data. The model is termed a semi parametric model as the underlying hazard model is not pre-specified allowing it to take any form, however it is assumed to be proportional over time . The relationship between covariates and hazard is then assessed using linear regression against a logarithmic scale. This has advantages over other methods which require the baseline hazard to be pre-specified as it removes the likelihood of modelling error as a result of an incorrectly specified baseline hazard.[177]

An important component of Cox regression analysis is the undertaking of diagnostics on the models to understand if violations of the underlying model assumptions have occurred or if models have included variables that are highly influential.[177]

As mentioned above, Cox models rely on the underlying proportional hazards assumption. This can be assessed using Schoenfeld residuals.[177] These residuals can be assessed for correlation with time to identify time varying effects. This is calculated both globally for the model and individually for each covariate. It may be possible to assess this numerically using the p value where a value of less than or equal to 0.05 is considered to be evidence of violation of the proportional hazards assumption. As detailed above, where large samples are used this may result in small p values in almost all cases rendering this approach unsuitable.[61] Instead the effects can be assessed graphically by plotting Schoenfeld residuals against time. The application of a smoothing curve with splines can then be used to assess the relationship. Where there is no violation, one would expect to see a flat horizontal line at or close to zero[176]. If the relationship shows large scale divergence from this pattern then it may be considered to have provided graphical evidence of violation of the proportional hazards assumption.

If this assumption is violated a number of approaches can be taken such as including interaction terms between the variable that does not have proportional hazards and time. Alternatively the cohorts may be stratified. Both approaches have potential limitations which are discussed in more detail in chapter 5. Alternatively approaches that do not rely on this assumption may be implemented.

As a semi parametric model the Cox model relies on linear assumptions. Although the baseline hazard model is not pre-specified, the relationship of covariates to this is assessed in a log linear manner[178]. It is important to therefore assess for evidence of whether covariates are related to outcomes in a non-linear way. This can be assessed using Martingale residuals which are compared with increasing value of that variable. Where the general trend strays from a straight line relationship, this suggest non-linearity.[177] Due to the nature of this assessment it is not applicable to categorical variables as these will be treated as dummy variables that are dichotomous.

Where linear assumptions are violated, the issues can be overcome with stratification by converting the continuous variable in grouped categories which are included as covariates. Alternatively transformation of the data such as to a log or rooted scale can be attempted to see if this overcomes the issue.

The final diagnostic is to identify any individual cases that may be disproportionately influential in their effect on the underlying Cox model. This can be assessed with the use of dfbefta residuals. Some heuristics may be applied here such that the maximum acceptable dfbeta is $2/\sqrt{n}$ where n is the number of cases. In individuals where this threshold is breached consideration should be given to excluding them from the analysis.[177]

2.4.3 - Competing Risks Analysis

When using time to event analysis for the purposes of survival outcomes a number of different metrics may be used. The most common and simplest is the overall survival, in these analyses the focus is on any cause of death. Within the oncological literature a number of other events may be used including progression, recurrence or time to next oncological treatment. Here, statistics on progression free survival, disease free survival and treatment free interval may be quoted. When looking at overall survival there may be limitations in the utility of this particularly in the context of multimorbid patients. If a patient has both cancer and a second condition, such as heart failure, and

the focus for the study is cancer outcomes, then a death from heart failure will compete with cancer as the potential cause of death. Once the patient has died from another cause it becomes impossible to know how long they would have survived for with respect to the cancer. In order to overcome this issue a competing risk analysis may be undertaken, with the aim of estimating the risk of death whilst accounting for the competing risks. There are two key ways of doing this analysis, namely, cause-specific competing risk and sub-distribution methods such as the Fine and Grey method .

In both cases there is a reliance on having cause of death data as this allows the attribution of death to the cause of interest or not. Within the UK all deaths are registered via a death certificate which lists the cause of death in section 1 and section 2.[183] Section one is considered the direct cause of death and section two comprises conditions that may have contributed but not directly. Within section 1 it is subdivided into section a, b and c. Section a must be completed and this is the direct cause of death. If appropriate, b is filled out and this is the direct cause of a and c can be used where there is a direct cause of b. By way of an example, a cancer patient who dies from a chest infection after recent treatment for lung cancer with known heart failure would have 1a: Bronchopneumonia, 1b) Iatrogenic neutropenia, 1c) Squamous cell lung cancer 2) Congestive cardiac failure. In this example the lung cancer leads to the delivery of treatment inducing neutropenia, which in turn lead to the susceptibility to infection which in turn caused the pneumonia which cause the death of the patient.

The accurate generation of any competing risk analysis is integrally linked to the accuracy of information about the contributors in our case cause of death. As such, misclassification of cause of death may introduce error into the analysis[184–186]. A more detailed description of this can be found in chapter 6. Within our study any cancer related cause of death found in section 1a, 1b or 1c was regarded as being a cancer related death. Cancer within section 2 was not included as this is not a direct cause.

Cause-specific competing risk attempts to estimate the hazard attributable to a specific cause through the use of altered censoring. In our above example the patient who dies of congestive heart failure prevents further estimation of that individual's cancer risk. As such, death by a competing risk to cancer can be thought of as a censoring event as opposed to an event. By altering how the data records events such that deaths from anything other than our cause of interest are excluded, it is possible to generate an estimate of the hazard for that particular cause. In addition to the issues of correct cause of death reporting, this method can suffer from a number of other pitfalls. If the treatment for the cause of interest increases the likelihood of death from other conditions, this may result in increased censoring, reduced numbers at risk and lower precision.[175] The increased death rate from other conditions may also be of interest despite their exclusion as they are indirectly cancer related, but this cannot be identified from the cause of death data. A further criticism of this approach is that if the same method is applied to multiple cause-specific risks the sum of those risks will exceed the overall risk of death for that population and may even exceed 100% of the population.[180] This stems from the fact that this approach is trying to estimate the risk in the absence of the competing risks and is not attempting to account for them within the analysis. As such this approach may be thought of as appropriate if the output of interest is the cause specific hazard and how individual covariates contribute to this. If however the outcome of interest is the cumulative incidence of events, then this approach would be unsuitable.

In this alternative scenario where the probability of an incident occurring is of interest the alternative sub-distribution approach may be more appropriate. This approach involves a modified form of Cox proportional hazards where competing events do not remove the patients from the at risk population, but are instead kept in the population with reducing weighting as time progresses.[181]

This allows the sub-distributions of the total hazard to be calculated resulting in accurate generation of the true cumulative incidence overall.[180] This difference can be illustrated with the example of wanting to understand the risk of developing a treatment toxicity at one year. If there were a population of 10 patients, 5 of whom die within the first year and 3 of whom get the side effect at a year, then the cause-specific approach would suggest that the probability of toxicity at one year is 60% that is 5 are no longer at risk and 3/5 have toxicity. The Fine and Grey approach would state that the level of toxicity is 30% because of all of those who could have got toxicity 3/10 do. Thus if the question is relating to how diabetes impacts on the risk of developing toxicity at one year relative to other groups the cause-specific approach is more appropriate. If however the research question focusses on how diabetes impacts on the cumulative incidence of toxicity at one year the Fine and Gray method would be more appropriate. It is however worth noting that in both methods where the exposure of interest is related to both the cause of interest and other causes of death, it is possible to generate highly misleading results from either method.[180]

This work is trying to demonstrate whether or not comorbidity is associated with differential outcomes directly in cancer irrespective of other causes and how at each time point the risk of cancer death differs between those who are still alive. As such, the cause specific approach is the more appropriate choice and the one that has been implemented and outlined further in chapter 6.

2.4.4 - Forest Methods

Random survival forests are a method first described in 2008 applying a modification of the Random Forest ensemble learning method to censored survival data.[109,187] The method is based upon multiple decision tree regressions with randomness integrated through introduced variation between the different decision trees. The analysis population is used to draw a random sample, usually two thirds, on which to conduct decision tree regression. The selected sample is split into two groups where log-rank is calculated and used to select the value of the variable where the log-rank score is greatest which maximises the survival difference between the two groups. This is done multiple times until the population has been divided into small groups. Each selection point of a variable number for splitting is termed a node, the split is called a branch and the end groups are termed leaves. In order to introduce further randomness not all variables are assessed at each node, instead a random selection of variables are chosen which are termed the “candidate variables”. Of those candidate variables the most discriminatory variable value is selected for that node and the data is then branched. The individual tree often has low predictive power, so to improve this the process is repeated multiple times. Due to randomness in the sample selection and candidate variable selection, when repeated, the subsequent trees will be different. The multiple trees can then be combined to create the “forest” where the combination of many weak learners in an ensemble aims to create a strong learner with better predictive accuracy.[188]

The model performance can be altered through a process of optimisation via hyper-parameter tuning.[189] Here the number of candidate variables selected can be changed, the number of trees built and changing the minimum number of cases in the terminal leaves altered. The optimal combination of these hyper-parameters can then be identified and used within the final model.

Random forest methods have a number of advantages over other methods. Firstly, they have no underlying model assumptions and can therefore be easily used to perform analysis when data has been shown to violate proportional hazards assumptions. Additionally, they can be used to identify non-linear patterns of association. The method has also been shown to be robust to collinearity and the inclusion of non-informative variables.[109]

Unlike other forms of machine learning the resulting model is interpretable rather than being a "black box". It is possible to identify the most important predictors using Variable Importance Scores[109]. Here each variable is randomly permuted and the impact on model performance assessed. Where this does not impact on model performance it suggests the variable is not particularly important, where permutation has a big impact, then it suggests the variable is important. Although the underlying model is not impacted by collinearity, the VIMP scores may be, such that correlation between variables may deflate VIMP scores.[190] This can be partly overcome by also looking at pairwise VIMP. Here the process is the same, however each combination of two variables are permuted together to identify pairings of variables that are important together. This also offers a method of assessing interactions between variables. Although this in theory could be expanded to increasing numbers of combined variables this is a computationally expensive approach and going beyond pairwise VIMP quickly becomes impractical.

An additional measure of information gain can be extracted and analysed using average minimum tree depth.[109] This involves calculating the average depth of node that a variable is first used across the forest. The closer to the root, the greater the information gain that variable has. A threshold for important information gain can be calculated in addition to identify those that are important variables.

The models can also be used to create stratified survival curves and conduct partial plot analysis where the relationship between individual variables and outcome can be assessed. The models can also be used to create predictions for new patients and if desired these predictions can also be given some level of interpretability using counterfactual reasoning or application of methods such as Locally Interpretable Model Agnostic Explanations (LIME).[191]

This study will apply the RSF method to PPM data and uses VIMP, pairwise VIMP and average minimum tree depth to identify variables and combinations of variables that are highly informative in generating predictions. The application of partial plot analysis to the variables of interest is used to identify how variables relate to the patterns of prediction across the cohort. In view of the predictive framework of this approach several key principles are applied in the interpretation of results.

1. Models will first be tuned to optimise predictions prior to interpretation.
2. Where model accuracy is shown to be poor, low weight will be given to the model outputs.
3. VIMP scores will be used to compare importance within a model but not between different models.
4. Important predictors will be used to identify baseline characteristics that define a group but not suggest analysis provides evidence for a mechanistic or causal link even if a plausible explanation exists.

2.4.5 - Predictive Accuracy

In order to compare models using a predictive framework it is necessary to use some metrics to compare accuracy. Until recently in the domain of survival prediction C index[192] was the method of choice with this approach assessing the discriminatory power of a model. It is based on creating all possible combinations of pairs of subjects from the population and estimating the proportion of the cases where the model accurately assess which patient will outlive the other. Recent studies have however shown that this approach is flawed as two correlated C statistics do not always converge to zero under the null hypothesis resulting in an increased risk of type 1 errors.[193] Although work around strategies have been suggested including using the confidence interval distributions of the C index, as opposed point estimates, this still relies on equal censoring distributions to occur in each

group which is most often not the case. As a result, although the C index has been used as a marker of discriminatory power within the model building process, it is not employed as a primary performance indicator.

A key issue with survival prediction is that models may perform differently at different time points and that different time points have different levels of importance clinically.[194,195] A key example is death within 30 days of diagnosis. In this cohort it is important not to deliver aggressive treatment as this will result in reduced quality of life but with no potential for benefit. Thus being able to identify patients who are in this group is of huge importance. If a model therefore has an overall accuracy of 90% over 5 years however systematically fails to identify these early at risk individual, it may be a less useful model than one with a lower overall accuracy but with better accuracy in identifying early events. As a result a series of tests and principles are applied to our model accuracy assessment

1. Separate cohorts will be used to develop and test the error rates.
2. Integrated Brier score will be used to assess overall model performance.[196,197]
3. Time dependent Brier will be used to assess time dependent variation in prediction.
4. C index will be assessed within model building but viewed as less important that the other performance metrics.
5. Clinical utility and constraints will be assessed alongside performance metrics to assess whether models are fit for purpose.

2.5 - Prognosis Research Strategy Group (PROGRESS)

Within section 2.1.1 a breakdown of classes of analysis and their relationship to types of conclusion is presented. This thinking has been applied to the particular area of health outcome research in the work published by the Prognosis Research Strategy (PROGRESS) group. Over the course of four publications this multicentre expert group attempts to frame both the value and minimum standards of research relating to prognostic modelling. This work builds upon the concepts presented in the section 2.1.1 and helps to contextualise the value of the material presented within this thesis, despite, by design, not including formal causal analysis.

Each of the four PROGRESS publications focusses on a different type of analysis namely, descriptive survival analysis, survival analysis for prognostic factor identification, survival predictions and survival predictions that include treatment information. PROGRESS 1[198] focusses predominantly on best practice when describing the outcome trajectory for patients with a given condition and this is analogous to the work presented within Chapter four with Kaplan-Meier methods. PROGRESS 2[199] focusses on research to identify prognostic factors in relation to the outcome of interest. This is applicable to the work presented in chapters 5 to 6 and the variable importance analyses conducted in chapter 7. PROGRESS 3[200] details prognostic modelling and the value of individual level outcome predictions. This aspect of their work relates to the remaining material covered in chapter 7. PROGRESS 4[201] related to stratified treatment outcome predictions which is not considered within this thesis.

The previous PROGRESS group publications are important in highlighting the value that prognostic information can play in clinical decision making, informing future trials, identifying the most at risk groups and identifying the factors that define the most at risk cohorts. This information is shown to be of clinical and patient benefit even outside of an explanatory framework that is intrinsic to causal modelling. This approach to clinical outcome research highlights how the results presented within this thesis have the potential to inform and improve clinical care and research in future whilst

avoiding the many pitfalls of attempting causal modelling with observational data and its many biases and pitfalls[120,121].

2.6 - Software and Tools Used

All data extraction, analysis and data visualisation was undertaken using R software. Many of the analyses required multiple packages not included within base R. A full list of all the packages used and their version numbers can be found within **Table 23** in the appendix.

2.7 - Ethics

This research was conducted after obtaining formal ethics review (IRAS: 277122). The project applied a number of data safeguards including removal of patients with recorded opt outs, de-identification and minimising the number of data items to include only those necessary to undertake the analyses. All analysis was undertaken using secure NHS infrastructure on password protected and encrypted computers. The project had regular oversite from The Leeds Teaching Hospitals Chief Clinical Information Officer, senior oncology clinicians, a number of academics with expertise in this area of research and two patient representatives. The projects is entirely compliant with all relevant laws including general data protection regulations (GDPR)[68] and the data protection act.[202]

Chapter 3 – Descriptive and Exploratory Analysis of the Research Dataset

3.0 - Introduction

3.0.1 - Context

This chapter builds upon the topics and concepts outlined within the first two chapters of the thesis. Before conducting any inferential or predictive analyses it is first important to undertake some initial descriptive and exploratory analysis to understand the dataset on which subsequent modelling will be conducted. This chapter will focus on describing the baseline characteristics of the study cohorts, as well as assessing key features such as missingness, collinearity, accuracy and identifying potential sources of bias. By understanding these features of the dataset, we can adapt and identify approaches to subsequent inferential and predictive analyses that minimise the effects of these, or alternatively, use this information to inform our interpretation of analysis output. Some aspects of these dataset characteristics are important for understanding limitations of data more broadly, thus informing interpretation of previously published literature in this area of research. Before detailing these aspects of the dataset, a brief description of the system from which the data is derived is presented in order to provide context to the data provenance on which all subsequent analyses are based.

3.0.2 - Data Provenance: The History and Origins of PPM and the Leeds Dataset

All of the data used within the subsequent analyses were derived from the Leeds Teaching Hospitals Trust (LTHT) EPR system. LTHT has been using electronic patient records for oncology care going back as far as the 1990s. The system originally in place was called Patient Pathway Manager (PPM) and was designed to capture the required data items needed for the national cancer registry. Over time this system was developed to create a full oncology electronic record system that included clinical notes, letters, results of investigations etc. The capability of the software was recognised by other clinical specialties, such that when in 2011 the hospital were looking to adopt a trust wide EPR, two thirds of specialties were already using PPM. The decision was therefore taken to develop PPM into a specialty agnostic EPR system for the entire hospital. This new iteration is known as PPM+[143]. The development history of the EPR is reflected in the changes in data available over time such that the depth and breadth of information has grown year on year. Key information, such as chemotherapy data and admissions are only available from 2004 onwards, whereas cancer diagnostic data is available from the 1990s.

As PPM has been adopted as the trust wide EPR, clinical software tools used for specialist purposes have been integrated into it. Each of these commercial software tools will have its own clinical dataset that underpins the running and utilisation of it. Although not all of these data items may be visible in the PPM front end the use of common identifiers across software databases means that these additional sources of clinical data can be used for the purposes of analysis. These are commonly referred to as linked systems with key examples being surgical data, pathology data, results of investigations and prescribing data, although many more exist. Throughout the thesis the dataset will be referred to as the PPM dataset which includes all the data available collected via both PPM+, the original PPM and other linked systems.

Crucially the PPM dataset is the shared care record system for the Leeds City region and is known as the Leeds Care Record. The PPM dataset therefore also contains the basic information of patients within the region even where they have not accessed hospital services.

3.0.3 - Aims and Objectives

Aim:

1. To describe the characteristics of the research dataset and examine how these may introduce, bias, confounding and error into subsequent analysis and interpretation.

Objectives:

1. Describe the basics demographics and compare to regional and national demographics.
2. Quantify the completeness of the dataset.
3. Visualise the impact of geography on data accuracy of dataset.
4. Quantify the accuracy of clinical coding for comorbidity identification using diabetes mellitus as a case study.
5. Assess the prevalence of comorbidity in the cancer population and compare to the background population prevalence.
6. Assess the relationship between age and comorbidity.
7. Assess the timing of the diagnosis of comorbidity relative to cancer diagnosis.
8. Quantify collinearity within the dataset.

3.1 - Methods

3.1.1 - Identification of the PPM cohort and sub-cohorts:

The PPM Cohort

This dataset includes all patients within the PPM dataset with whom LTHT have a legitimate care relationship. De-identified demographics information was extracted programmatically using SQL (within R) and R to include information about age, gender, postcode sector, year of entry into the PPM dataset and death status. This information was subsequently joined with external open source data such as the English Index of multiple deprivation[149] for each postcode sector. Patients without an NHS number were excluded.

The Cancer Cohort

A subpopulation of the PPM cohort was identified to include all patients with a cancer diagnosis. The PPM database contains information on cancer diagnoses that are used as the basis of the national cancer registry submission. All patients with a "C" ICD-10[125] code for a malignancy were identified from this dataset. As this population is derived from the PPM Cohort all those without an NHS number were already absent from the dataset. Diagnoses were limited to confirmed primary malignancies, with all recurrence and progression events excluded. Data was extracted relating to the basic demographics to match the information detailed above for the general PPM cohort. Where a patient was found to have multiple malignancies their first episode by date was kept as their diagnosis date and other subsequent diagnoses excluded. Thus patients with multiple malignancies are only represented once within the cohort.

Cancer Site Specific Cohorts.

These were derived from the full PPM cohort applying the same exclusion criteria as the cancer cohort. Prior to the exclusion of multiple malignancies patients were subdivided into the most common cancer diagnoses as discussed in chapter 2. This was achieved using the specific 3 digit ICD-10 definitions for each cancer as detailed in **Table 24** of the appendix. Patients were able to appear only once within each site specific cohort such that only the earliest episode of that cancer was retained. As each site was treated as a separate cohort the same patient may be present multiple times across the cohorts if they have had multiple malignancies in different tumour sites.

3.1.2 - Demographics and Basic Descriptors

Analysis was undertaken to understanding the basic make-up of the whole PPM patient population in addition to the cancer patient cohort. Comparison was made to the population of the UK and the population of the Yorkshire and Humber Region by making use of open source data from the Office of National Statistics.[203] Summary statistics and exploratory analyses were produced using R to include population pyramids, median age, median IMD and percentage by gender. Information on the date of registration in PPM was converted into a year of registration with annual and cumulative patients entering the dataset calculated from 2004 onwards.

3.1.3 - Missingness

In order to assess missingness[158] the proportion of missing data for age, gender, deprivation, histology, stage and grade were assessed. To assess changes in data quality over time missingness of stage and grade data was assessed annually.

3.1.4 - Patient Geography

To assess geographical variation in data availability, information was extracted by Colin Johnston from LTHT. This information included the proportion of events for patients from each postcode where LTHT was the originating / referring unit, i.e. their initial management was in LTHT.

This was converted into an average percentage by postcode sector and then joined onto the whole PPM and cancer cohorts. The longitude and latitude of each postcode sector was obtained from ONS and used to produce mappings of the proportion of patients from a given region that had their care delivered at Leeds Teaching Hospitals.

3.1.5 - Accuracy of Clinical Coding

3.1.5.1: Study Population and Diabetes Mellitus Definition:

This analysis made use of the cancer cohort and the site specific cohorts described above. Diabetes mellitus (DM) was identified through the programmatic evaluation of clinical coding in PPM or abnormal HbA1c results recorded at LTHT. All clinical coding events within a patient's clinical records were analysed to identify any instance of a diabetic ICD-10 code within their coded events (see appendix for ICD-10 code DM data definitions). The first occurrence of a DM code was taken as being indicative of the diagnosis date for DM. Patients with HbA1c results of greater than or equal to 48 mmol/mol[204] were classified as being diabetic, with the first abnormal HbA1c results taken as the date of diagnosis. Subsequently there groups were defined for those identified as having a diagnosis of DM: i. Those identified with abnormal HbA1c only, ii. those identified through clinical coding only, and iii. those with a hybrid of either abnormal HbA1c or clinical coding (**Figure 2a**). In the case of the hybrid approach, if a patient had both abnormal bloods and clinical coding, then the earlier of these two events was treated as the DM diagnosis event.

To assess the impact of patients living on the edge of the hospital catchment area a sub-population of patients was created to include only those patients living within the geography for which LTHT's blood laboratory process General Practice/primary care blood samples, which has been termed the "LTHT blood catchment area". This area was defined by identifying primary care practices that had provided 10,000 or more blood samples to the LTHT lab and including all patients registered at these practices. The date of DM diagnosis as per the three definitions was compared to the date of cancer diagnosis. Those patients with a diabetic diagnosis date on or before the date of cancer diagnosis were treated as pre-existing diabetics. As comparisons were made between clinical coding and HbA1c, patients were limited to those with a cancer diagnosis from 2005 onwards as blood data was only available from 2004.

Subgroup Analysis

In order to assess aspects of diabetes identification rates and timings of different DM definitions the analysis was conducted on subsets of the hybrid definition DM population. Patients were divided into their route to inclusion in the hybrid diabetic cohort (**Figure 2b**), namely, uniquely identified by HbA1c, uniquely identified by clinical coding or universally identified (Identified by both coding and HbA1c).

Temporal Analysis:

A comparison between the date clinical coding and HbA1c blood results first suggest a DM diagnosis was conducted by analysing those patients in the universally identified group of the hybrid DM cohort. The earliest indicator of DM from bloods and coding were compared and the time difference calculated. The percentage of patients with evidence of diabetes prior to cancer diagnosis that was not identified by clinical coding was calculated for the all cancer cohort.

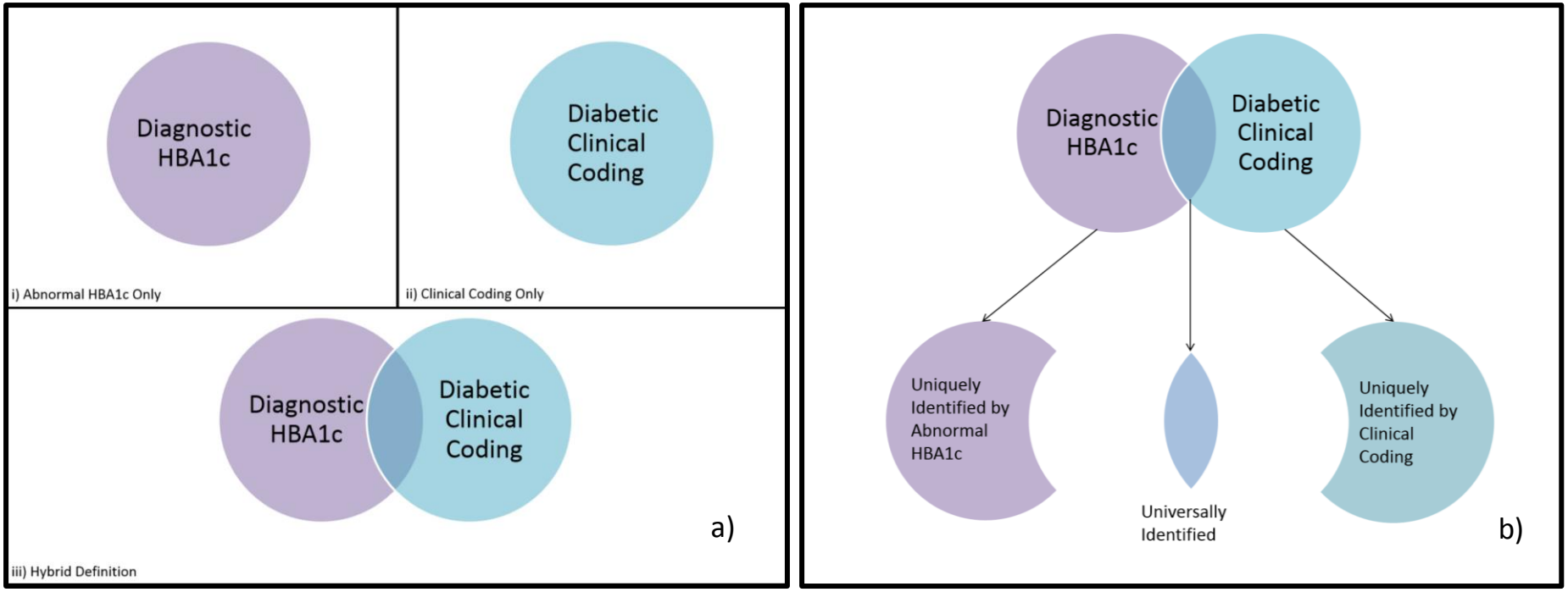


Figure 2: Graphical Representation of Diabetes Mellitus Data Definitions and Subgroups *–a) Graphical representation of the three data definitions of Diabetes Mellitus used to assess the accuracy of clinical coding data. b) Graphical representation of the subgroups of the hybrid definition cohort used in the assessment of patient numbers and timing of diabetes mellitus diagnosis. Note that the circle area is not scaled to the patient numbers in the dataset*

Survival Analysis:

Survival analysis was conducted using Cox proportional hazard adjusting for age, gender (where relevant) and Index of Multiple Deprivation (IMD) quintile.[149] Four Cox proportional hazards[108] models were built using a dataset of all patients and datasets of patients defined by each of the three definitions of DM. Resultant survival trajectories were compared across the dataset visually with survival curves and comparisons of median survival and hazard ratios for clinical significance. Statistical significance was not tested due to DM populations having incomplete pairing with some cases appearing in multiple diabetes definition groups and some in only one. Hazard ratios were extracted from the resultant models in order to identify differences in the hazard estimates produced by each of the diabetic definitions.

3.1.6 - Prevalence of Comorbidity

ICD-10 code definitions for the comorbidities of interest (see chapter 2 and appendix for more details) were used to identify patients with a record of these diagnoses. All clinical coding data for each patient was extracted with partial string matching used to identify each event with a code of the comorbidity of interest. The earliest coded event was used as the date of diagnosis for that comorbidity in that patient. Diabetes Mellitus data was enhanced using HbA1c data, where available, such that a result of greater than or equal to 48mmol/mol was treated as diagnostic for diabetes.[204] Obesity data was enhanced using height and weight data. Implausible or extreme BMI results of below 10 and above 100 were excluded along with results where height was implausible or extreme namely above 3 meters and below 1 meter. Patients with their comorbidity diagnosed on or before the date of cancer diagnosis were treated as having the condition as a pre-existing condition. Those where the diagnosis was recorded after their cancer diagnosis had the condition treated as a late effect.

Data for common comorbidities from the NHS Digital Quality Outcomes Frameworks data[156] was used to determine their prevalence in England and the North East and Yorkshire. These were used as comparators to the all cancer cohort at diagnosis and the all cancer cohort at any time. Additionally the number of comorbidities per patient was calculated with histograms generated to visually demonstrate the distributions.

3.1.6 - Age and Comorbidity:

The relationship between age and comorbidity was assessed by calculating the median age of cancer site specific cohorts and comparing this to the percentage of the cohort with one or more comorbidities. Formal assessment of this relationship was conducted using a zero inflation Poisson model.[171] The coefficients were then extracted and analysed to reveal the relationship between age and being a patient with no comorbidity and the relationship between age and the comorbidity count. In order to assess whether the zero inflation model performed better than a standard Poisson model both were created and then compared using the Vuong test to assess for superiority or non-superiority.[173]

3.1.7 - Timing of comorbidity analysis

Analysis of the timing of comorbidity diagnosis in days before or after cancer diagnosis was calculated for each patient. Initial analysis included all comorbidities together without distinguishing between the specific diagnoses. Data was represented visually using a histogram with 1 month bin width and separately with daily bin width. The analysis focussed on the year of diagnosis ranging from 6 months prior to six months after cancer diagnosis. The analysis was conducted using the all cancer cohort and the site specific cohorts to identify differences in patterns between cancers. A

further analysis was undertaken using the all cancer cohort assessing the pattern of diagnosis in the 15 most common comorbidities by prevalence. Where comorbidities had overlapping definitions i.e. any diabetes mellitus and type one diabetes mellitus, the broadest criterion was included and others excluded.

3.1.8 - Collinearity of Variables

To assess collinearity within the dataset pairwise correlation was calculated across comorbidities, age, gender, and deprivation quintiles. This was conducted using a Spearman's correlation estimate along with a significance test. Results are presented with a significance threshold of p=0.05 and additionally with a Bonferroni correct p value threshold. The latter is included due to the potential for erroneous conclusions to be drawn based on multiple comparisons, this is described in more detail in chapter 2.

3.2 - Results

3.2.1 - Demographics

PPM

The PPM dataset contained 2,764,613 patients when assessed on 05/10/2020. Of these, 459,025 were deceased, leaving 2,305,588 living patients **Figure 3**. This equates to a population size equivalent to 49.1% of the estimated total population of the Yorkshire and Humber Region. It is important to note however that due to migration this is not a true reflection of this proportion of the Yorkshire and Humber population. **Figure 4** shows the population pyramids for the UK, Yorkshire and Humber Region and The PPM cohort. This demonstrates little by way of variation between the national and regional data but marked differences in the PPM cohort. PPM contains a lower proportion of patients under 20, but more patients between 20 and 55 years of age. Similar proportions are seen in patients in the over 65 age group across all cohorts. Within the PPM cohort the reduced number of under 20s translates to a 5 year increase in median age when compared to the national data which are 45 and 40 respectively. The PPM cohort demonstrates a higher proportion of women, with 53.04% being female compared to 50.63% nationally.

The development of the PPM cohort over time can be seen in **Figure 5**. This shows the number of registrations since 2004, which is not consistent over time. 2012 is the largest growth year, with 659,993 new registrations representing 23.87% of all patients registered on PPM. This is over twice as many as the next biggest year of 2015 with 295,465 new patients representing 10.69% of PPM patients.

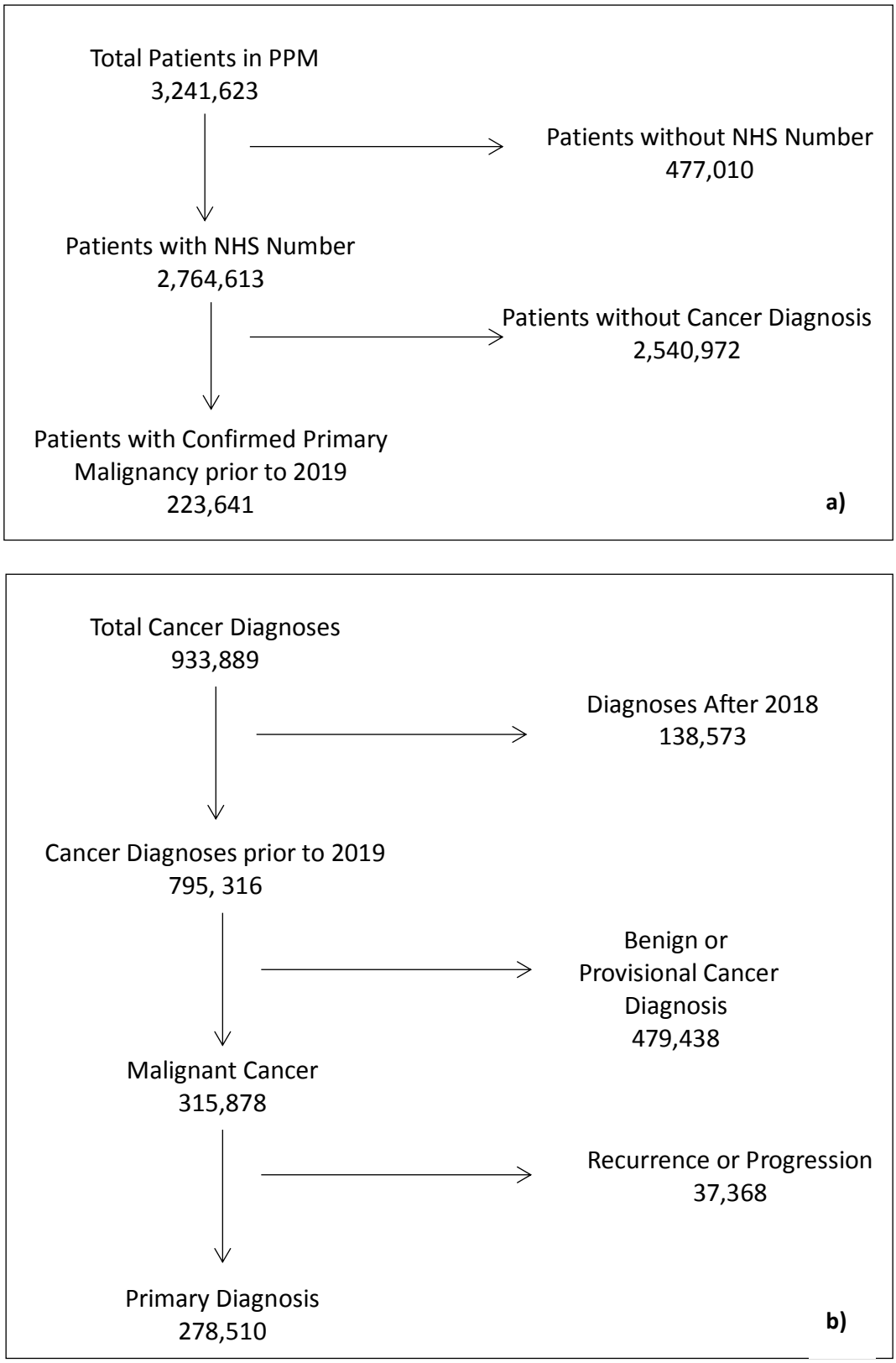


Figure 3: Consort Diagram for PPM Population *- Summary of a) Patient numbers within the PPM cohort and those with a confirmed cancer diagnosis b) Number of cancer diagnoses within the PPM dataset*

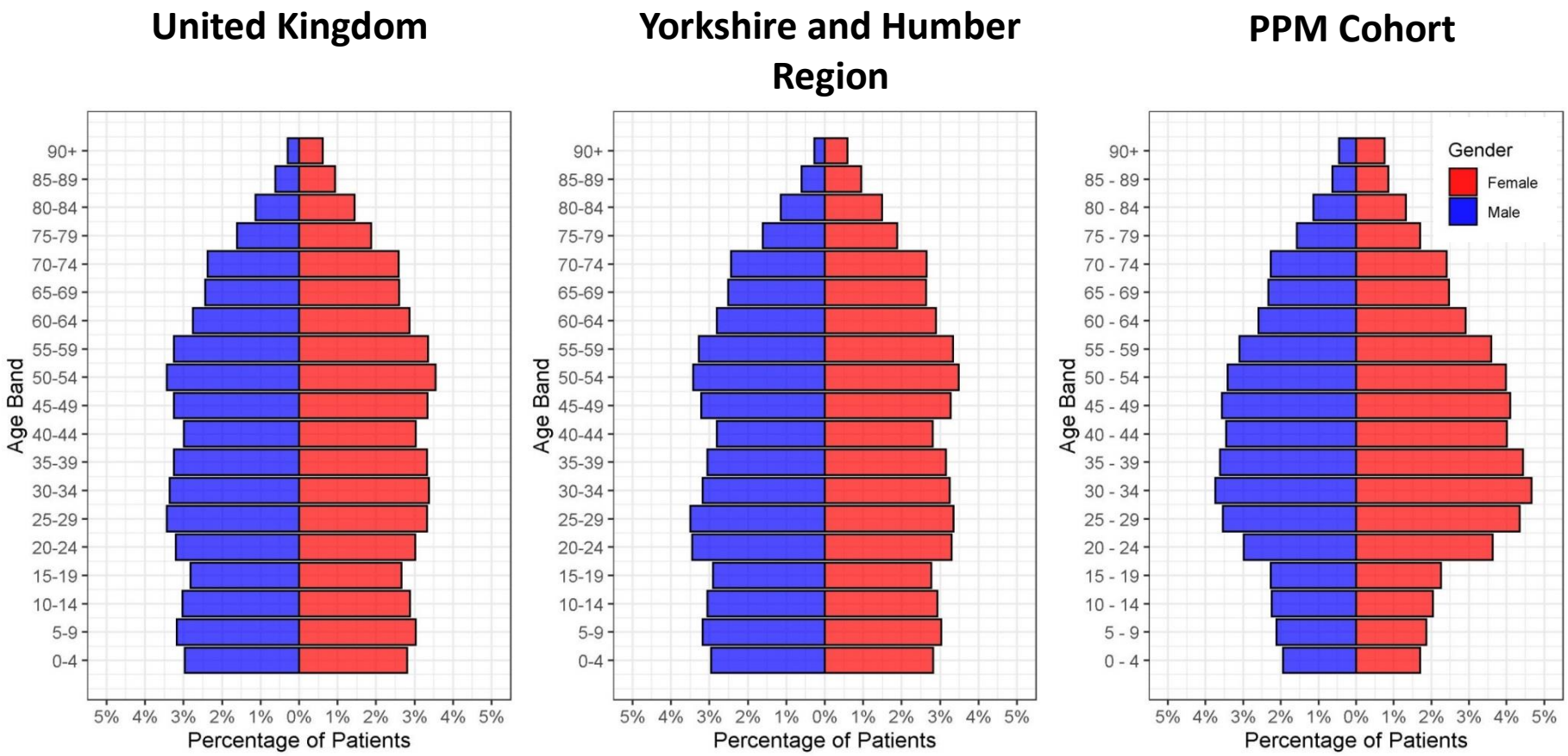


Figure 4: National, Regional and Local Population Pyramid Comparison *- Population pyramids for the UK, Yorkshire and Humber and PPM populations. The UK and Yorkshire and Humber results are derived from ONS data and the PPM results are derived from the local PPM dataset*

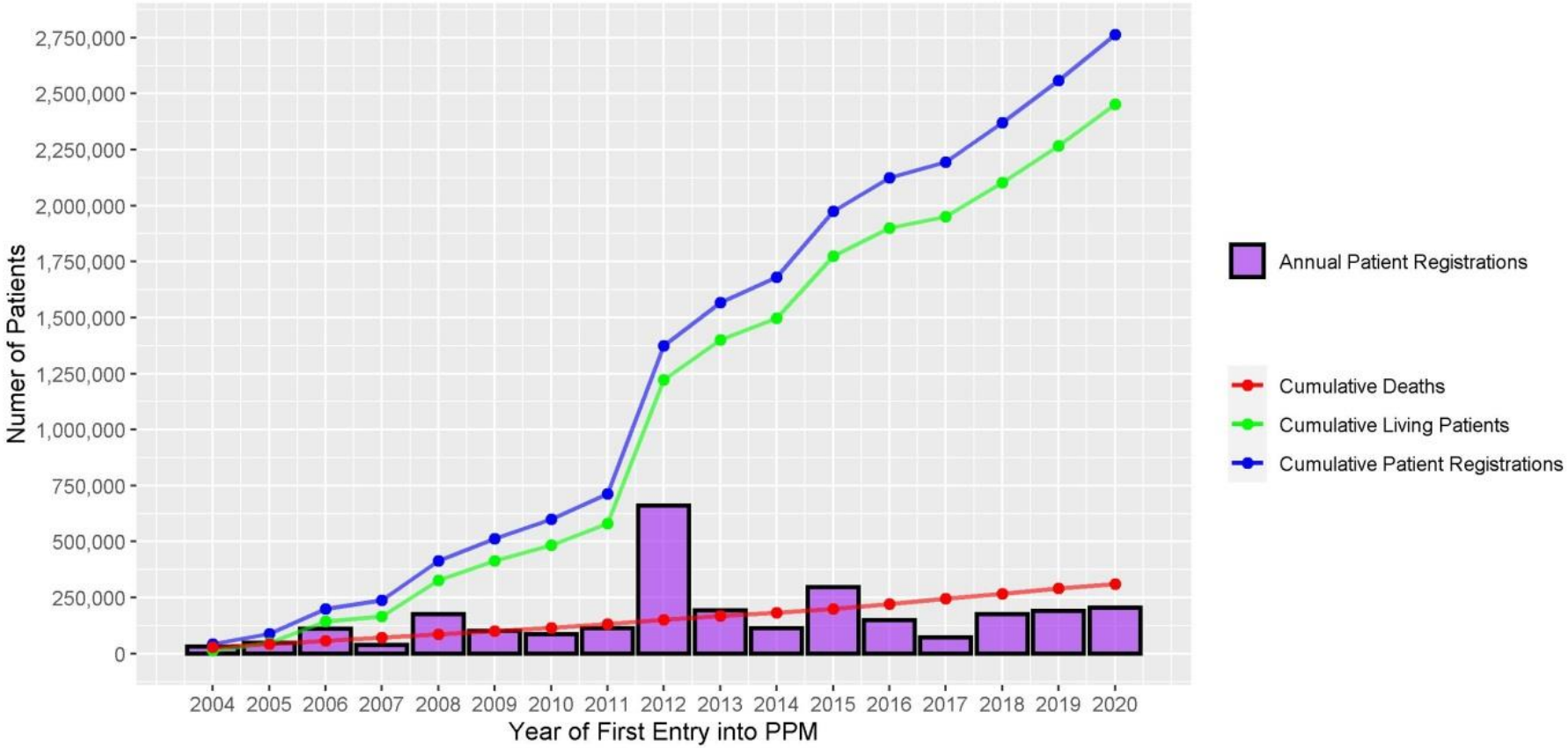


Figure 5: Case Registration Numbers in PPM *– Summary of the number of patient registrations, deaths and living patients within the PPM dataset by year. The purple bars represent annual registrations with the blue line representing cumulative registrations. The red line represents cumulative deaths registered in PPM and the green line the cumulative number of living patients within the PPM dataset. The values provided include only those from 01/01/2004 onwards.*

Cancer Cohorts

Cohort	Median Age	Median IMD	Male n (%)	Female n (%)	No Comorbidity n (%)	Comorbidity n (%)	Median Comorbidities	Total Cohort Size
All Cancer	67	3	112,550 (50.6)	109,866 (49.4)	179,646 (80.77)	42,770 (19.23)	0	222,416
Bladder	74	3	5,293 (70.85)	2,178 (29.15)	5,569 (74.54)	1,902 (25.46)	0	7,471
Brain	58	3	4,106 (60.26)	2,708 (39.74)	5,126 (75.23)	1,688 (24.77)	0	6,814
Breast	60	3	218 (0.69)	3,1484 (99.31)	28,512 (89.94)	3,190 (10.06)	0	31,702
Cervical	45	2	1 (0.03)	3,093 (99.97)	2,897 (93.63)	197 (6.37)	0	3,094
Colorectal	69	3	25,121 (58.76)	17,630 (41.24)	33,582 (78.55)	9,169 (21.45)	0	42,751
Connective	63	3	1,046 (58.27)	749 (41.73)	1,462 (81.45)	333 (18.55)	0	1,795
CUP	71	2	2,345 (48.64)	2,476 (51.36)	3,556 (73.76)	1,265 (26.24)	0	4,821
Endometrial	66	3	0 (0)	5,763 (100)	4,850 (84.16)	913 (15.84)	0	5,763
Kidney	67	3	3,113 (62.27)	1,886 (37.73)	3,583 (71.67)	1,416 (28.33)	0	4,999
Laryngeal	66	2	1,696 (80.88)	401 (19.12)	1,717 (81.88)	380 (18.12)	0	2,097
Leukaemia	63	3	3,130 (58.79)	2,194 (41.21)	4,205 (78.98)	1,119 (21.02)	0	5,324
Liver	67	3	2,509 (68.76)	1,140 (31.24)	2,535 (69.47)	1,114 (30.53)	0	3,649
Lung	71	2	16,588 (55.18)	13,473 (44.82)	21,150 (70.36)	8,911 (29.64)	0	30,061
Lymphoma	63	3	4,419 (54.83)	3,641 (45.17)	6,423 (79.69)	1,637 (20.31)	0	8,060
Melanoma	60	4	3,213 (46.45)	3,704 (53.55)	6,089 (88.03)	828 (11.97)	0	6,917
Myeloma	69	3	1,555 (57.25)	1,161 (42.75)	2,002 (73.71)	714 (26.29)	0	2,716
Oesophageal	70	3	4,277 (68.15)	1,999 (31.85)	4,946 (78.81)	1,330 (21.19)	0	6,276
Ovarian	63	3	0 (0)	3,751 (100)	3,196 (85.2)	555 (14.8)	0	3,751
Pancreatic	70	3	2,260 (51.04)	2,168 (48.96)	3,039 (68.63)	1,389 (31.37)	0	4,428
Prostate	69	3	24,538 (99.97)	7 (0.03)	20,194 (82.27)	4,351 (17.73)	0	24,545
Skin	73	3	13,475 (53.1)	11,900 (46.9)	18,473 (72.8)	6,902 (27.2)	0	25,375
Stomach	72	3	2,993 (63.57)	1,715 (36.43)	3,559 (75.59)	1,149 (24.41)	0	4,708
Testicular	34	3	2,407 (99.96)	1 (0.04)	2,322 (96.43)	86 (3.57)	0	2,408
Thyroid	49	3	670 (26.83)	1,827 (73.17)	2,097 (83.98)	400 (16.02)	0	2,497

Table 4: All Cancer Cohort and Site Specific Cohorts Demographics Summary – *Summary of patient numbers and baseline characteristics at the point of cancer diagnosis for each of the research cohorts including each of the PPM site specific cohorts and the PPM All Cancer Cohort*

The all cancer cohort contains a total of 222,416 unique patients with a total of 223,641 diagnoses across the site specific cohorts. As shown in **Figure 6** the population make-up of the all cancer cohort is different from that of both the PPM population and the national population. It shows much higher proportions of patients over the age of 50 and far fewer patients under the age of 20. This is demonstrated by the median age in this cohort which is 67 compared to 45 in the PPM cohort. The cancer cohort also demonstrates a high burden of ill health with 19.23% of patients having evidence of one or more significant health issues at the point of cancer diagnosis. Across the site specific cohort there is heterogeneity of baseline characteristics with median age ranging from 34 in testicular cancer to 74 in bladder cancer. Similarly, variation is seen in gender balance and deprivation levels with different cancers. Cervical, CUP and lung cancer are seen to have higher levels of deprivation, where melanoma has lower levels of deprivation. **Table 4** also highlights potential evidence of inaccurate data, with patients of the wrong gender having a gender specific cancer such as prostate, testicular and cervical cancer.

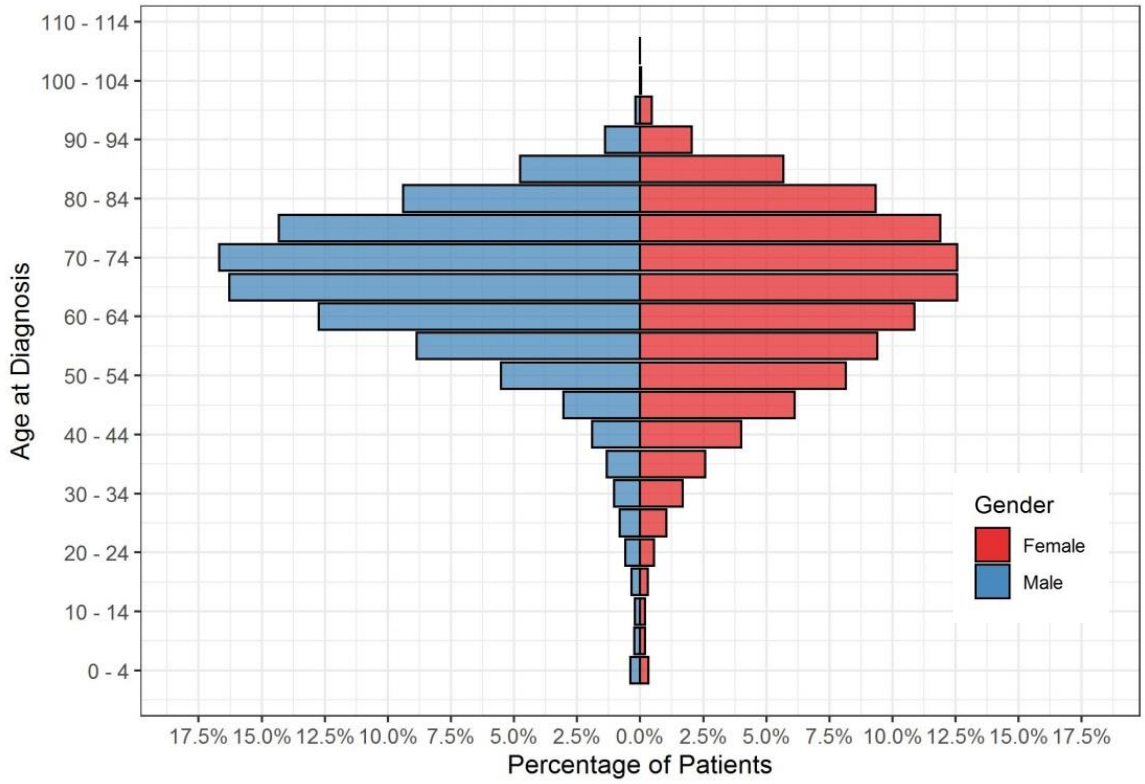


Figure 6: Age and Gender Distribution of the PPM All Cancer Cohort – *Population pyramid for the PPM All Cancer Cohort patients at the point of cancer diagnosis.*

3.2.2 - Missingness

Site	Gender	Age	Grade	IMD	Stage	Histology
Bladder	100	100	79.64	94.66	10.44	100
Brain	100	100	3.08	94.86	18.52	100
Breast	99.99	100	83.4	94.28	46.75	100
Cervical	100	100	71.59	94.21	88.62	100
Colorectal	99.97	100	61.22	95.11	49.33	100
Connective	100	100	50.75	95.32	12.2	100
CUP	100	100	24.02	94.4	1.78	100
Endometrial	99.98	100	89.4	93.72	81.12	100
Kidney	100	100	36.81	94.8	8.96	100
Laryngeal	100	100	68.24	95.57	11.4	100
Leukaemia	100	100	28.93	95.17	0.56	100
Liver	100	100	12.33	95.45	6.39	100
Lung	99.99	100	36.47	95.28	27.15	100
Lymphoma	100	100	28.06	94.69	23.5	100
Melanoma	99.99	100	14.25	93	75.21	100
Myeloma	100	100	29.46	94.88	10.79	100
Oesophageal	99.98	100	58.12	93.74	18.38	100
Ovarian	100	100	70.22	93.55	81.07	100
Pancreatic	99.95	100	18.69	94.22	10.99	100
Prostate	100	100	60.21	94.24	8.64	100
Skin	100	100	33.6	96.32	4.42	100
Stomach	99.98	100	60.99	94.12	10.49	100
Testicular	100	100	11.17	92.48	72.92	100
Thyroid	100	100	31.56	95.07	12.05	100

Table 5: Percentage of Complete Data for Site Specific Cohorts *–Summary of the percentage of data that is complete for gender, age, grade, deprivation quintile (IMD), stage and histological subtype for each of the PPM site specific cohorts.*

Table 5 quantifies the proportion of cases in each site specific cohort with complete data for age, gender, deprivation, histological subtype, cancer stage and grade. Across all cancer sites age, gender and histological subtype are available for all or nearly all cases. Deprivation data is available for the majority of patients with at worst 7.52% missing in the testicular cancer cohort. Cancer grade and staging data has a lower level of completeness with wide variation between cancer sites. Endometrial cancer has on average the most complete data with 10.6% missing grade data and 18.88% missing cancer staging data. By contrast primary brain tumours have the greatest levels of missing data with 96.92% missing a grade and 81.48% missing staging data.

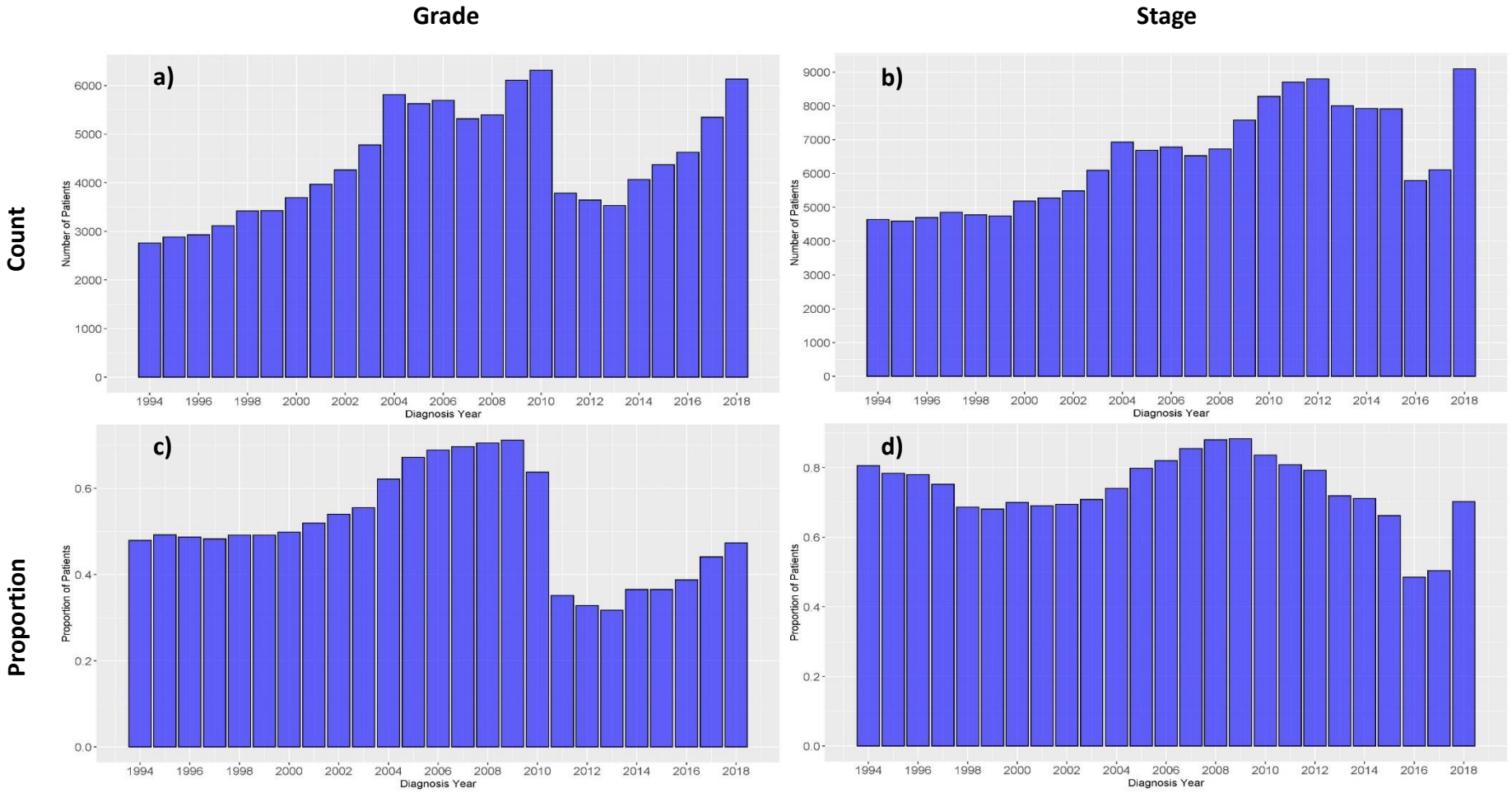


Figure 7: Annual Missing Cancer Stage and Grade data in the All Cancer Cohort of the PPM Dataset– *a) The annual number of patients with missing cancer grade data b) The annual number of patients with missing cancer staging data c) The annual proportion of patients with missing cancer grade data d) The annual proportion of patients with missing cancer staging data*

Figure 7a and **Figure 7b** show the number of patients with missing grade and stage data annually across the cancer cohort. **Figure 7*c*** and **Figure 7*d*** represents these same data as a proportion of cases. This demonstrates annual variation in the percentage with missing staging data and grade data. 2009 has the lowest levels of completeness for both data items, where 2013 has the most complete grading data and 2016 has the most complete staging data. No clear pattern appears to be identifiable in relation to completeness over time.

Beyond cancer specific data, further data items are also missing in a high proportion of cases. One key example is that of ethnicity which is recorded and unknown, unanswered or missing in 65.7% of the total PPM population.

3.2.3 - Patient Geography

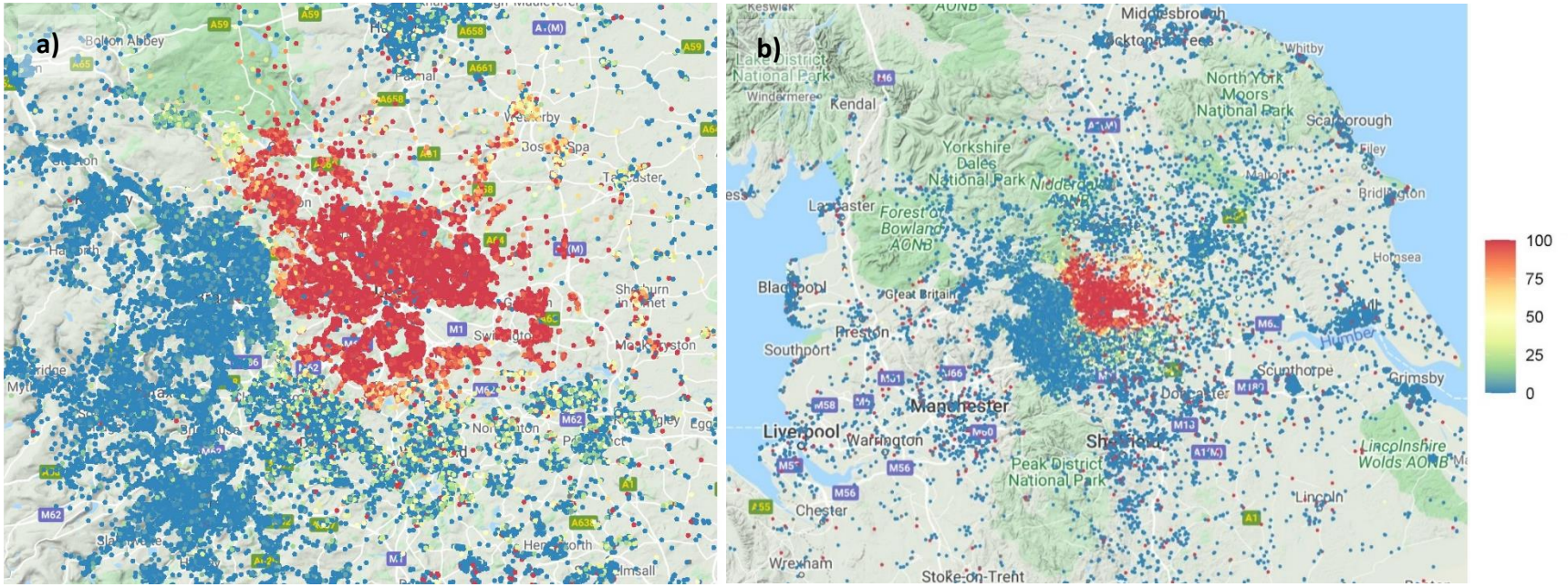


Figure 8: Boundary Effects *- Percentage of the All Cancer Cohort from each postcode whose original cancer diagnosis was Leeds Teaching Hospitals in a) Leeds City Region, b) North of England.*

The geographical mapping of the oncology population in Figure 8 demonstrates that patients in the Leeds city centre are almost exclusively managed within the Leeds Teaching Hospitals network of hospitals. The further the distance someone lives away from the hospital, the lower the chances that their care is exclusively within LTHT. Areas in close proximity to other hospitals such as Bradford Royal Infirmary, Airedale Hospital, Pinderfields and Harrogate General appear to have a smaller proportion of patients managed in LTHT than their distance from Leeds might otherwise suggest. There are a number of locations where despite being over 25 miles from Leeds a large proportion of cases are managed at LTHT, this includes some of the areas around Manchester.

3.2.4 - Accuracy of Clinical Coding

A total of 191,420 unique patients were identified in the more limited time period cohort used for these analyses. This included 14,206 (7.4%) with a DM diagnosis at any time and 8,567 (4.5%) with a DM diagnosis at or before cancer diagnosis.

The abnormal HbA1c only and clinical coding only definitions identified different individuals as having DM, although similar overall numbers of cases (**Figure 9**). Consequently, the hybrid definition identifies a greater number of cancer patients diagnosed with DM at any point. When the population analysed is limited those living within the geography served by the LTHT blood laboratory, the DM diagnostic accuracy of the abnormal HbA1c definition improves from 73.3% to 91.9%. Conversely, the accuracy of the clinical coding only definition declines from 81.1% to 69.9%.

Temporal analysis of the time DM was first recorded by clinical coding and abnormal HbA1c identified low levels of agreement with 64.4% of patients having a date difference of greater than one year and 20.4% greater than 5 years. The use of clinical coding alone to identify DM resulted in 17.5% (over 1 in 6) being incorrectly classified as being diagnosed with DM post cancer diagnosis when abnormal HbA1c data identified DM being diagnosed pre-cancer diagnosis. A smaller 4.4% of patients are incorrectly classified as being diagnosed with DM post cancer by abnormal HbA1c data where clinical coding identified DM being diagnosed pre-cancer diagnosis.

Cox models generated for each of the DM defined cohorts were used to generate survival estimates and then compared these to the total population including both diabetic and non-diabetic patients (**Figure 11**). This identifies differences in the estimated survival trajectories for patients with DM dependent on the definition utilised. Abnormal HbA1c produced the most optimistic DM median survival of 3.9 years, a 36.5% reduction in median survival time compared to the 6.1 year overall median survival for all cancer patients. Clinical coding was the most pessimistic with an estimated median survival of 2.6 years (57.2% reduction from baseline median survival) and the hybrid definition was in between at 3.4 (43.9% reduction in median survival time). The differences in survival curves can be represented by the hazard ratio attributed to DM by each definition (**Figure 12**). *Clinical coding* estimates the hazard to be 3.9 times greater than abnormal HbA1c in the all cancer cohort with an excess hazard of 32.6% for the former and 8.3% for the latter.

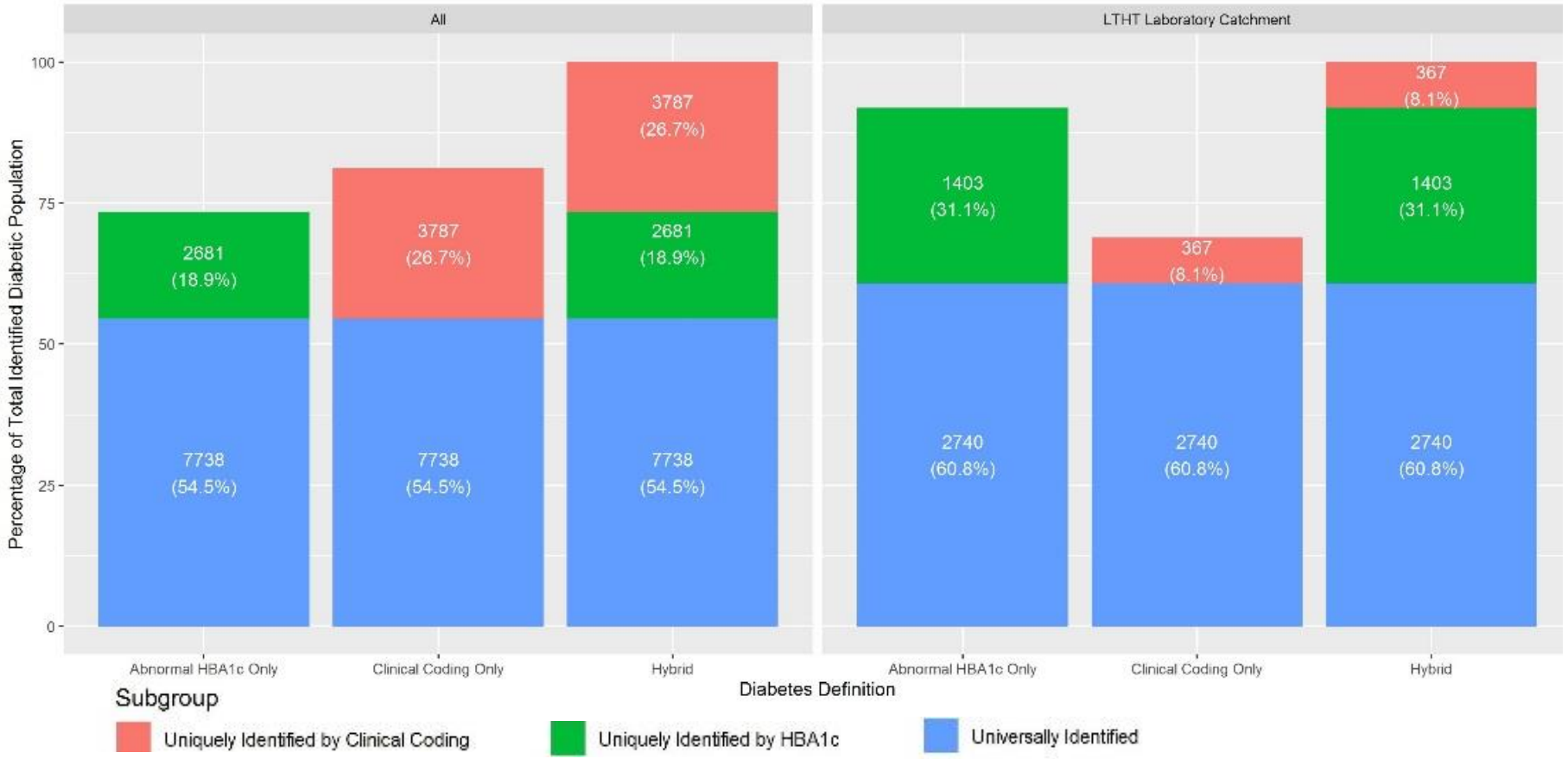


Figure 9: Fidelity of DM Identification *- Stacked bar chart demonstrating the number of diabetes mellitus patients and proportion of total diabetes mellitus population identified by each of the data definitions. Colour is used to identify the subgroups within that data definition. The left hand plot is based on the All Cancer Cohort. The right hand plot is based on the members of the All Cancer Cohort registered to a GP practice that uses Leeds Teaching Hospitals to process its blood samples.*

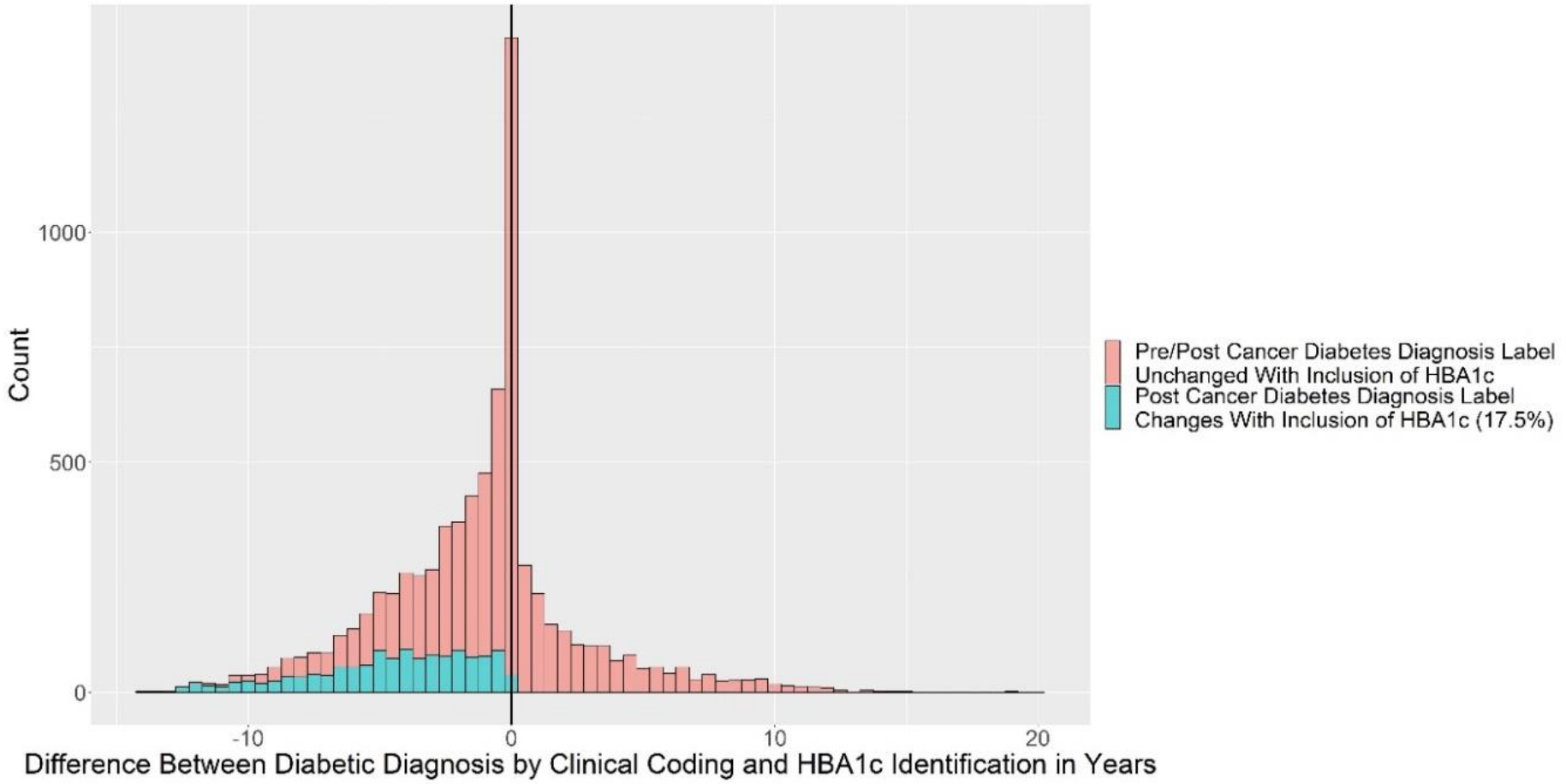


Figure 10: Difference in First Diabetic Diagnosis Indication from Clinical Records Comparing Clinical Coding and Abnormal HbA1c – *Red = patients defined by both an abnormal HbA1c and diabetic clinical coding where the inclusion of HbA1c does not change diagnosis from post cancer to pre cancer. Blue = patients defined by clinical coding as post-cancer diabetics but with abnormal HbA1c pre-cancer. Left of black line = abnormal HbA1c earlier than clinical coding. Right of black line = clinical coding earlier than abnormal HbA1c.*

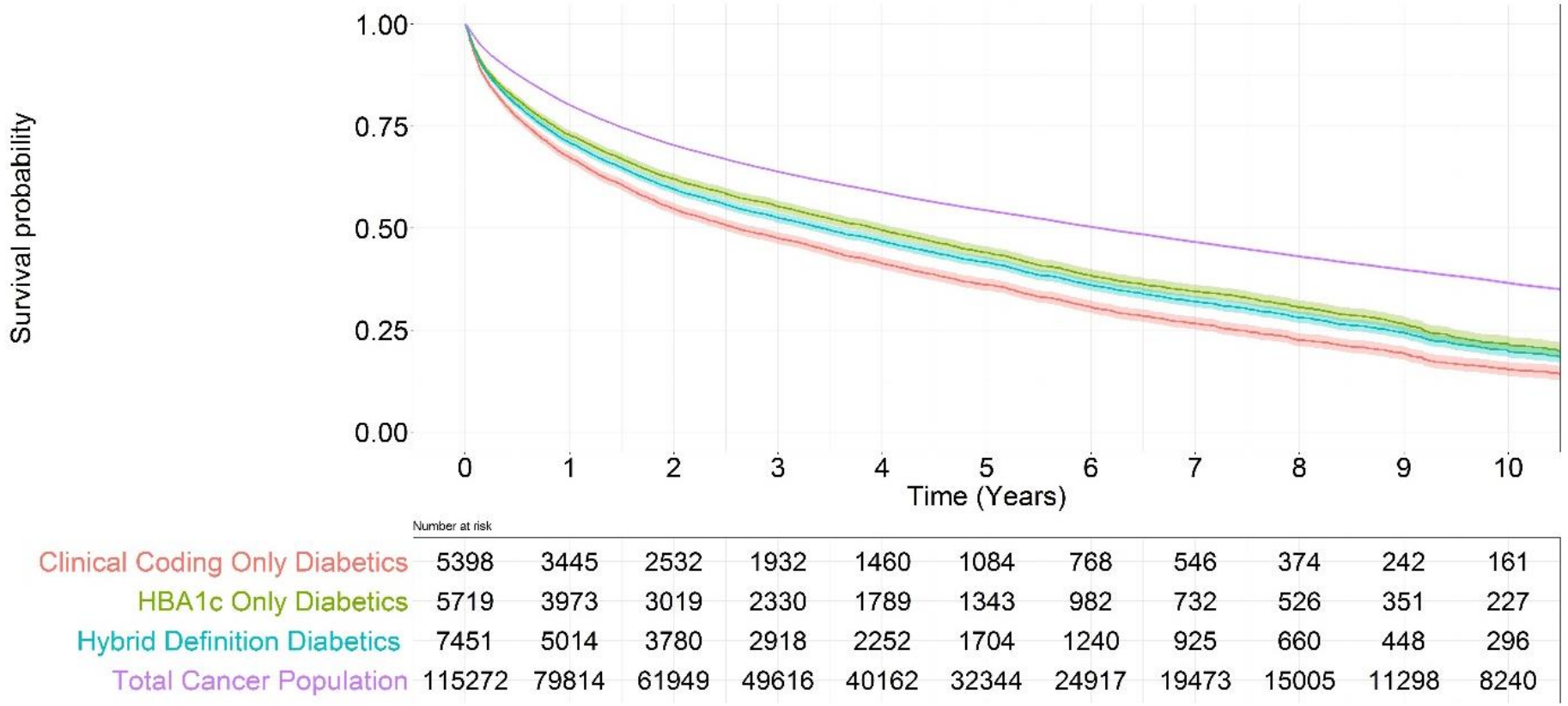


Figure 11: Survival Trajectories for Each Diabetic Data Definition and All Patients in the All Cancer Cohort *– Cox Proportional Hazard Survival estimates plotted for four separate populations i) All patients within the All Cancer Cohort ii) Patients from the All Cancer Cohort identified as diabetic by HbA1c, iii) Patients from the All Cancer Cohort identified as diabetic by Clinical Coding iv) Patients from the All Cancer Cohort identified as diabetic by the Hybrid definition. In all cases estimates were adjusted for age, gender and deprivation. Note that this is four separate survival curves plotted on a single set of axes, not a stratified Cox as each data definition produces a subgroup of the All Cancer Cohort yielding a partially paired and partially unpaired population.*

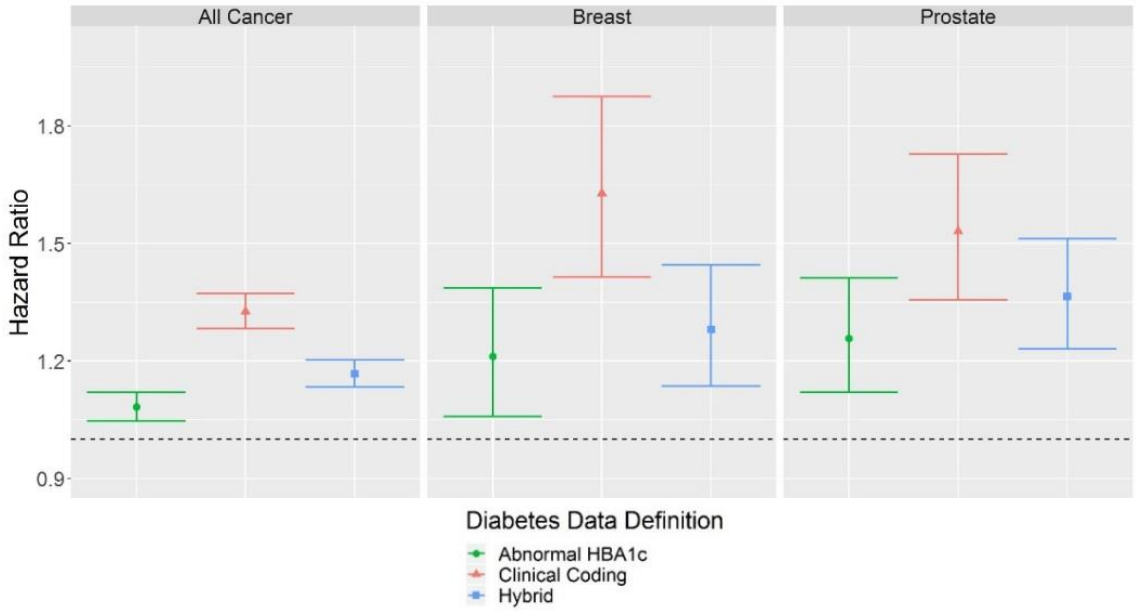


Figure 12: Impact of Data Definitions on the Cox Derived Hazard Ratios for the Impact of Diabetes *- Comparison between the estimated point estimate and confidence intervals for the hazard ratio associated with diabetes mellitus in the All Cancer Cohort and the Prostate and Breast Cancer Site Specific Cohorts for the three data definitions of diabetes mellitus. Dotted line at 1 represents equal hazard in both groups, above the line suggested increased hazard associated with diabetes mellitus and below the line suggests an associated decrease in hazard associated with diabetes mellitus. Adjustment for age, and deprivation was conducted for all analyses and gender in the case of the Breast and All Cancer Cohorts.*

3.2.5 - Prevalence of Comorbidity

Condition	England	North East and Yorkshire	All Cancer Cohort at Diagnosis	All Cancer Cohort at or After Diagnosis
Coronary Heart Disease	3.16	3.85	6.66	11.84
Congestive Cardiac Failure	0.83	0.98	1.54	4.81
Hypertension	13.94	14.87	9.41	20.65
Peripheral Arterial Disease	0.59	0.78	1.01	2.14
Stroke or TIA	1.76	2.07	1.27	1.27
Asthma	5.93	6.37	1.86	4.30
COPD	1.90	2.50	2.73	6.36
Obesity	9.75	11.92	1.00	6.81
Chronic Kidney Disease	4.10	4.48	1.12	4.16
Diabetes Mellitus	6.79	7.12	4.55	9.45
Dementia	0.76	0.84	0.60	2.78
Rheumatoid Arthritis	0.75	0.82	0.55	1.14

Table 6: Comparison of Comorbidity Prevalence in Local, Regional and National Data *–Data for population prevalence as a percentage of the population. The national (England) and regional (North East and Yorkshire) percentages are derived from to 2017-2018 NHS Digital GP Quality Outcomes Framework publically available data. The remaining two columns relate to the population prevalence within the all cancer cohort. The prevalence is expressed at the time of cancer diagnosis and at any time either at or after cancer diagnosis. The comorbidity groupings for the all cancer cohort were adapted to create comparable definitions and do not map onto the comorbidity definitions used elsewhere in analysis*

Table 6 identifies key differences between the prevalence of comorbidity within the All Cancer Cohort and population estimates regionally and nationally. The North East and Yorkshire have a higher prevalence of all of the comorbidities when compared to the English average, suggesting a greater burden of ill health. In some cases the regional relative excess is clinically significant such as 32% for COPD and peripheral arterial disease, 22% for coronary heart disease and obesity with the smallest difference seen in hypertension and asthma where there is a 7% relative excess within the region.

When comparing the rates of ill health at the point of cancer diagnosis to the regional prevalence the pattern is highly variable with some conditions being more common and other less so. Coronary heart disease has the highest relative excess prevalence at 73% above the regional average. Three other conditions have a higher relative excess prevalence, namely congestive cardiac failure at 57%, peripheral arterial disease at 29% and COPD at 9%. Some conditions have large relative underrepresentation, particularly obesity at -92% and chronic renal disease at -75%.

If the lifetime prevalence of comorbidity for the cancer cohort is examined, this results in sizable, clinically relevant shifts, with on average higher levels of comorbidity with the exception of asthma (-32%), obesity (-43%) and chronic renal dysfunction (-7%). The largest relative excess prevalence is for congestive cardiac failure which is 391% higher than the regional average. Other notable elevated rates are for dementia (231%), coronary heart disease (208%), peripheral arterial disease (174%), and COPD (154%).

As the publically available data for comorbidity is different to the comorbidity groups used within the wider analysis of our cohorts, the prevalence of each comorbidity, as per our study definition, is presented in **Figure 13**. This demonstrates the wide variation in how common comorbidities are. A number of comorbidities demonstrate a clear bias towards first detection after cancer diagnosis in particular obesity, CCF, and dementia.

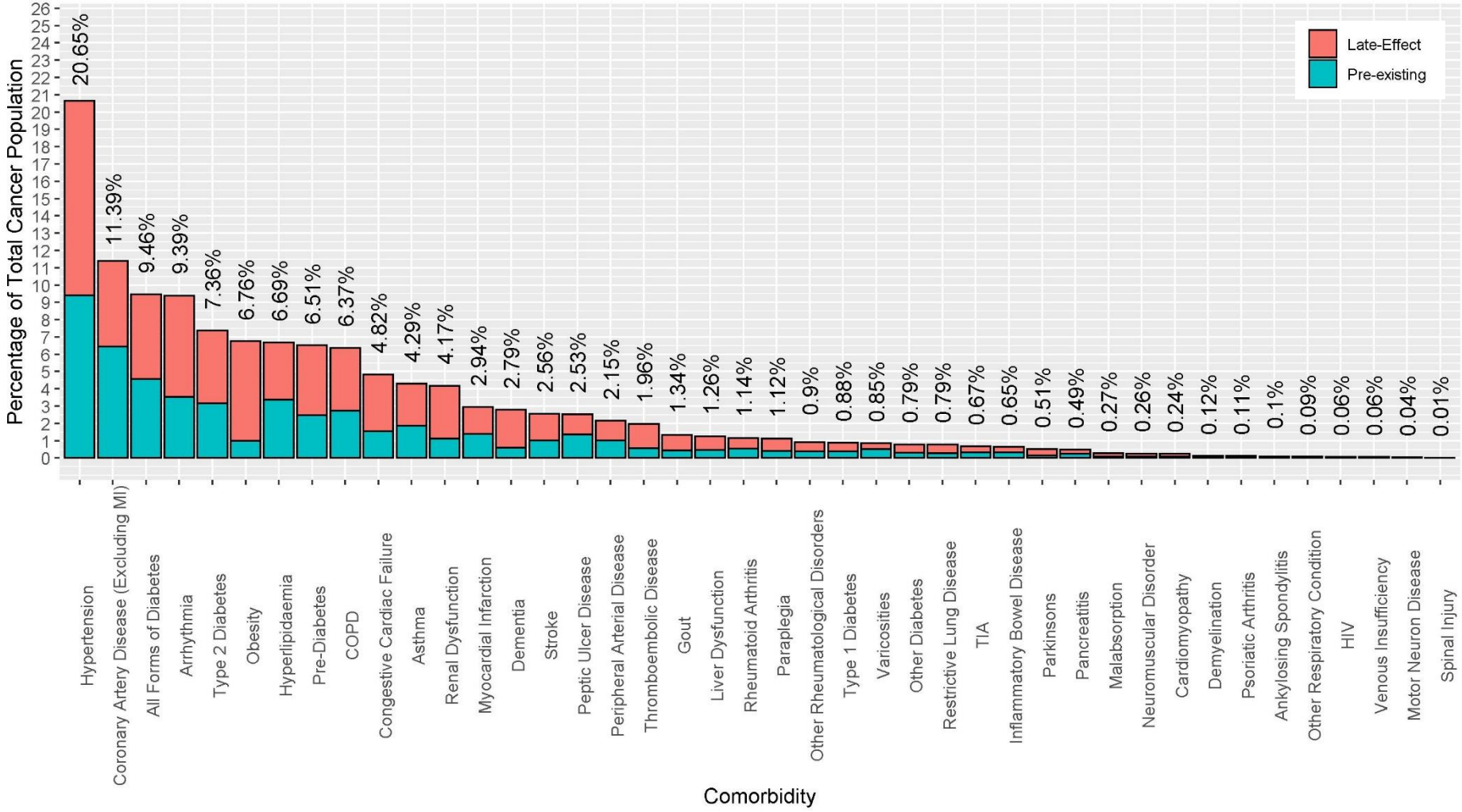


Figure 13: Prevalence of Comorbidity in the All Cancer Cohort - *The percentage of the All Cancer Cohort identified as having each of the comorbidities of interest. Each condition is divided by colour into those where the comorbidity was recorded at or before cancer diagnosis (Pre-existing) and those where the comorbidity was recorded after cancer diagnosis (Late-Effects)*

3.2.6 - Timing

Figure 14 identifies the timing of comorbidity diagnosis in a more granular way than the binary allocation of timing into before and after cancer diagnosis. When assessing the all cancer cohort on a monthly basis a near symmetrical distribution is identified with peak incidence around the time of cancer diagnosis. The peak is in the month leading up to cancer diagnosis representing 6.44% of all comorbidity diagnoses. The month after cancer diagnosis remains high, accounting for 5.45% of comorbidity diagnoses. The six month period around cancer diagnosis, from three months before to three months after, represent 21.25% of all diagnosed comorbidities. The rapid fall off in comorbidity diagnosis either prior to, or after this time period means that extending the time window to the year of cancer diagnosis only increases the proportion of comorbidity identified to 27.18%.

Figure 15 focusses on the year around the cancer diagnoses for the All Cancer Cohort. This demonstrates how increasing the granularity further to the daily proportion of comorbidity identifies not a single peak, but a biphasic pattern of comorbidity diagnosis. This pattern includes an exponential daily increase from a baseline of 0.02% at 3 months before cancer diagnosis to a peak of 2.03% on the day of cancer diagnosis. This then rapidly falls to a new baseline of 0.07% three days after cancer diagnosis, before a more gradual increase to a smaller second peak of 0.16% 56 days after cancer diagnosis. Despite the smaller maximum value of the second peak compared to the first, its more gradual increase and decrease means that the 3 months after cancer diagnosis accounts for 10.63% of comorbidity diagnoses, versus 8.58% for the three months prior. Indicating a larger area under the curve for the second peak compared to the first.

When the timing is assessed by separating out different comorbidities (**Figure 16**) the first peak is apparent in all however the second peak is less pronounced in several cases. Asthma has the highest first peak of 3.55% on the day of cancer diagnosis, closely followed by hypertension with 3.44%, By contrast, obesity shows a much smaller first peak of 1.79% but a much larger second peak 0.76% 34 days after cancer diagnosis. The proportion of comorbidity diagnoses daily conducted in the site specific cohorts (**Figure 17**) shows variation across cancer sites with the largest first peak seen in primary brain tumours and thyroid cancer. The large second peaks are seen in breast, cervical and testicular cancer.

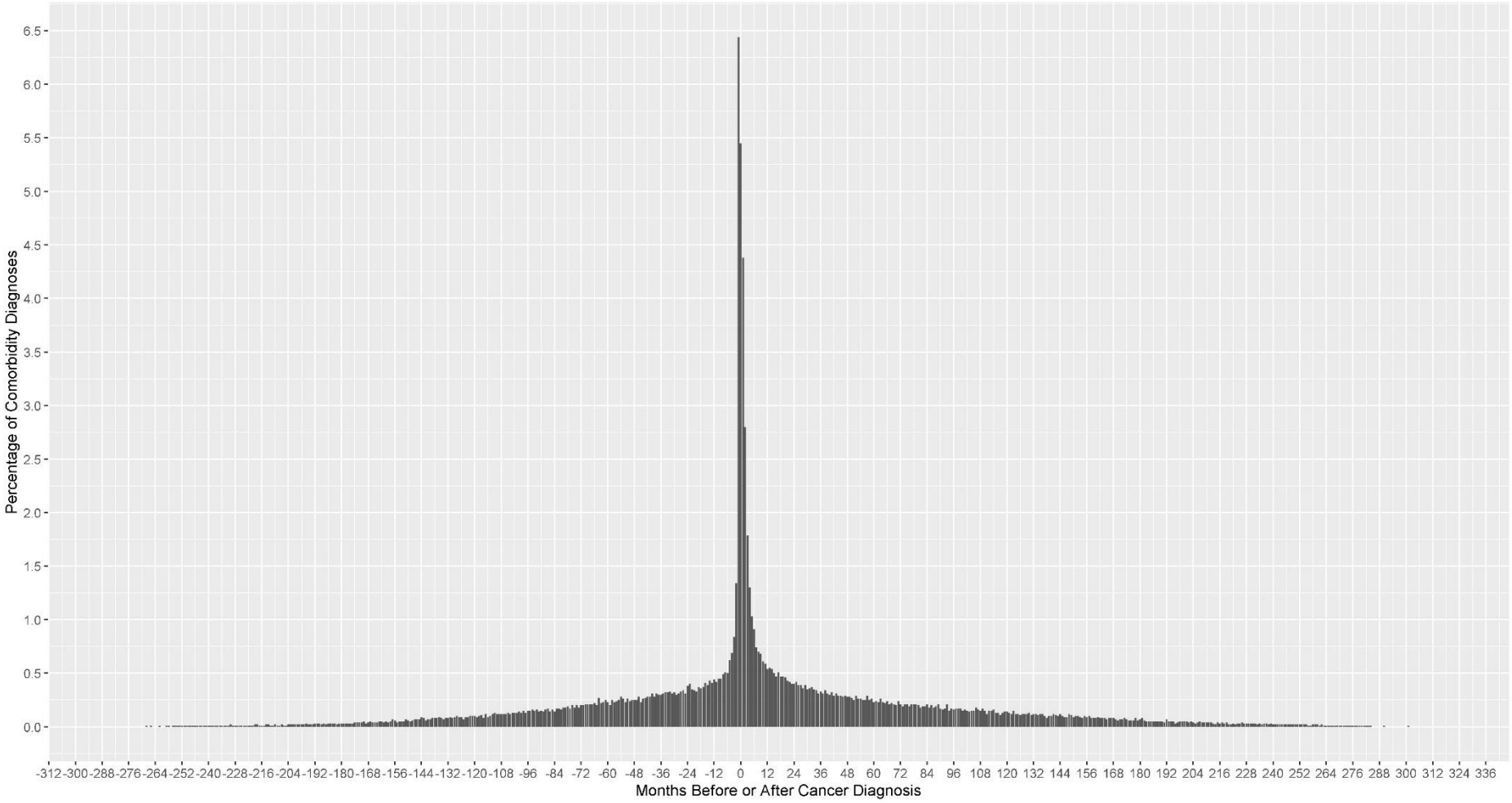


Figure 14: Distribution of Comorbidity Diagnosis Event Relative to Cancer Diagnosis *- Proportion of all comorbidity events in the All Cancer Cohort that were diagnosed each month over time expressed as a percentage. Time zero is the month of cancer diagnosis with negative months being prior to cancer diagnosis and positive months being after cancer diagnosis*

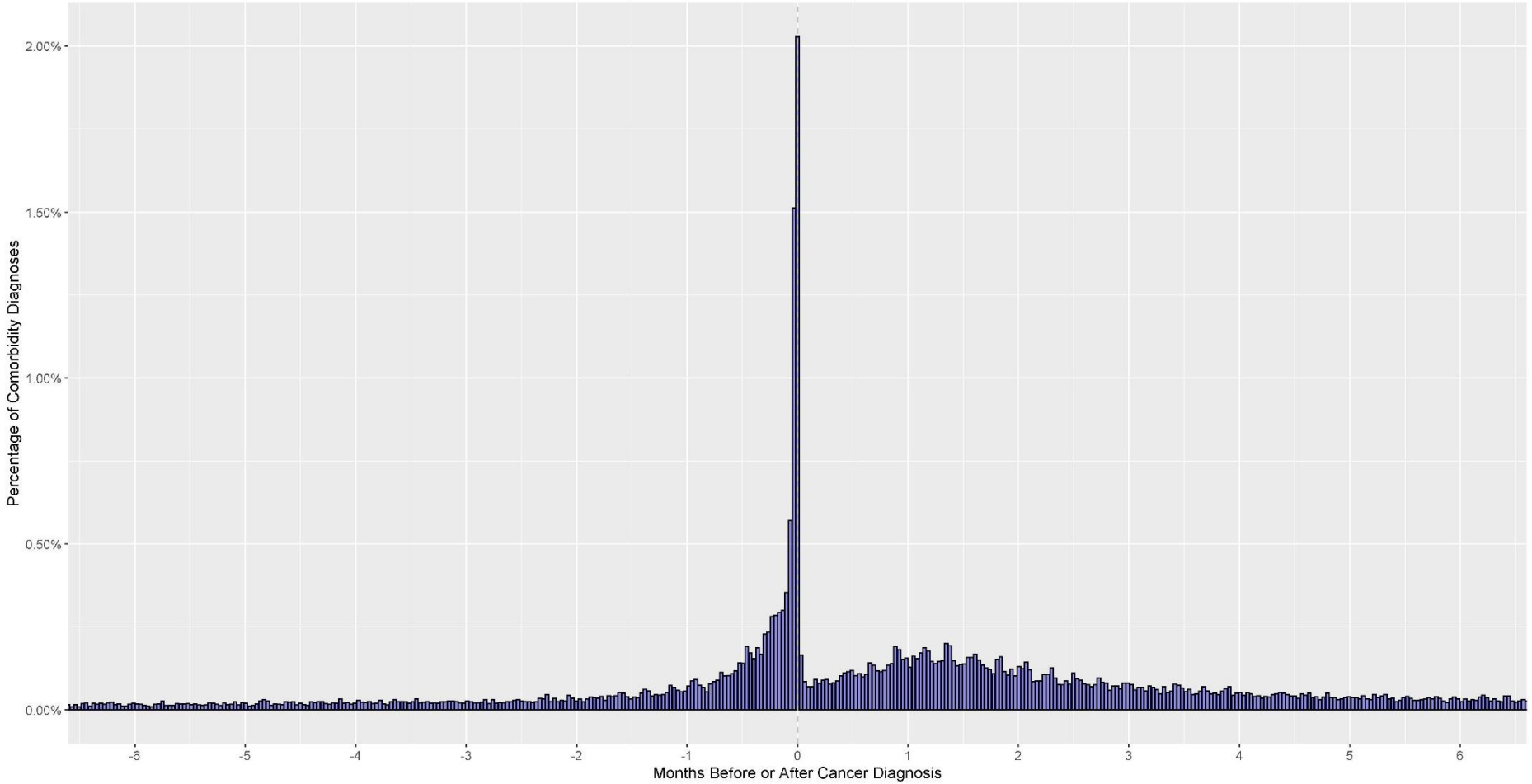


Figure 15: Distribution of Comorbidity Diagnosis Events Relative to Cancer Diagnosis Daily in the Year of Cancer Diagnosis *- Proportion of all comorbidity events in the all cancer cohort that were diagnosed each day over time expressed as a percentage. Time zero is the day of cancer diagnosis with negative days being prior to cancer diagnosis and positive days being after cancer diagnosis. Although only 6 months before cancer diagnosis to six months prior is plotted the proportions are expressed as function of the number of comorbidities at all times.*

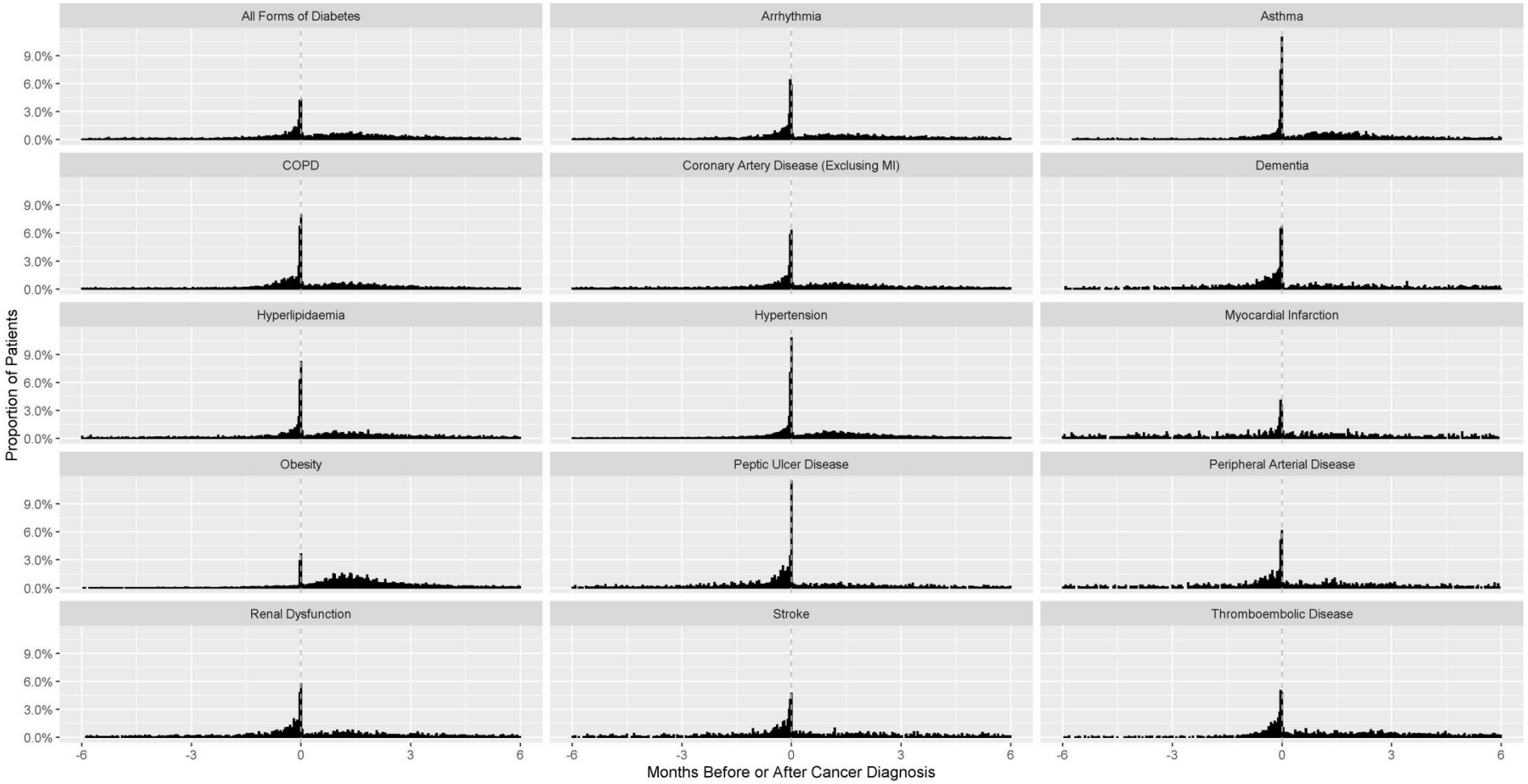


Figure 16: Distribution of Comorbidity Diagnosis Events Relative to Cancer Diagnosis by Comorbidity *- Proportion of each of the 15 most common comorbidities by prevalence diagnosed daily in the all cancer cohort. Time zero is the day of cancer diagnosis with negative days being prior to cancer diagnosis and positive days being after cancer diagnosis. Although only 6 months before cancer diagnosis to six months prior is plotted the proportions are expressed as function of the number of comorbidities at all times*

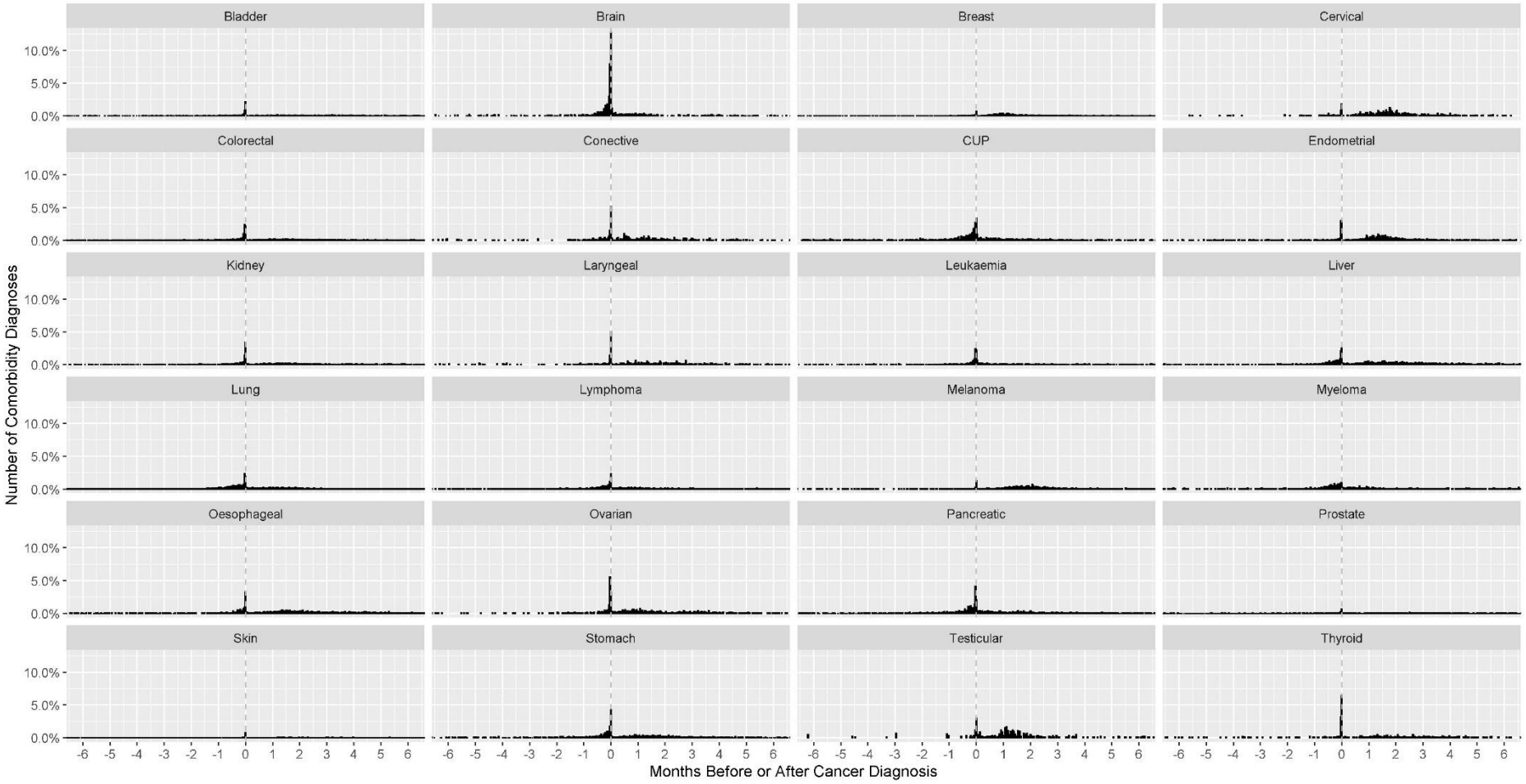


Figure 17: Site Specific Distribution of Comorbidity Diagnosis Events Relative to Cancer Diagnosis Daily in the Year of Cancer Diagnosis *- Proportion of all comorbidity events in the Site Specific Cancer cohorts that were diagnosed each day over time expressed as a percentage. Time zero is the day of cancer diagnosis with negative days being prior to cancer diagnosis and positive days being after cancer diagnosis. Although only 6 months before cancer diagnosis to six months prior is plotted the proportions are expressed as function of the number of comorbidities at all times.*

3.2.7 – Collinearity

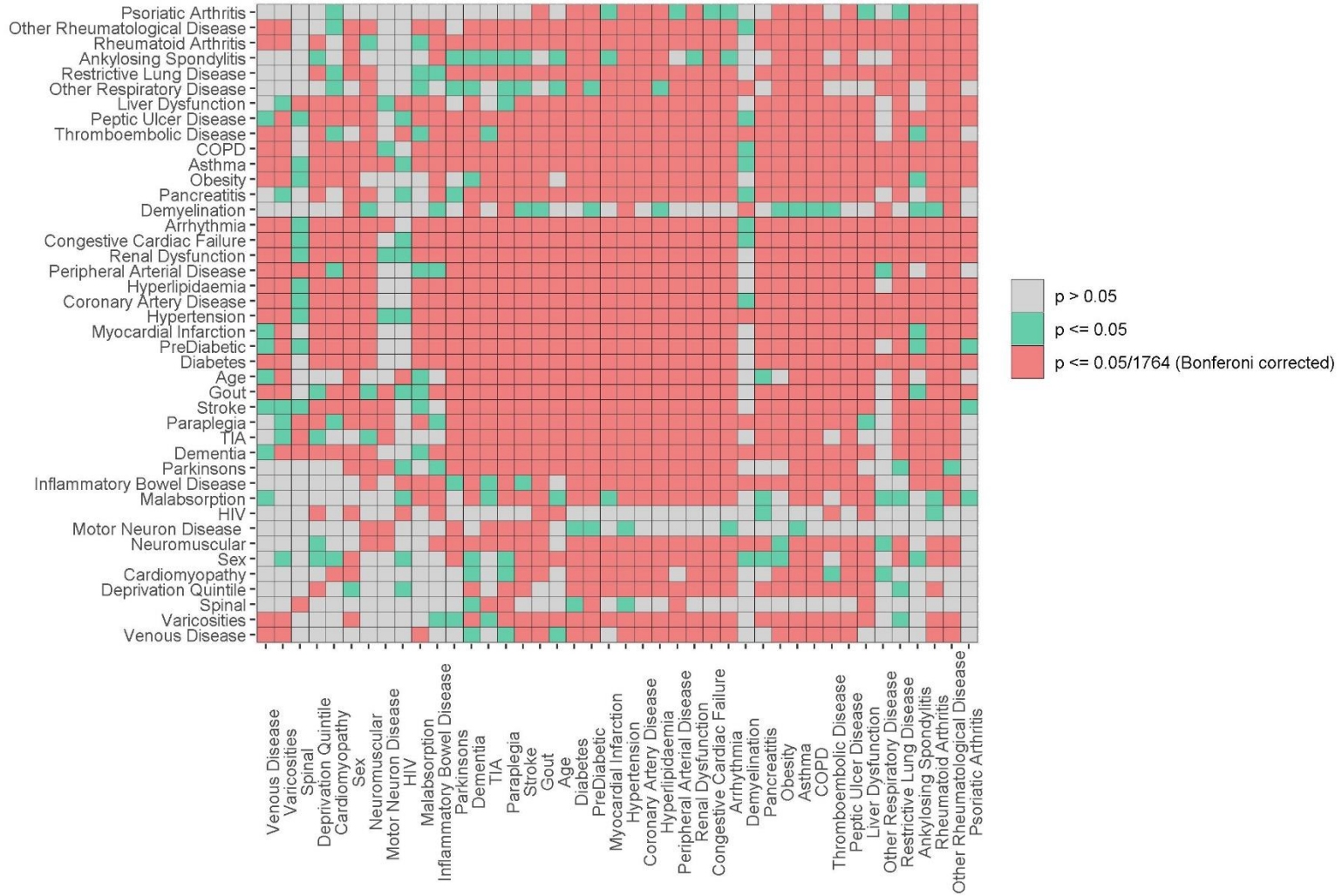


Figure 18: Statistical Significant of Pairwise Spearman's Correlation *– p value results for pairwise comparisons of variables in the All Cancer Cohort. Results are shown as whether they are at or below the Bonferroni corrected value, at or below 0.05 but above the Bonferroni corrected value or above 0.05.*

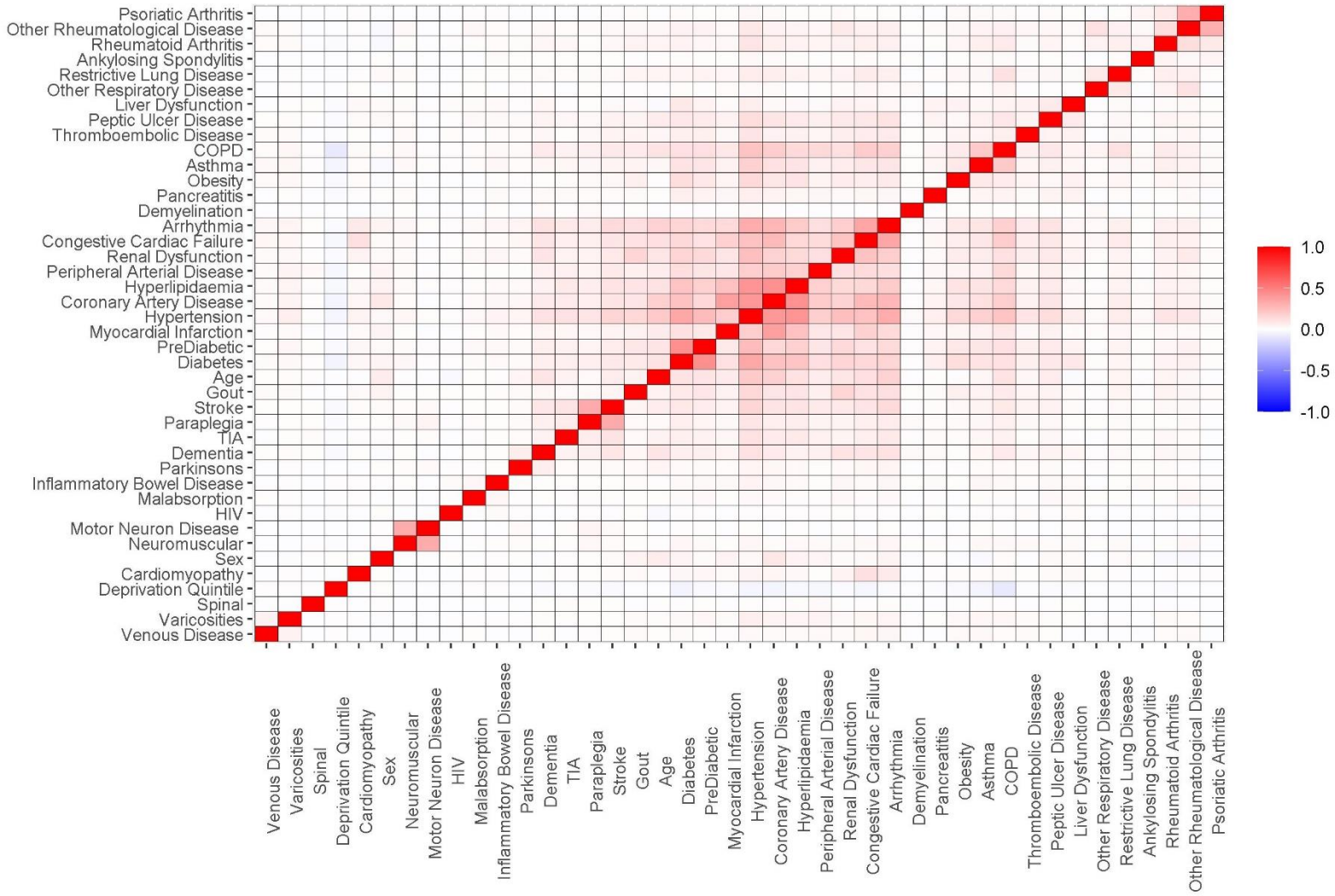


Figure 19: Spearman's Correlation Coefficients for Pairwise Comparisons – *Spearman's correlation coefficient results for each pair of variables in the All Cancer Cohort.*

Figure 19 shows the correlation for of each of the pairwise comparisons. The majority of estimates yielded a zero or low positive correlation. A number of more strongly positive correlations were apparent particularly in relation to cardiac comorbidity where arrhythmias, hypertension, coronary artery disease, heart failure and myocardial infarction were all interrelated. Age was demonstrated to have weak positive correlation with almost all health conditions. **Figure 18** shows the same comparisons in terms of their significance tests, using both a standard and adjusted p value threshold. This highlights that although most of the correlations are only modest in scale, they are highly statistically significant in many instances. The five strongest associations were pre-diabetes and diabetes (0.43), hyperlipidaemia and coronary artery disease (0.43), hyperlipidaemia and hypertension (0.41), hypertension and coronary artery disease (0.40) and myocardial infarction and coronary artery disease (0.37). Scatter plots for each comparison were also checked and identified no examples where they were dichotomous.

2.2.8 - Interaction between Age and Comorbidity

As demonstrated in **Figure 18** and **Figure 19** age is significantly associated with the rate of multiple comorbidities. Rather than looking at individual comorbidity an alternative approach would be to look at a count of comorbidity. **Figure 20** highlights that although having one or more comorbidities at the point of diagnosis is common, it is still a minority group representing 19.23% of the All Cancer Cohort. This explains why consistently the median number of comorbidity across all the site specific cohorts is 0 (**Table 4**). In order to assess the relationship between age and comorbidity a linear regression was fitted in **Figure 21** demonstrating that with increasing median age at diagnosis, there is a trend towards a greater proportion of patients with one or more significant health conditions at the point of cancer diagnosis. Compared to the overall trend, some cancers show higher levels of ill health than the modelling might anticipate. Examples of this include primary brain tumours, pancreatic cancer and liver cancer. In other cases the level of ill health relative to age is lower such as breast cancer, melanoma and ovarian cancer.

The relationship between age and count of comorbidity was assessed using a zero inflated Poisson regression model[171]. A standard Poisson regression model was applied to the data and compared to the zero inflation model using the Vuong test.[173] This demonstrated superiority of the zero inflated model with a significance level of $<2.22 \times 10^{-16}$ for raw values, AIC and BIC. Analysis of the zero inflated model coefficients demonstrated a relative risk of 0.964 (0.963-0.964) for the logistic regression element and a risk ratio of 1.018 (1.017-1.019) for the count model. This can be interpreted as with each one year increase in age, the chances of being a patient with no comorbidity reduces by 3.6% and the count of comorbidity increases by 1.8%. This identifies a clear relationship between the age of diagnosis and the number of comorbidities.

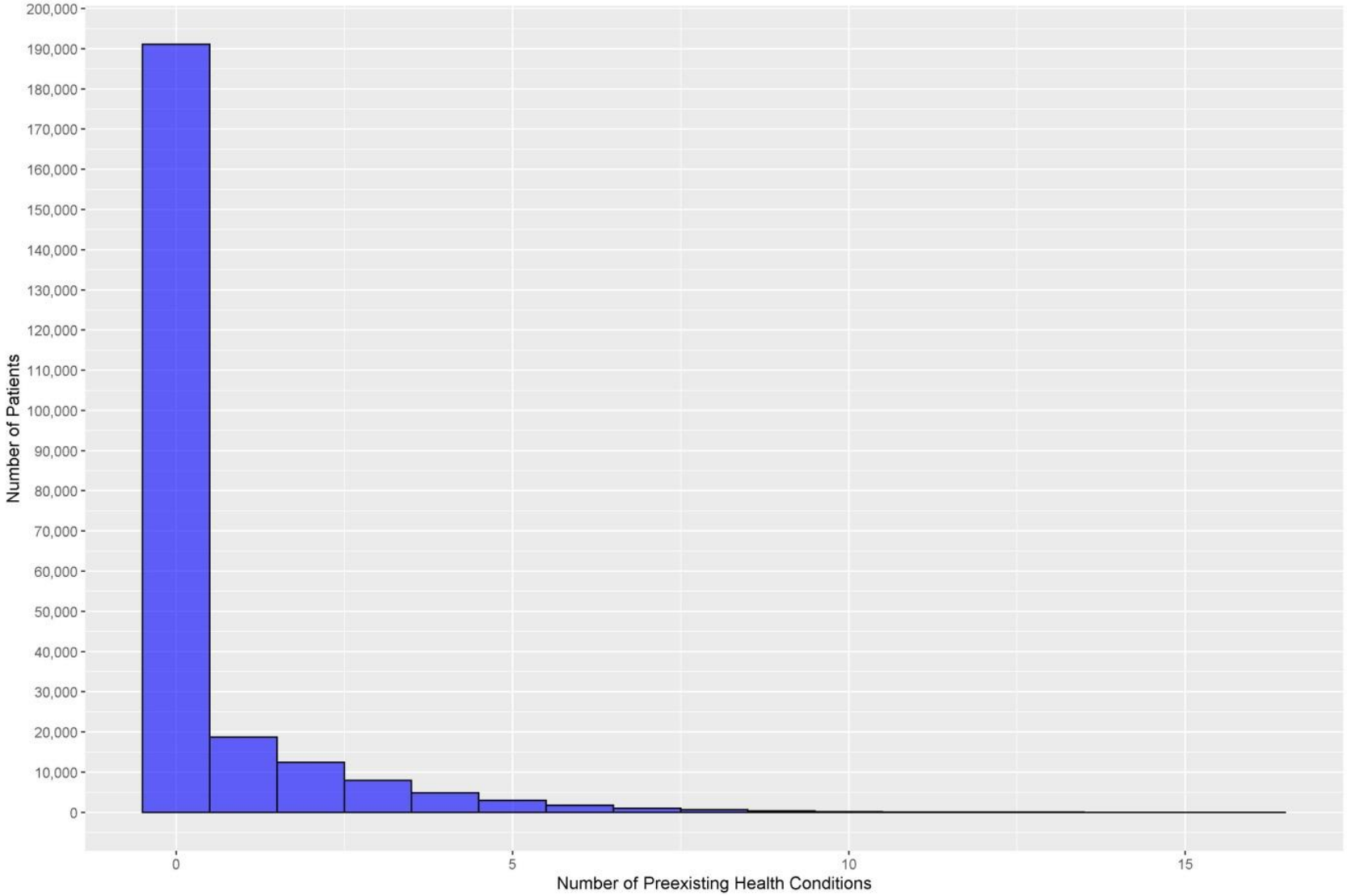


Figure 20: Number of Comorbidities per Patient in the All Cancer Cohort *- Histogram of the count of comorbidity at the point of cancer diagnosis for the All Cancer Cohort. Where conditions of interest overlap e.g. hybrid defined diabetes mellitus and type 2 diabetes mellitus, these were only counted once.*

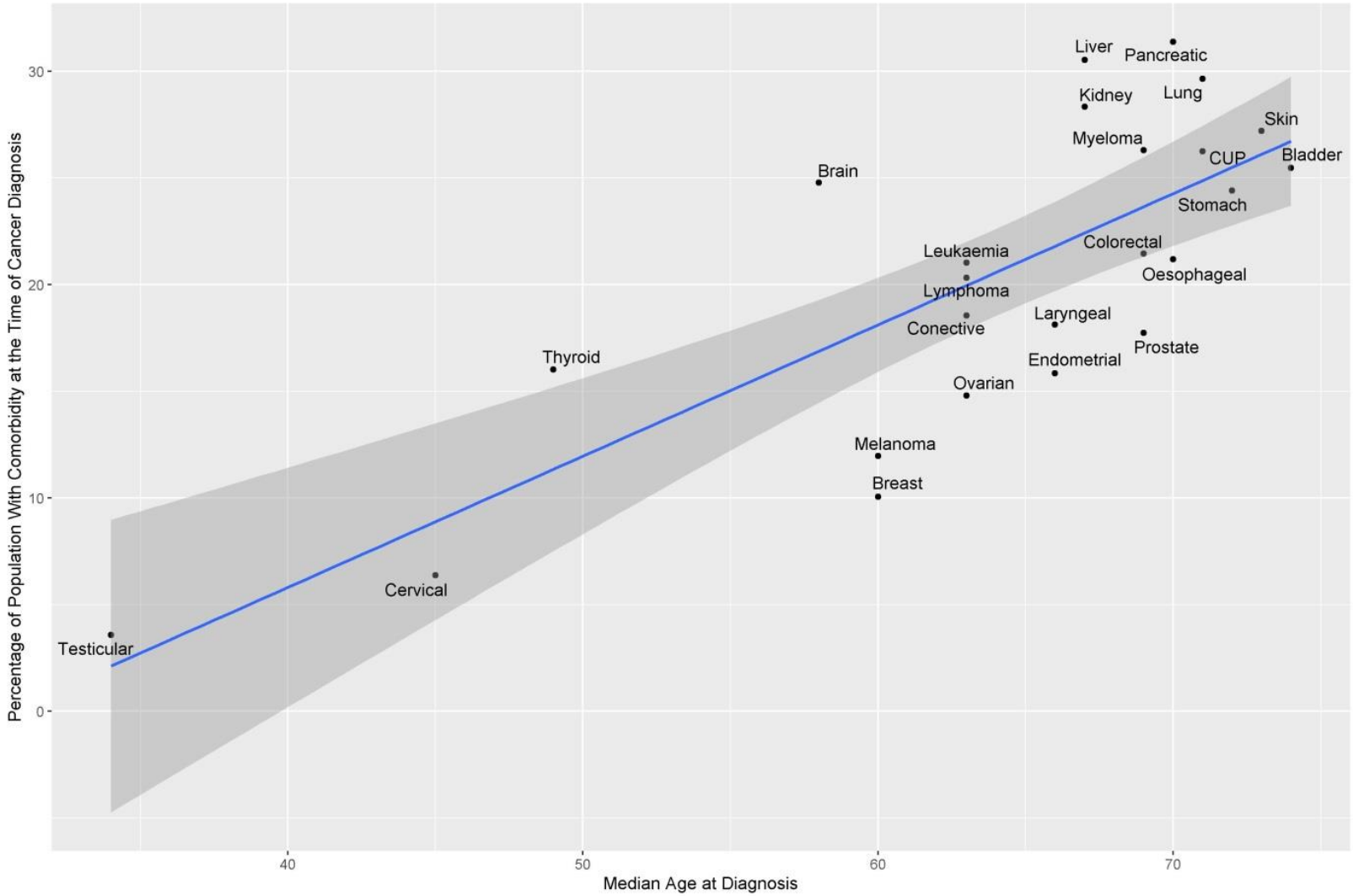


Figure 21: Relationship Between Median Age at Diagnosis and and the Percentage of Patients with One or More Comorbidities *– The median age at diagnosis and the percentage of the population with one or more comorbidities of interest at the point of cancer diagnosis is represented for each of the Cancer Site Specific Cohorts. The line applied to the plot is a linear regression model of x~y with a 95% confidence interval shown in grey*

3.3 - Discussion

3.3.1 - Demographics:

PPM Cohort

It is important to consider the basic composition of the population on which our analysis is based, to identify whether there is any form of overt selection bias present. The basic demographic data demonstrates differences between both the PPM cohort and the cancer cohorts, when compared to the national and regional population data.[203] The PPM dataset demonstrates a bias towards patients in the 20-50 age range, with a decrease in the proportion of patients in the under 20s. It is important to consider here the differences in the way these population numbers are generated. The national data is an estimate of the number of people currently living in the area. The PPM numbers, are based on patient registrations, not their current geography. As such, patients who move to the area, register with the hospital, and move away, will still be represented in the PPM data, where they won't in the regional data. This creates a bias, such that patients of an age where they move more geographically, are more likely to be represented. Government data shows those aged 20-55 are the most common in recent house movers[205], likely because it includes both students and economically active individuals. It is therefore possible that this registration effect is a large contributor to the differences seen. Additionally, it is important to note that the median age calculations on the national data were based on 1 year increment data. By contract the PPM data was only available in 5 year bands, with the figure based on the lowest value in the age band. As such, the metrics are not in fact entirely comparable. These differences, although potentially explainable, should be considered when interpreting any subsequent results. Further research is needed to identify if the highlighted demographic differences seen in the PPM cohort are similar in other health regions, to understand whether this is a systematic selection bias across secondary healthcare datasets in general, or a local phenomenon.

The assessment of PPM registrations shows non-linear patterns of growth over time, with 2012 being particularly notable, accounting for almost a quarter of total patient registrations. These differences from year to year are likely to be not only because of demand differences, but additionally process differences. A key example of this is highlighted in section 2.1.2 which identifies that in 2011 PPM was chosen to be used as the main EPR system. As such it is likely that the 2012 large increase is as a result of adding a significant volume of new patients, as the new EHR was rolled out across the trust. Other large increases in other years may represent system changes as opposed to demand driven differences. Although this data was not available this could be confirmed through the use of hospital activity data such as outpatient clinics, scans etc. and could be a potential focus of future research. It does however serve to illustrate the importance of understanding data provenance in interpreting data analysis.

Cancer Cohort

The All cancer Cohort Demonstrates a number of differences compared to both the PPM cohort and the regional population[1]. The significant increase in median age compared to the background population is a reflection of the underlying pattern of disease presentation, with oncological processes usually taking many years to develop[206]. As such, cancer is predominantly a disease of older age.[19,42,207] The generally older population being studied introduces a number of other considerations as age is associated with other forms of chronic health problems beyond cancer[42]. As such, using general population data as a baseline for mortality would be inappropriate as this would not account for the differences in the makeup of the cohort. Although this could be mitigated through the use of age stratified population data, there are other factors that would need to be

considered such as gender[208–212] and deprivation[213–215] levels both of which have been shown to be impactful on survival outcomes in both the cancer and non-cancer related settings. An alternative approach would be to use a matched cohort based on these important characteristics.[216] This process however fell outside of the ethics and scope for this study and has therefore not been conducted. It does however identify an area of potential future study. The link between age and comorbidity in the cancer cohort has been analysed further and can be found below in section 3.3.6.

The results presented for the All Cancer Cohort, as described above, treats all cancer patients as a single cohort. Although this highlights how the cancer cohort demographics diverge from the general population, differences between subpopulations of the cancer cohort also occur. The literature clearly outlines differences in the age of presentation of many of the most common cancers by prevalence. These differences are demonstrated in the PPM cancer data when assessing the population characteristics of the site specific cohorts. **Table 4** highlights these differences with testicular cancer showing a preponderance of younger patients compared to both the all cancer cohort and other cancers such as prostate cancer. The overall trend towards older patients in the all cancer cohort is largely driven by the fact that the most common cancers (lung, breast, colorectal and prostate cancer) are more common in older patients. The site specific cohort analysis highlights differences in the gender make up of different cancers. Some are gender specific such as ovarian and prostate, where others simply show a preponderance of a gender such as women and breast cancer. The noted cases of gender specific cancers showing patients of the opposite gender may represent data errors or how the hospital codes data for people who have are transgender.

The heterogeneity seen between different cancers is a key reason for the need to conduct site-wise analysis, rather than analysing the entire cohort as one. Heterogeneity is also seen in deprivation levels across the cancer sites. As age, gender and deprivation vary between cancers, are temporally precedent to the cancer diagnosis, and impact on survival, they can be considered potential confounders and be incorporated into future analyses.

3.3.2 - Missingness

The proportion of patients with missing data is important in determining how to appropriately analyse and interpret each data item.[157,158] In cases where data is missing, a number of approaches may be taken including complete case analysis and imputation all of which are detailed in chapter 2.[159,161217]. In these approaches there is a fundamental assumption that must be made, which is that those cases with missing values are missing at random or completely at random. This is important as otherwise the exclusion of these cases will introduce a selection bias, or in the case of imputation, results of imputed values may be highly unrepresentative of the true unmeasured value. When considering this assumption, there are plausible reasons to suggest that the missing data is missing not at random. Although the analysis does not demonstrate a clear relationship with missingness and year of diagnosis (**Figure 7**), there are a number of other possible ways in which missing data may in fact occur systematically. In many cases the stage and grade data will be input at review during a multidisciplinary team (MDT) meeting. Patients who may bypass this process will include those too frail to proceed with full investigation and treatment, or those who are at such an early stage of disease that they are never sent to an MDT. Systematic variation in how data is input will therefore introduce bias by ignoring the fact that the data is missing. An alternative approach in such a situation is to treat the missing value as a category of that variable. This will enable the analysis to identify whether or not the presence of a missing value is in some way influential in the outcome of interest.

Beyond the pattern of missingness, consideration of the scale of missing data is also required. The high level of complete data for age, gender and deprivation makes these variables excellent candidates for the inclusion in any subsequent modelling strategy. It is however worth noting that as highlighted above, data items may be complete but still inaccurate as potentially demonstrated in **Table 4** where gender specific cancers have patients of the opposite gender. The scale of missing data in relation to grade and staging however poses significant issues. The large percentage that is missing makes imputations of little value as in many cases we would be imputing a greater proportion of cases than those being used to generate the imputations. This is unlikely to result in reasonable or reliable estimation of the missing values. A general rule of thumb that is advocate by some literature is that the cut off for imputation should be 20% missingness[162] as after this point there is a high risk of bias being introduced.

It is also worth considering that as per the features of confounders detailed in chapter 2 grade and stage may in fact be mediators of effect in later analysis of the impact of comorbidity.[165] Thus in addition to concerns over missing data their relationship with both our exposure of interest and outcome might cause increased bias via adjustment. This concept is explored further in chapter 5. To further assess this a sensitivity analysis was undertaken for the results of chapter 5 and 6 with and without grade and stage data, the results can be found within the appendix (**Table 27**).

The analysis of missing grade and stage data, along with the proportion of missing ethnicity data, highlights one of the common challenges and pitfalls of routinely collected data. Although they commonly contain a diverse range of data items, much of the information is found within the free text of clinical documents rather than in structured data or are simply missing all together. The structured datasets derived from RCD are therefore commonly sparse, requiring important data items with a previously demonstrated link to disease outcomes to be excluded from analysis. This is a methodological requirement with high levels of missing data, however represents a clear and potentially significant limitation.

3.3.3 - Patient Geography

Figure 8 highlights potential ways in which geography may impact on data accuracy. In locations where patients live approximately equidistant between two hospitals then the total cohort of patients in that area may well be distributed between the two different hospital trusts. Individual patients may have their care split, such that on some occasions they attend one hospital, and on other occasions attend another. This will impact on the accuracy of data held within the PPM research dataset, as those closest to the hospital are more likely to have complete data, as compared to those further away. This can be demonstrated using a fictional patient that is known to be diabetic, is admitted to Bradford Royal Infirmity with a heart attack, but subsequently is diagnosed with lung cancer which is managed in Leeds. In this situation the patient will erroneously be attributed to the non-myocardial infarction and non-diabetic group, as the clinical coding data from the patient's acute admission is siloed in Bradford and therefore not within the research dataset.[65,66,126]

This effect may be more pronounced in cases where care is a supra-regional service. Examples of this include radiotherapy provision, ovarian cancer care and sarcoma patients. These patients may travel large distances for their cancer care, which is unlikely to be true for other routine care.

There are additionally a number of locations a significant distance away from Leeds which suggest that the majority of patients from this area are exclusively managed by Leeds. This is implausible and may be explained by how the data is generated. The postcode allocation within the analysis is based on current or most recently documented postcode, rather than postcode at the point of diagnosis.

This means that patients treated in Leeds who subsequently relocate will appear in their new location. The further away the patient is, the less likely it is that others will be referred to Leeds and thus it may be that many of these LTHT predominant locations far away, represent a single or small number of relocated individuals, rather than a systematic referral process from that location to LTHT.

3.3.4 - Accuracy of Clinical Coding

The assessment of diabetic cohort size identified variation in numbers across each of the DM definitions employed. In the all cancer cohort both HbA1c and clinical coding fail to identify a large proportion of diabetic patients, resulting in smaller sample sizes, misclassification error and therefore biased estimates of subsequent analysis.[218] When the population was limited to the LTHT blood catchment area, there was a meaningful improvement in the performance of HbA1c, reducing missed diabetic patients from 26.7% to 8.1%. When considered in the context of the boundary effect results described above, it is likely that a significant proportion of the diabetic patients missed by HbA1c are due to their routine blood sampling being analysed at a different hospital. The variation in distance travelled by cancer site might mean that the level of inaccuracy of HbA1c might vary by cancer site, although this was not directly assessed.

Clinical coding by contrast has a lower fidelity when used in the LTHT blood catchment area, with a missed diagnosis rate increasing from 18.9% in the full cohort to 31.1% in the LTHT catchment area. This fall off in performance may suggest that even when combining HbA1c and clinical coding a proportion of diabetic patients are still being missed as the diagnostic blood results that could identify them are not contained within the dataset. The performance of clinical coding in the full cohort may therefore be worse than estimated, as there is ongoing misclassification error for diabetes which is not able to be assessed or quantified. In either case, the use of the hybrid definition substantially increases the diabetic population size and assuming that the coding and blood results are not due to an error, then it will additionally minimise misclassification bias.

When assessing the cohort of patients identified by both HbA1c and clinical coding, it is possible to identify that clinical coding also introduces error through the timing of diagnosis. Large discrepancies are identified, with HbA1c evidencing diabetes on average earlier than clinical coding. This becomes particularly important in analyses relying on an index date, such as the date of cancer diagnosis. The data identifies that 17.5% of pre-cancer diabetic patients are missed due to the diagnostic timing error of clinical coding. This further compounds the issue of missed diagnoses introducing further misclassification error.

The subsequent modelling data identifies that these areas of inaccuracy are not only of theoretical importance, but are also impactful in terms of analysis outcomes. **Figure 11** and **Figure 12** identify clinically meaningful differences in the risk that is attributable to DM depending on the definition used. Of interest, clinical coding is shown to be the most pessimistic, HbA1c the most optimistic and the hybrid definition somewhere in between the two. It is possible that this is due to the patients being missed by each group being systematically different i.e. they are missing not at random. As clinical coding requires an admission, these patients may be more likely to have serious health issues that have warranted a hospital stay. Those patients with well controlled diabetes and no ill effect will be managed in the community but identified on blood testing. Thus those missed by clinical coding may be the healthy patients with potentially better outcomes. By contrast the patients missed by HbA1c have had an admission and are more likely to be from out of area, increasing the likelihood that their issues are serious enough to warrant admission but that their routine care and bloods are being done outside of the tertiary centre. This means that the missed patients are likely

to be amongst the sickest of patients. By making use of the hybrid definition it is possible to remove some of the selection bias and misclassification error introduced by any one definition alone. The error in diagnosis timing is also overcome by using this approach. It is worth noting that irrespective of the definition used DM is associated with worse outcomes in the All Cancer Cohort with variation in the scale of effect, but not the direction of effect. A more detailed analysis and discussion of diabetes and cancer outcomes can be found in subsequent chapters and is not covered in detail here as the main focus is data accuracy.

The highlighted accuracy difference based on geography does identify a limitation of HbA1c used in a single centre setting. The robustness of the hybrid definition could be further enhanced by including data from multiple centres. If all clinical coding from all hospitals and all blood results were available it is likely that the accuracy would be further enhanced. If primary care coding was included, this would likely further improve the identification of the diabetic cohort. The caveat to this is that with increasing numbers of sources, there is an increased chance of an erroneous coding or blood results causing misclassification in the other direction.

The differences in patients being identified by each of the DM definitions highlight potential areas of concern surrounding previous research in this area or how research output is used in clinical practice. In the case of studies assessing the impact of diabetes on health outcomes in cancer or otherwise, a reliance on clinical coding may lead to overestimation of the true hazard. This could have a profound effect on patient and clinical decision making if a patient's risk of an adverse outcome are in fact smaller than those quoted in the literature.

Additionally, it suggests that the true diabetic cohort and the coded diabetic cohort differ in terms of outcomes. As such, comorbidity scores that include diabetes for outcome prediction, may produce incorrect risk estimates when deployed in the clinical setting.[38,39,139] If a score is developed on clinical coding but used in practice by collecting information from patients, and these two populations differ, then the risk estimates will be inaccurate.

Although our analysis focusses on diabetes it is plausible that similar accuracy issues may occur in other conditions when using clinical coding alone. If this is the case, then comorbidity scores may be even more severely affected, due to the combined effect of errors from each comorbidity. This provides weight to the argument that scores developed using coded data should only be used prospectively on coded data, unless it has been specifically validated in a separate analysis using other sources of comorbidity information. This identifies the realistic possibility that many current clinical risk scores used in clinical practice may be providing incorrect outcomes estimations for many individuals.

It is important however to caveat these concerns with the limitations of this data. Firstly, this is based on a single centre and might reflect more inaccurate clinical coding locally compared to nationally. Furthermore, many risk scores are based on international data such as US claims data, and the errors in these forms of data may differ. Further study on national and international data is therefore needed to assess if this issue is local or more widespread. Issues with clinical coding accuracy have been identified extensively in the literature previously[219–222]. In each of these cases the analysis has been based on the accuracy of coding in patients after an admission event. It appears that this is the first analysis of clinical coding on a population level that includes patients both with and without admission events.

As a result of this analysis of coding accuracy, the hybrid definition of diabetes has been used in subsequent analyses. Results for clinical coding alone is still presented with type 1, type 2 and other

diabetes as the hybrid definition is not able to distinguish these. Although this analysis provides an argument that other conditions should have had blood test results used to enhance clinical coding, this was not undertaken for other conditions. This was for two key reasons, firstly the lack of a clear single diagnostic test for many conditions and secondly due to limited access to the results of other blood tests.

3.3.5 - Prevalence of Comorbidity

Table 6 identifies important differences in the cohorts analysed. In general, it appears as though the population of the North East and Yorkshire has a higher level of ill health than the English Average. The underlying pathological processes that drive each of these conditions will be multifactorial including genetic, environmental and lifestyle factors.[42] The differences seen could therefore represent regional differences in genetics due to high numbers of particular ethic groups within the region. An example of this is the population of people with heritage from the sub Indian continent is higher in Yorkshire than the national average. Previous research has demonstrated an association between this group and increased cardiovascular risk.[223–226] It is import to note however that the literature fails to come to a consensus as to whether this is truly a genetic phenomenon, due to lifestyle differences, or a combination of the two.

A further explanation may be environmental, particularly in view of the high levels of manufacturing, mining and heavy industry that have taken place within the region in previous decades. This may resulted in a larger proportion of the population having significant occupational exposures that may impact on their health.[227–229] Furthermore, lifestyle differences and socioeconomic differences may lead to behaviours that increase the risk of ill health. The higher rates of obesity within the region are highly suggestive of this as lifestyle and environment are thought to be greater contributors than genetics.[230]

When comparing the rates of comorbidity in the cancer population to the regional population there is a less clear overall trend. When interpreting these results it is important to consider the data generation process for each. The NHS Digital regional data is reliant of GP clinical coding.[231] Thus a condition where a patient is managed within the community is likely to be well recorded as are important hospital events with good survival rates, such as myocardial infarction. The cancer cohort data is predominantly based on hospital clinical coding.[126] As this data is only generated for patients with an acute admission, it may fail to capture conditions managed in primary care or in the outpatient setting. Thus patients who are otherwise healthy at the point of cancer diagnosis and whose cancer does not require admission, might never have had a hospital admission and thus have no clinical coding. Patients who are subsequently treated have a much higher probability of admission to hospital, either electively for surgical procedures, or acutely with the effects of their cancer or its treatment.

These different data collection models may explain the differences seen when comparing regional prevalence to the point of cancer diagnosis and comparing the point of cancer diagnosis to the lifetime cancer cohort prevalence. At the point of cancer diagnosis many patients may never have had an admission, thus the lower rates shown for several health conditions may not be a true lower prevalence, but simply a reflection of a lack of coding. The plausibility of this explanation is increased by virtue of the fact that MI, which forms part of coronary artery disease is a condition requiring admission and hospital diagnosis. This comorbidity is more highly represented in the cancer cohort than the background population. An alternative explanation for this would be the number of common risk factors for multiple health conditions.[28,42] Obesity, smoking, lack of physical activity, diet and a number of other lifestyle factors are common risk factors for both cancer and heart

disease. As a result, the high levels of heart disease in the cancer population could be due to shared pathogenesis.

When assessing the differences between baseline levels of ill health and lifetime levels in the cancer cohort there is a general trend towards large increases in ill health over time. This could be due to cancer and its treatment causing side effects and long term damage increasing multimorbidity. Many cancer treatments have well established side effects that can lead to acute or chronic problems. This includes the impact of surgery, radiotherapy and systemic anticancer treatment. Examples of this include lung fibrosis from radiotherapy[232], cardiovascular, renal and respiratory complications from systemic anti-cancer treatments[233]. Additionally cancer itself can cause issues by direct structural damage or by altering physiology such as increasing thrombotic risk.[234] The high levels of ill health over the lifetime of cancer patients when compared to the general population could be due to the disease and its management. The large shift from baseline ill health in the cancer cohort to lifetime risk would initially seem to lend weight to this conclusion. This large shift in prevalence may however be partly artefactual and due to the way that the data is collected. As mentioned previously, hospital clinical coding requires an admission and the risk of admission is increased by having cancer and treating it. As a result, longstanding health conditions may be first coded after cancer diagnosis even though they have been affecting the patient for a significant period of time. The previous analysis of HbA1c and clinical coding diagnosis date would add weight to this theory.

Thus far the focus of discussion has been on the data derived from the all cancer cohort. Results from the site specific cohorts demonstrate variation in the levels of comorbidity seen. Some cancers were associated with lower levels of comorbidity for almost all conditions when compared to the all cancer cohort, these included breast, cervical and testicular cancer. Other sites such as bladder, renal and pancreatic cancer show typically higher levels of comorbidity. This difference might be explained by the average age at diagnosis, with a greater proportion of young patients in those with lower levels of comorbidity. It may also be due to a bias in the all cancer cohort towards cancers with greater numbers, such that the high comorbidity burden seen in lung and colorectal cancers pulls the All Cancer Cohort average up. If the class imbalance seen was addressed with up sampling lower frequency cancers or down sampling higher frequency cancers the results might shift.

Making a clear conclusion as to what is true effect, bias and data artefact is made even more challenging in the adult cancer population due to the difficulty in separating out the effects of aging and cancer itself[42]. As demonstrated in **Table 4** and **Figure 6** the cancer population is on average older than the general population. The older someone is the, greater their lifetime exposure to risk factors for ill health, thus increasing the risk of developing significant health conditions. This is also particularly relevant given the common risk factors for cancer and many common health conditions. The higher rates of ill health in the cancer population could therefore simply be due to the selection bias introduced by limiting the analysis to the cancer population. To further understand what may underpin the patterns seen a further analysis of the association between age and comorbidity has been conducted. Further analysis has been undertaken using more granular data on the timing of comorbidity diagnosis both of which are discussed in more detail below.

3.3.6 - Age and Comorbidity

The analysis of the relationship between age and comorbidity count demonstrates an overall trend towards increased levels of comorbidity with increasing median age at diagnosis. As demonstrated in **Table 4** the majority of patients have no comorbidities at the point of cancer diagnosis, however a significant minority have one or more conditions. Using a standard Poisson model to estimate the relationship between age and comorbidity count may therefore be prone to errors as a result of an

excess of zero count patients. A zero inflated Poisson mode was therefore developed which in effect fits two models, the first a logistic regression to estimate the relationship between age and the likelihood of having no-comorbidity and a subsequent Poisson model to estimate the relationship between age and the comorbidity count, accounting for the excess of zeros. The results clearly show that increasing age results in a reduced chance of having no comorbidity and additionally increasing age is associated with a small but significant increase in the chances of having a higher comorbidity count. The results of the zero inflated model were compared to those of a standard Poisson regression model using the Vuong test. This demonstrated that the zero inflated model performed better, however it should be noted that several previous research papers has identified limitations to the Vuong test.[173]

The nature and scale of this relationship between age and number of comorbidities along with the multiple pairwise correlations identified suggests that it may be advantageous to consider comorbidities in isolation when analysing the data for the purposes of effect estimates. This is because of the high levels of collinearity between these interrelated variables. The counter argument could however be made that in doing so there is a risk that part of that effect size may be due to another comorbidity and that the condition used is in fact showing all or some of its effect as a result of confounding as opposed to a true direct effect. Due to the complex interplay between comorbidities, age and other variables of interest along with variation in the temporal ordering of the development of comorbidities, subsequent inferential analysis will focus on individual comorbidities.

3.3.7 – Timing of Comorbidity Diagnosis

The biphasic nature of the distribution of the diagnosis of concomitant illness in cancer patients identified in **Figure 15** raises important questions about how patients with cancer present, and how cancer and its treatment may impact on the long term health of patients. A possible explanation for the patterns seen is that as patients develop cancer they attend community care or acute hospital services where previously unknown health conditions are identified as part of the investigations, which ultimately result in the diagnosis of cancer. Similarly, as patients are treated for cancer in the months after their cancer diagnosis, the exposure to significant health interventions such as surgery, radiotherapy and chemotherapy results in the triggering of significant health events. The body of research around ill health in cancer would seem to agree with this explanation of the patterns seen, with multiple research papers in both the hospital and community setting showing how cancer triggers significant health events with treatment and identifies ill health during the diagnostic process.[233]

There are however several other possible explanations which would suggest that the patterns seen are entirely artefactual. As detailed above, clinical coding is only generated by a hospital admission. Many of the diagnoses identified by clinical coding may have been present for significant periods of time prior to a patient's first admission. This has been clearly demonstrated by our analysis of diabetes mellitus diagnosis. As a result, the peaks seen may in fact represent peaks not in diagnoses but peaks of hospital admissions, with many patients being inpatients before their cancer diagnosis and being admitted as part of their post diagnosis treatment.

In the case of obesity, where identification includes the use of height and weight measurements these peaks may simply represent the first time a patient is weighed in the hospital setting, as part of their initial diagnostic pathway, or as part of their post diagnosis treatment pathway. As the development of obesity is a gradual process the initial spike seen at the point of diagnosis with consistently low levels prior to this would seem to be incongruous with the manner in which

individuals put on weight. Although there is an association between obesity and the development of cancer[235], it would be more likely that at the point of cancer diagnosis patients would experience unexplained weight loss as opposed to rapid unexplained weight gain.

It is important to recognise the differences in how different illnesses are diagnosed and how this may impact on their identification within clinical coding. Myocardial infarction for example is a condition which necessitates hospital investigation and intervention. As such, most patients with an acute ischaemic event will be admitted to hospital at the time of the event. This means that theoretically, one would expect the clinical coding to capture most cases with reasonable accuracy. By contrast, hypertension is usually diagnosed in the community and often requires long term follow up of patients to confirm the diagnosis. Furthermore, unless the level of hypertension is extreme, it does not necessitate admission at the point of diagnosis or at a later stage. As such, it is more likely that hypertension is accurately coded in primary care and is inaccurately coded in secondary care systems. When comparing the diagnosis date patterns of these two conditions the initial hypertension peak at the point of cancer diagnosis is three time higher than that of MI. Additionally, MI has no discernible second peak. The theoretical higher accuracy of MI may suggest that the pattern seen here is more likely to be the true pattern where hypertension may include more artefact. Key differences are a smaller initial peak of MI and the absence of a second peak in the months after diagnosis. It could therefore be argued that the likely true pattern is a smaller but true peak of diagnosis around the point of cancer diagnosis.

A reasonable counterargument to this is that different conditions may have very different patterns of presentation. Certain oncological treatments are known to cause risks of significant health events, for example 5-flurouracil and related drugs can cause acute coronary artery vasospasm[236], lung radiotherapy is associated with the development of lung fibrosis[232], Herceptin can cause cardiac dysfunction and cardiac failure.[237] Thus comparing different conditions in the manner conducted above could equally lead to the false impression of artefact where none does in fact exist.

Further research in this area is needed to better understand the relationship between the timing of concomitant illness and cancer. This area of research into the late effects of cancer has been noted to be particularly challenging in the adult population.[37] Distinguishing between the effects of aging and the effects of oncological diagnoses and treatment can be extremely challenging. Research in paediatric population can however act as a guide as the influence of aging is less problematic. Here, research has demonstrated clear links between cancer and a significant increase in the burden of ill health amongst survivors. In order to analyse this area of oncology in the future two further sources of information will be needed. Firstly access to coding from both primary and secondary care. This would enable the differences in the setting in which ill health is identified to be largely overcome. This would however fail to overcome the issue of inaccurate data and coding. An additional source of information would be a non-cancer control cohort. This could include age, gender and deprivation matched controls. This would allow the rates of comorbidity and the patterns of comorbidity to be compared to identify either an excess of events in the cancer cohort or differences in the time to the diagnosis.

3.3.8 - Collinearity

Our pairwise association analysis identifies positive correlation between most of the variables analysed. Although many were only weakly correlated several showed more sizable correlation. In almost all cases the correlations were highly statistically significant. The value of these significance tests is questionable in part due to multiple comparisons but also due to the large numbers of patients used in the analysis. Despite this the identification of multiple pairwise associations

highlights the issue of multi-collinearity in health data research. Many conditions share similar risk factors such as genetics, smoking, and diet. In addition to this, many condition's underlying pathogenic process may cause other forms of ill health such as cardiovascular disease due to hyperglycaemia induced endothelial damage in diabetes. The interconnected nature of multimorbidity means that collinearity is almost certain when using multiple diseases in a single model.

There is a wealth of evidence of the impact of introducing collinear variables into a regression model. It results in both bias and also imprecise estimation of the effect size.[163] Both of which are important if we aim to estimate the impact of comorbidities on survival in cancer cases. Although the scale of correlation between individual pairs is in most instances is low, the inclusion of many variables with low levels of correlation could result in significant errors in the effect sizes obtained by analysis. It would be impossible to know how imprecise the estimate was and the direction of effect of the bias. As a result, including multiple comorbidities in a single inferential model has the potential to yield extremely biased results. Additionally using aggregate scores of the count of comorbidities could have similar effects due to the association between age and comorbidity.

3.4 - Summary

This initial descriptive and exploratory analysis has highlighted a number of potential sources of bias within the underlying research dataset. Basic demographic data highlights a selection bias with an older population with high levels of comorbidity. Several data items show information bias with high levels of missing data, misclassification error and issues surrounding data accuracy. Within the individual data items there are meaningful associations and correlation which highlight potential issues with multicollinearity in later modelling. For some of these issues methods will be introduced to mitigate them through enhancement of comorbidity data, analysing comorbidities singly and omitting data items with large sale missing data. Despite these strategies much of the potential bias identified will persist and thus it is import that these issues are regarded as limitations of the analyses and incorporated into the interpretation of subsequent findings.

In order to complete our exploratory analysis the next chapter will focus on outcomes data and apply Kaplan Meier estimates to demonstrate the overall pattern of survival outcomes in patients with and without comorbidity.

Chapter 4 – Survival Outcomes in Comorbid Cancer Patients Using a Univariate Kaplan Meier Approach

4.0 - Introduction

Within the medical literature one of the most commonly employed approaches to survival analysis is the Kaplan Meier method[107,238]. Chapter 2 provides a detailed summary of this time-to-event analysis approach. The application of this method will enable the estimation of the incident rate of the event of interest, in this case death, making use of information relating to all the individuals at risk for the event. This is an important feature, as although this analysis seeks to use 15 years of follow up data, individual patients may not have data available for this full period. This would be the case for patients lost to follow-up, or patients who are alive and have received their diagnosis within the last 15 years. This issue is overcome by making use of censoring[175,176] to include all patients, for as long as data is available for them, allowing the use of the entire at risk population, at every time point.

This approach may be used to create a time-to-event estimate in a given population as a function of a single variable of interest. This approach is commonly employed in the literature to compare the outcomes after the use of a given medical intervention, where the median survival or time point specific survival may then be compared for each group[174]. By taking the same approach it is possible to generate a Kaplan Meier estimate of overall survival as a function of the presence or absence of comorbidities of interest. The resulting survival functions can then be used to identify conditions associated with meaningful differences in survival outcomes. As this approach does not try to make adjustments for potential confounding, this may be considered as an exploratory or descriptive analysis[89]. This approach is an estimate of the survival of the cohort divided into groups based on an identifying characteristic. As detailed in chapter one, the literature in this area is almost universal in suggesting that common comorbidities are associated with worse survival outcomes.[19,239,240] Despite this, no comprehensive analysis has been conducted on a single large dataset to identify the effect of multiple individual comorbidities, in multiple common cancers. Within this chapter the analysis builds on the initial descriptive and exploratory analysis conducted in Chapter 3 and will describe the survival outcomes of groups defined by the presence or absence of the previously detailed comorbidities in the site specific cohorts. The results of these analyses will be used to identify any differences in the survival outcomes overall within our dataset and compare these to national averages. The comorbidity analyses will guide which cancers and which comorbidities should be investigated in more detail with more sophisticated methods in the subsequent chapters.

4.0.1 - Aims and Objectives

1. Quantify differences in survival outcomes within the Leeds Cancer Center dataset compared to national survival data.
2. Describe the survival outcomes of cancer patients with and without comorbidity.
3. Identify which cancers and comorbidities have consistent survival effects for further detailed analysis.

These will be delivered through the following objectives:

a. Estimate the median survival for each site specific cohort.
b. Compare local median survival to published site specific cancer outcomes.
c. Estimate median survival for the site specific cohorts when stratified by the presence or absence of comorbidities of interest.
d. Statistically assess differential survival outcomes using log rank test.

e. Assess differential survival for clinical significance.
f. Quantify the number of cancer sites across each comorbidity with significant differences, to determine those comorbidities which are associated with the consistent survival differences.
g. Quantify the number of comorbidities in each cancer site with significant differences to determine the sites, which are associated with the consistent survival differences in patients with comorbidity.

4.1 – Methods

4.1.1 - Survival Methods, Whole Population Survival and Summary Statistics

Analysis was undertaken using the site specific cohorts detailed in chapters 2 and 3. Kaplan Meier survival estimates were obtained using the R "*survival"* package and plots to visualised survival were produced using the "*survminer"* and "*ggplot2*" packages. Full details of R packages are found in **Table 23** of the appendix. Summary survival estimates were extracted in the form of median survival from the Kaplan Meier estimates produced. In both the all cancer cohort and site specific cohorts, initial estimates were produced on the whole cohort without stratification. In each case a whole population overall median survival was extracted.

Population survival statistics were identified from Macmillan Cancer support publications.[241,242] Where cancer site statistics were available from both our analysis and Macmillan data, the two were compared with the relative percentage difference calculated in each case.

4.1.2 - Stratified Survival of Pre-defined Cancers and Comorbidities

In each of the site specific cohorts KM estimates were generated as a function of the presence or absence of each comorbidity of interest. Median survival for each strata was extracted, with the survival difference calculated in time and as a percentage, using the non-comorbid strata as the reference group for calculations.

Statistical significance was assessed using log rank score comparing the survival trajectory for those with and without a given comorbidity. A p value of 0.05 was used to determine statistical significance however due to the large number of comparisons results are demonstrated with a second p value adjusted threshold of 0.00005.[168]

Due to the large number of comparisons undertaken, detailed summary information was limited to key cancers and key comorbidities. The top 4 cancers by incidence namely breast, colorectal, prostate and lung cancer were selected with the impact of all comorbidities assessed in these populations. Comorbidities with a high likelihood of accuracy in our dataset were selected which included MI, stroke and diabetes for analysis of their impact across all cancers. A further description of these choices can be found in the discussion section below.

To improve the efficiency and ease of interpretation of analysis, results were converted into summary plots by cancer site and by comorbidity. In the case of cancer sites the plots demonstrate the percentage change in median survival in the comorbid group and whether this was statistically significant according to the log rank test applied. In the case of summary by comorbidity, these same summary plots were produced, along with a second summary plot with the median survival difference described by time. This allows for comparisons between cancers, which may often have highly variable median survival times, which is not clearly represented when using percentage change alone.

4.1.3 - Identifying Comorbidities and Cancer Sites of Focus

In order to identify other important cancer sites and comorbidities to conducted further analysis on, summary statistics were analysed to identify comorbidities where 90% or more of cancer sites had a statistically significant log rank score at the 0.05% level and at least 75% of cancers at the 0.00005% level. In the case of cancer sites the same thresholds were applied, but assessing the percentage of comorbidities meeting the threshold. Any comorbidities or cancer sites meeting this threshold were included for detailed analysis with summary plots as described above. Those fulfilling these criteria were additionally subjected to further detailed analysis in subsequent chapters.

4.1.4 - Assessing the Relationship between Comorbidity Impact and Whole Population Survival

Visual comparisons were made between the median survival of the total cohort and the survival difference caused by each comorbidity to assess the relationship between the two. In each case a line was fit to estimate the relationship between these two values. Formal correlation between median overall survival and percentage difference in median survival between stratified groups, was assessed using Pearson's correlation.

4.2 – Results

4.2.1 - Total Population Survival and Summary Statistics

Figure 22 demonstrates the overall median survival for each of the 24 site specific cohorts. The plot shows the bars in order of shortest to longest survival with cancer of unknown primary (CUP) patients' having the worst outcomes with 0.2 year median survival. The longest quantifiable overall median survival is cervical cancer, with a median survival of 23.8 years. Notably both testicular and thyroid cancers have no quantifiable median survival due to the high survival rates seen. This prevents the survival estimate ever reaching or falling below 50% survival, thus preventing median survival being estimable (**Figure 23**). This does not preclude analysis of significance in stratified curves as the log rank test can still be applied to compare the strata. Results for significance of stratification of these cancer sites is therefore available in subsequent sections where appropriate.

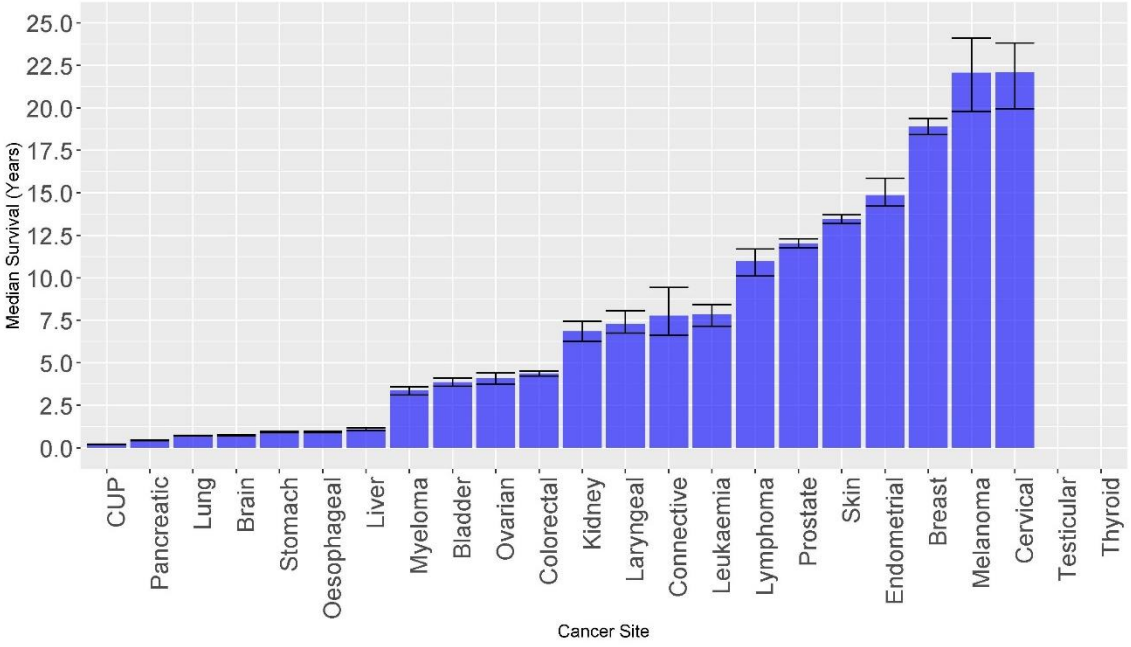


Figure 22: Overall Median Survival for Site Specific Cohorts – *Kaplan Meier derived median survival estimates in years and associated confidence interval for the whole population of each Cancer Site Specific Cohort*

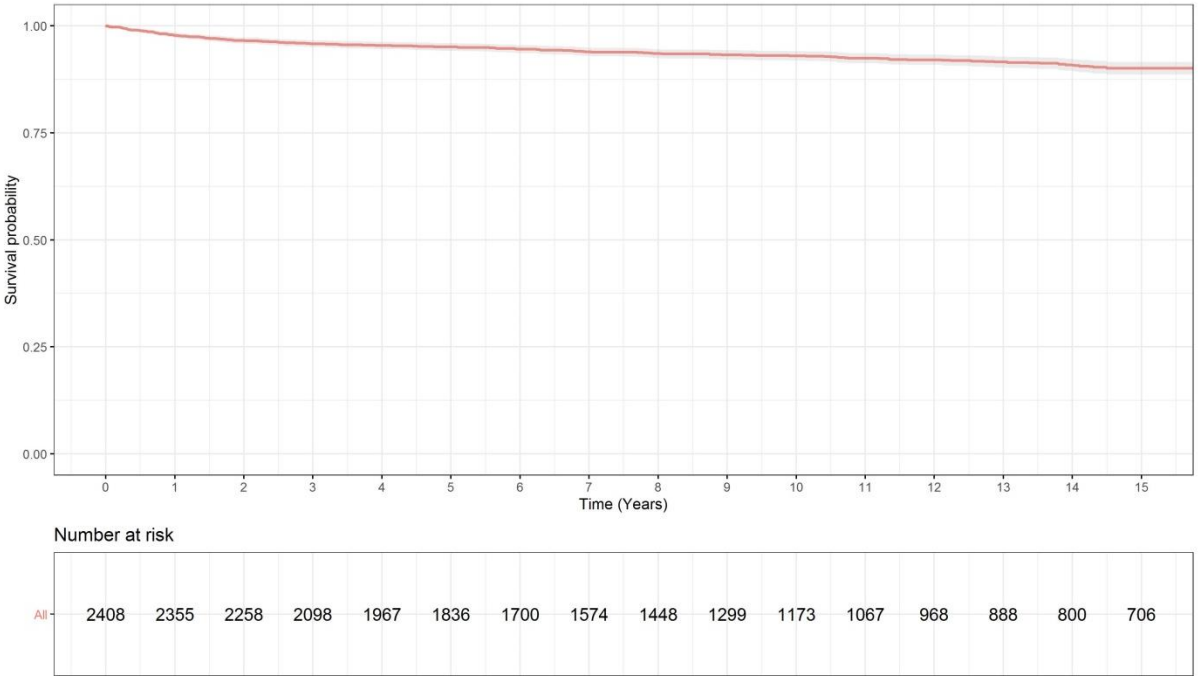


Figure 23: Overall Survival for Testicular Site Specific Cohort – *Kaplan Meier estimate for testicular cancer patients using the PPM Cancer Site Specific Cohort.*

Site	Published Median in Years	Calculated Cohort Median in Years (CI)	Percentage Difference
Bladder	9	3.9 (3.6-4.1)	-57%
Brain	0.5	0.7 (0.7-0.8)	40%
Kidney	5.3	6.9 (6.3-7.4)	30%
Leukaemia	3	7.9 (7.2-8.4)	163%
Lung	0.4	0.7 (0.7-0.7)	75%
Myeloma	2.5	3.4 (3.1-3.6)	36%
Oesophagus	0.7	0.9 (0.9-0.1)	29%
Ovary	3.1	4.1 (3.7-4.1)	32%
Pancreas	0.2	0.4 (0.4-0.5)	100%
Stomach	0.7	0.9 (0.9-1.0)	29%

Table 7: Comparison of Published Survival Estimates to Calculated Survival Estimates – *Median Survival in publically available data in years, median survival from PPM site specific cohort in years and the percentage difference in survival comparing the local data to national UK data. Not all cancers are included as only those with identical groupings in both the Macmillan data and site specific cohorts were included in the comparison*

When comparing the locally derived survival estimates to those found within the public domain[241], all but one site has improved survival locally. The most marked of these by percentage, are pancreatic cancer and leukaemia. In the case of leukaemia where overall survival is longer, this equates to a survival difference of 4.9 years longer in the local cohort. Conversely in pancreatic cancer where survival is shorter the large percentage difference represents a survival improvement of just 2.4 months. The survival from bladder cancer is however much lower in the local data representing a 4.9 year reduction in median survival compared to the population estimates.

4.2.2 - Stratified Survival

KM curves were generated as a function of each comorbidity of interest, for each cancer site of interest. With 24 cancer sites and 40 comorbidities plus the analysis of the all cancer cohort, this yielded a total of 1000 stratified KM curves. **Figure 24** is provided as an example of the output of this analysis. This volume of granular data is inefficient to analyse in its raw form, so was converted into summary plots as described in 4.1.2. A summary plot was produced for each cancer site and each comorbidity. The results presented however focusses on the predefined common cancers and robustly recorded comorbidities.

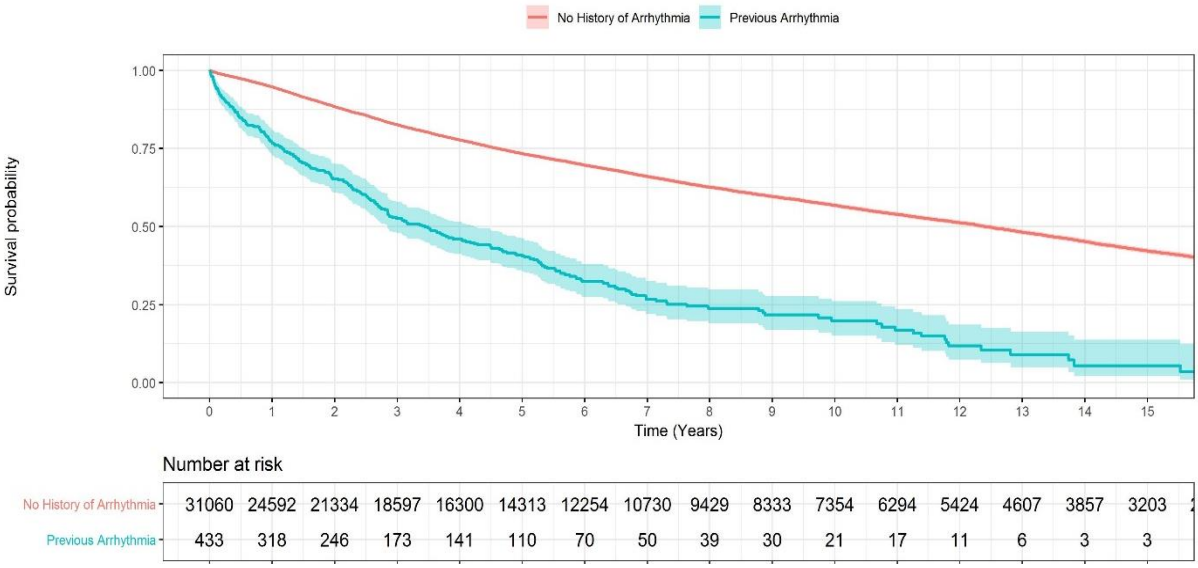


Figure 24: Breast Cancer Survival Stratified by History of Arrhythmia *- Graphical representation of Kaplan Meier estimate conditioned on the presence of an arrhythmia ICD-10 code within a patient's clinical records prior to the date of their breast cancer diagnosis*

In the breast cancer site specific cohort analysis (**Figure 25**), all comorbidities were associated with reduced survival times compared to the non-comorbid strata. Five comorbidities had an insufficient number of cases for estimates to be generated (cardiomyopathy, neuromuscular disorders, motor neuron disease, HIV and psoriatic arthritis). Four further comorbidities were found to be associated with reduced survival, however were not statistically significant (inflammatory bowel disease, malabsorption, pancreatitis and other respiratory disease). Of those with a statistically significant difference, the largest was spinal damage, highlighting a reduction in the median survival by over 99% ($p = 1.22 \times 10^{-25}$) taking the survival in this group to 0.2 years compared to 18.9 years in the non-spinal injury group . Large differences with a reduction of over 80% in median survival are seen with congestive cardiac failure (-89.2%, $p = 8.5 \times 10^{-185}$), restrictive lung disease (-82.6%, $p = 0.0053$), renal dysfunction (-83.2%, $p = 2.38 \times 10^{-67}$), dementia (-90.5%, $p = 7.12 \times 10^{-159}$), gout (-81.7%, $p = 1.09 \times 10^{-12}$), ankylosing spondylitis (-92.7%, $p = 3.72 \times 10^{-6}$) and peripheral arterial disease (-84.9%, $p = 3.54 \times 10^{-31}$).

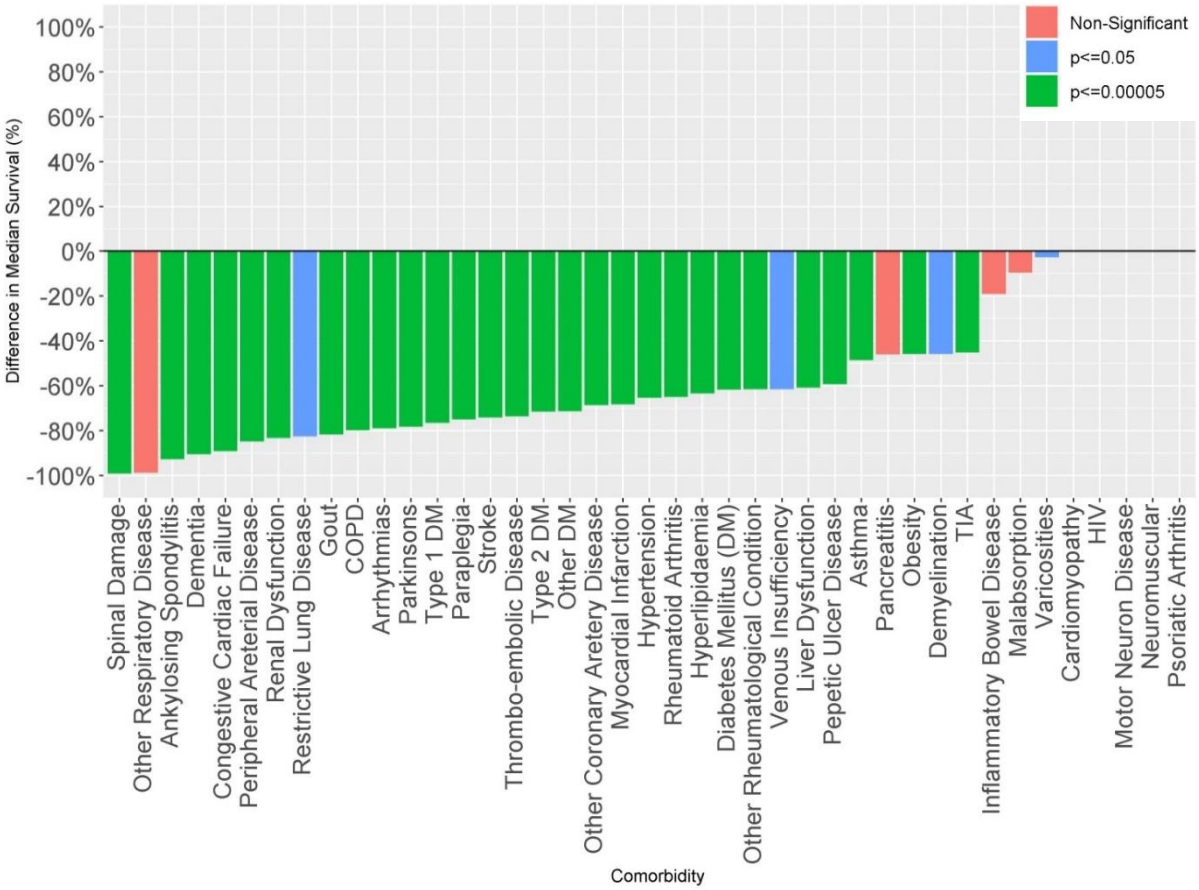


Figure 25: Impact of Comorbidity in Breast Cancer - *Percentage difference in median survival in comorbid group compared to non-comorbid group for each comorbidity of interest in the PPM Breast Cancer Site Specific Cohort. Negative values represent a reduced survival in the comorbid group. Significance levels are demonstrated by colour*

The colorectal cancer site specific cohort analysis (**Figure 26**) demonstrates a different pattern with some comorbidities being associated with improved survival outcomes. Varicosities were found to be associated with an 88.4% improvement in median survival (p= 0.029). Obesity demonstrated a 26.3% increase in survival, although this was not found to be statistically significant. Nine other comorbidities were found to have non-significant differences including motor neurone disease, ankylosing spondylitis, cardiomyopathy, HIV, pre-diabetes, psoriatic arthritis, other forms of DM, malabsorption and inflammatory bowel disease. The most impactful comorbidity by percentage was dementia (-86.6%, p = 9.34 x 10^{-43}) accounting for a 3.8 year drop in median survival. No other comorbidities had an impact beyond 80%. As with the breast cancer cohort some conditions had insufficient numbers to generate percentage change estimates, including spinal damage and venous insufficiency

The impact of comorbidity on lung cancer (**Figure 27**) was less pronounced with 14 comorbidities demonstrating non-significant effects. The direction of effect was more varied with three statistically significant improvements in survival seen in the cases of hypertension, (1.9%, p = 0.01), asthma (29.0%, p = 2.98 x +10^{-5}) and obesity (437.3%, p = 2.01 x 10^{-66}). As the median survival in lung cancer is relatively short, the survival time benefit for hypertension and asthma is modest representing 5 days and 2.4 months respectively. The estimated effect size in obesity is substantially larger equating to an over 3 year improvement in the median survival. The largest statistically significant negative impact was in motor neurone disease (-76.2%, p = 0.01) equating to just over 6 months reduction in median survival.

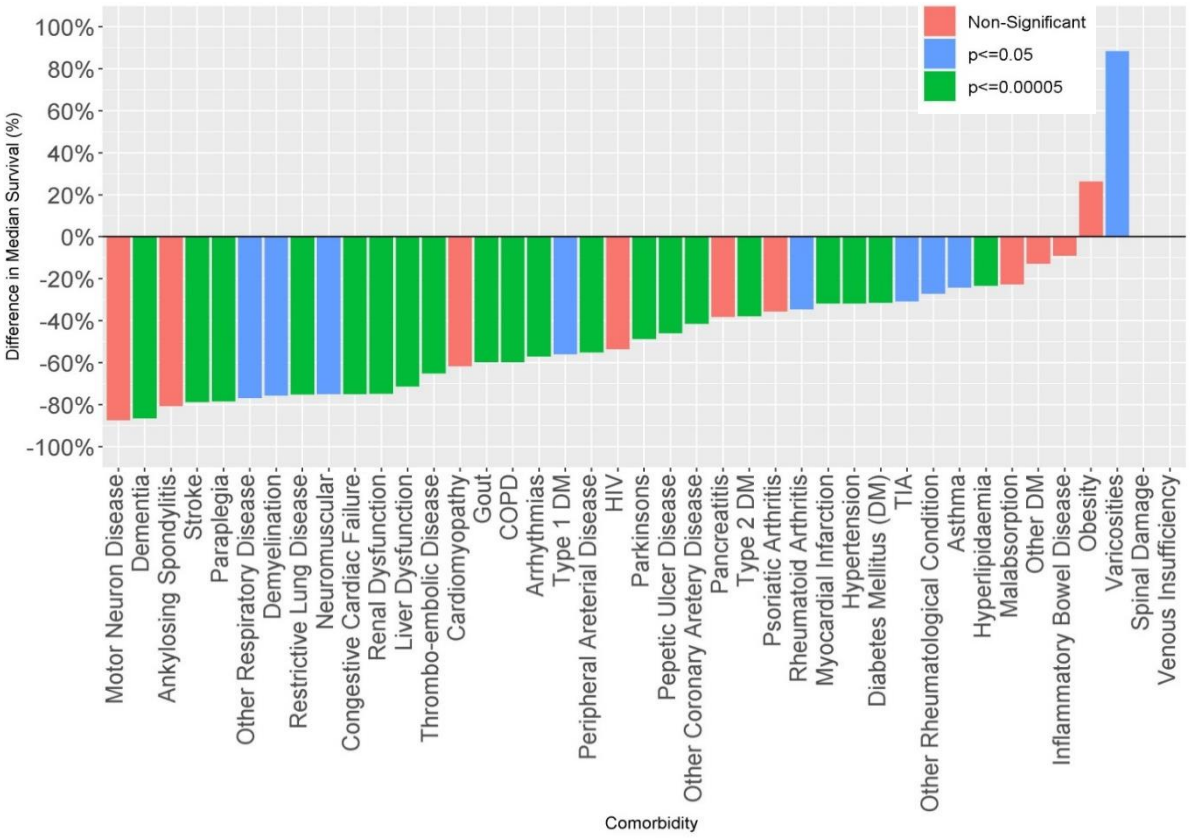


Figure 26: Impact of Comorbidity in Colorectal Cancer *- Percentage difference in median survival in comorbid group compared to non-comorbid group for each comorbidity of interest in the PPM Colorectal Cancer Site Specific Cohort. Negative values represent a reduced survival in the comorbid group. Significance levels are demonstrated by colour*

Prostate cancer (**Figure 28**) demonstrates a more uniform direction of effect with no statistically significant increases in median survival across the comorbidities of interest. Restrictive lung disease has the largest effect size of -82.0% ($p = 1.21 \times 10^{-17}$) equating to a reduction in median survival of 9.9 years compared to those patients without restrictive lung disease. Due to the relatively long median survival times in prostate cancer patients, even the least impactful statistically significant comorbidity, thromboembolic disease, results in clinically meaningful changes in survival (-2.9 years).

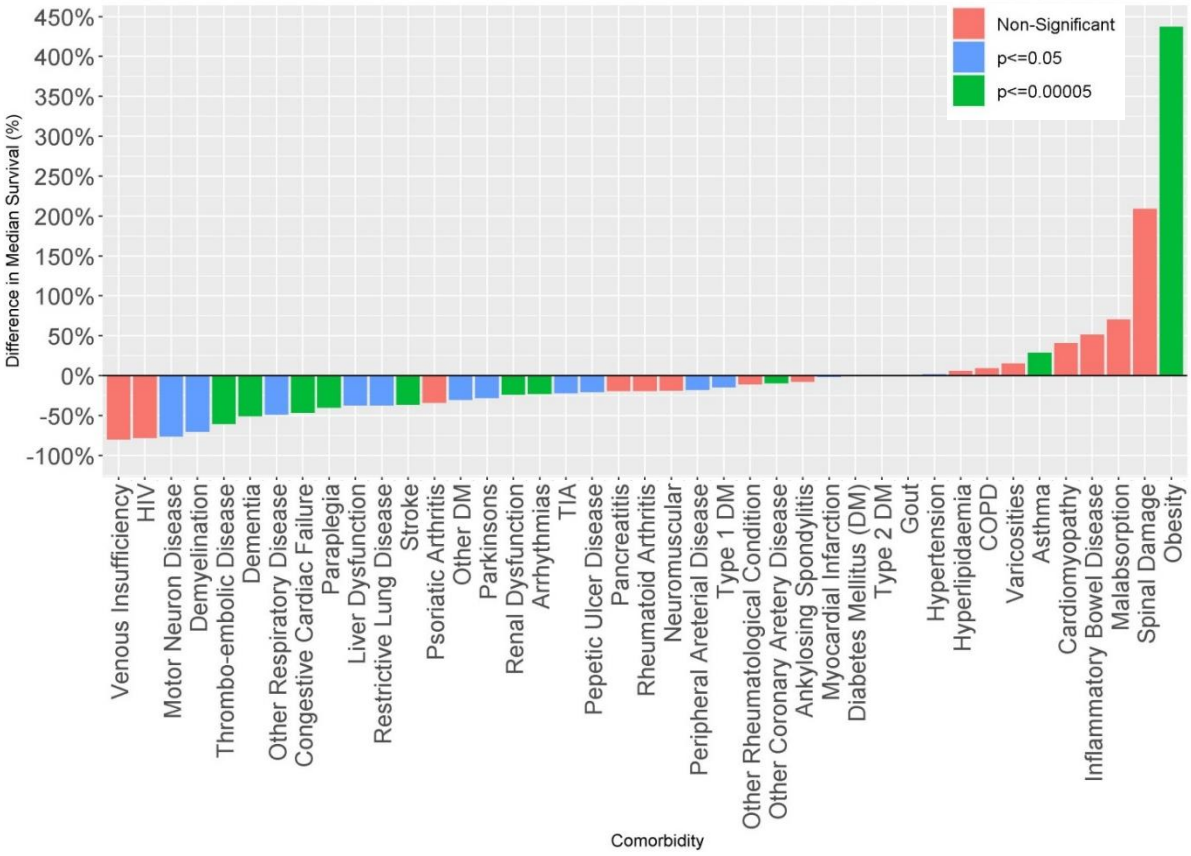


Figure 27 : Impact of Comorbidity in Lung Cancer *- Percentage difference in median survival in comorbid group compared to non-comorbid group for each comorbidity of interest in the PPM Lung Cancer Site Specific Cohort. Negative values represent a reduced survival in the comorbid group. Significance levels are demonstrated by colour*

The alternative approach to summarising the results data included analysing the impact of our three pre-specified comorbidities in all site specific cohorts. The results of the analysis for diabetes can be found in **Figure 29a**. In the case of thyroid and testicular cancer the lack of median survival estimates prevented the calculation of median survival based summary statistics. In both cases however diabetic patients has a statistically significant reduction in survival demonstrated by their stratified curves and log rank testing. Four cancers sites were found to have no statistically significant survival differences, namely; stomach, ovarian, renal, breast and laryngeal cancer. 15 cancer sites were associated with worse outcome in the diabetic group with the largest effect by percentage seen in cervical cancer representing a 20.7 year reduction in median survival. Due to the large differences seen in the median survival of cancers, it is helpful to consider the impact across different sites by time rather than percentage. This is demonstrated in **Figure 29b** which highlights than in many cases large percentage changes in median survival often represent more modest changes in time. This is demonstrated when looking at the lymphoma and breast cohorts which have similar percentage changes of -64.8% and -66.7% respectively, but very different median survival time differences of -7.4 and -11.8 years. Liver cancer demonstrates a different pattern to the other statistically significant comorbidities and is associated with a median survival improvement of 61.4% or 0.64 years (p= 0.00184).

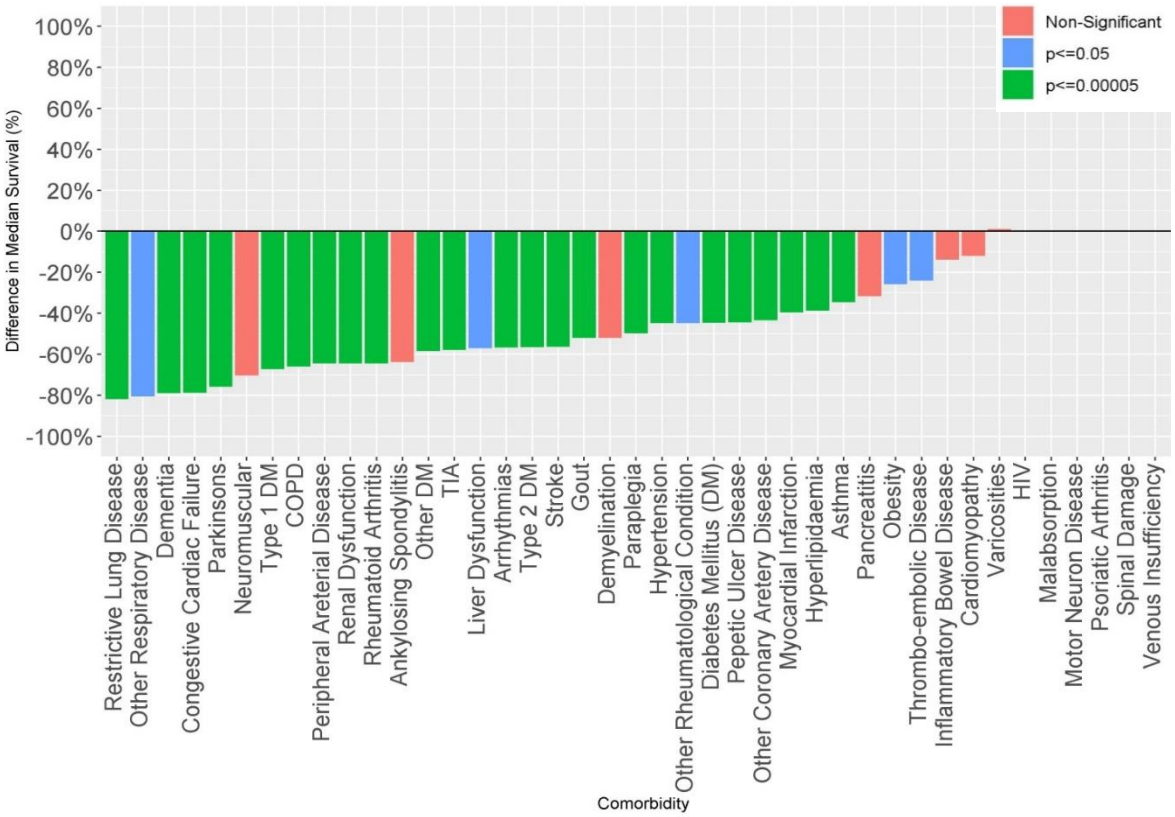


Figure 28: Impact of Comorbidity in Prostate Cancer *- Percentage difference in median survival in comorbid group compared to non-comorbid group for each comorbidity of interest in the PPM Prostate Cancer Site Specific Cohort. Negative values represent a reduced survival in the comorbid group. Significance levels are demonstrated by colour*

The analysis of patients with previous stroke demonstrates a uniform direction of effect, with all cancers being associated with decreased survival in this group (**Figure 30a**). In the case of testicular cancer there were insufficient stroke patients for an estimate to be generated. Two cancers were associated with non-significant differences, thyroid and laryngeal cancers. The largest impact by percentage was seen in lymphoma patients with a -95.5% difference ($p = 2.6 \times 10^{-19}$) closely followed by cervical cancer with a -94.4% difference ($p=1.04 \times 10^{-10}$). The smallest percentage effect was seen in pancreatic cancer with a 32% ($p= 0.008$) decrease in survival seen. When these results are transformed into time differences (**Figure 30b**) the level of effect is markedly different with laryngeal (-3.6 years) cancer showing a lesser impact of stroke when compared to cervical cancer (-20.9 years), breast (-14.1 years), melanoma (-18.2 years) and endometrial cancer (-11.5 years). Thirteen of the cancer sites do however show a decrease in survival of one year or more which is highly clinically relevant.

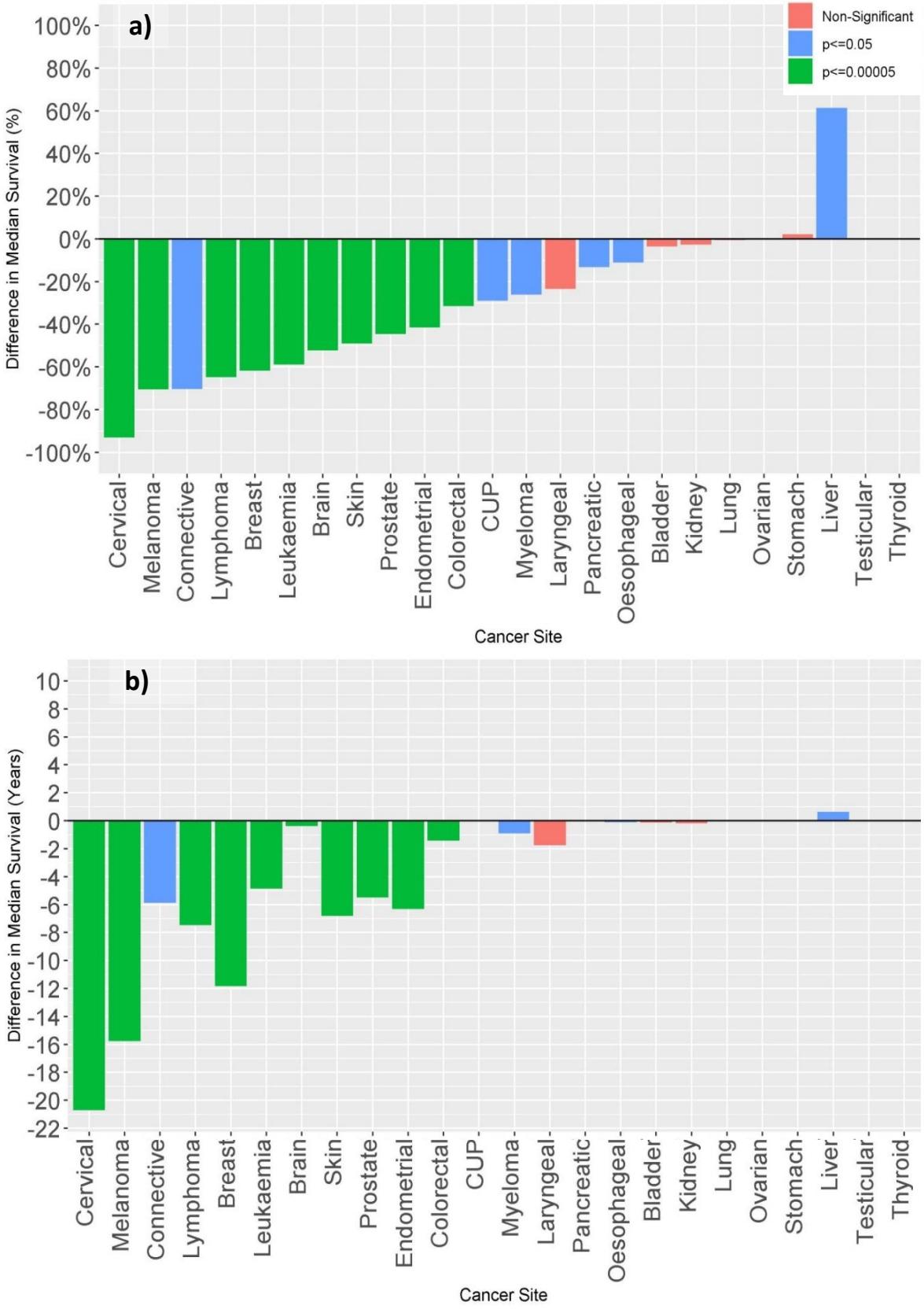


Figure 29: Association of Diabetes Mellitus to Median Survival in Site Specific Cohorts *– a) Difference in the median survival in years between patients with evidence of diabetes prior to cancer diagnosis and those without evidence of prior diabetes mellitus in each of the Cancer Site Specific Cohorts b)Difference in the median survival in percentage between patients with evidence of diabetes prior to cancer diagnosis and those without evidence of prior diabetes mellitus in each of the Cancer Site Specific Cohorts Significance levels at 0.05 and corrected for multiple comparisons are denoted by colour*

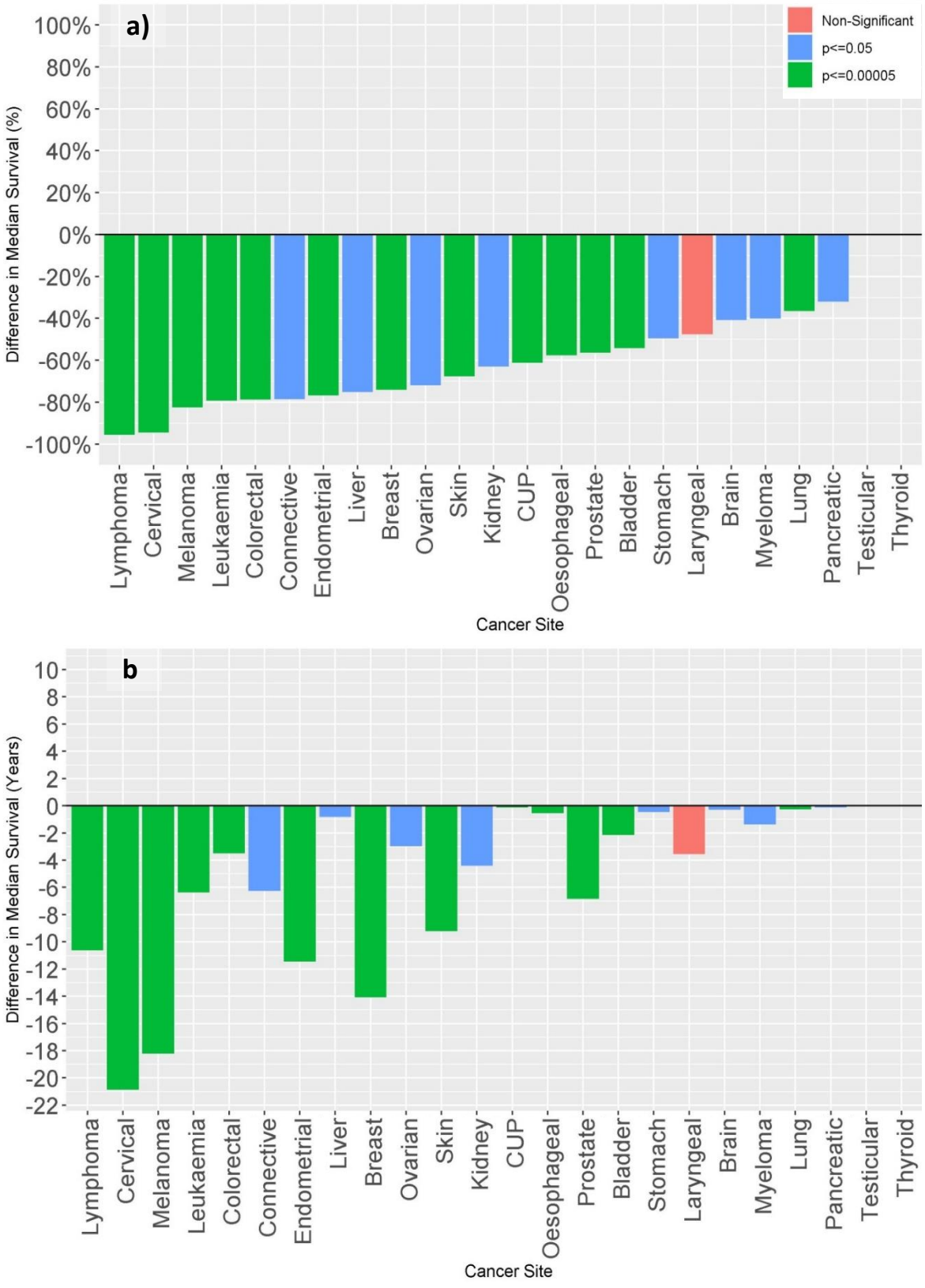


Figure 30: Association of Stroke to Median Survival in Site Specific Cohorts *– a) Difference in the median survival in years between patients with evidence of stroke prior to cancer diagnosis and those without evidence of prior stroke in each of the Cancer Site Specific Cohorts b)Difference in the median survival in percentage between patients with evidence of stroke prior to cancer diagnosis and those without evidence of prior stroke in each of the Cancer Site Specific Cohorts Significance levels at 0.05 and corrected for multiple comparisons are denoted by colour*

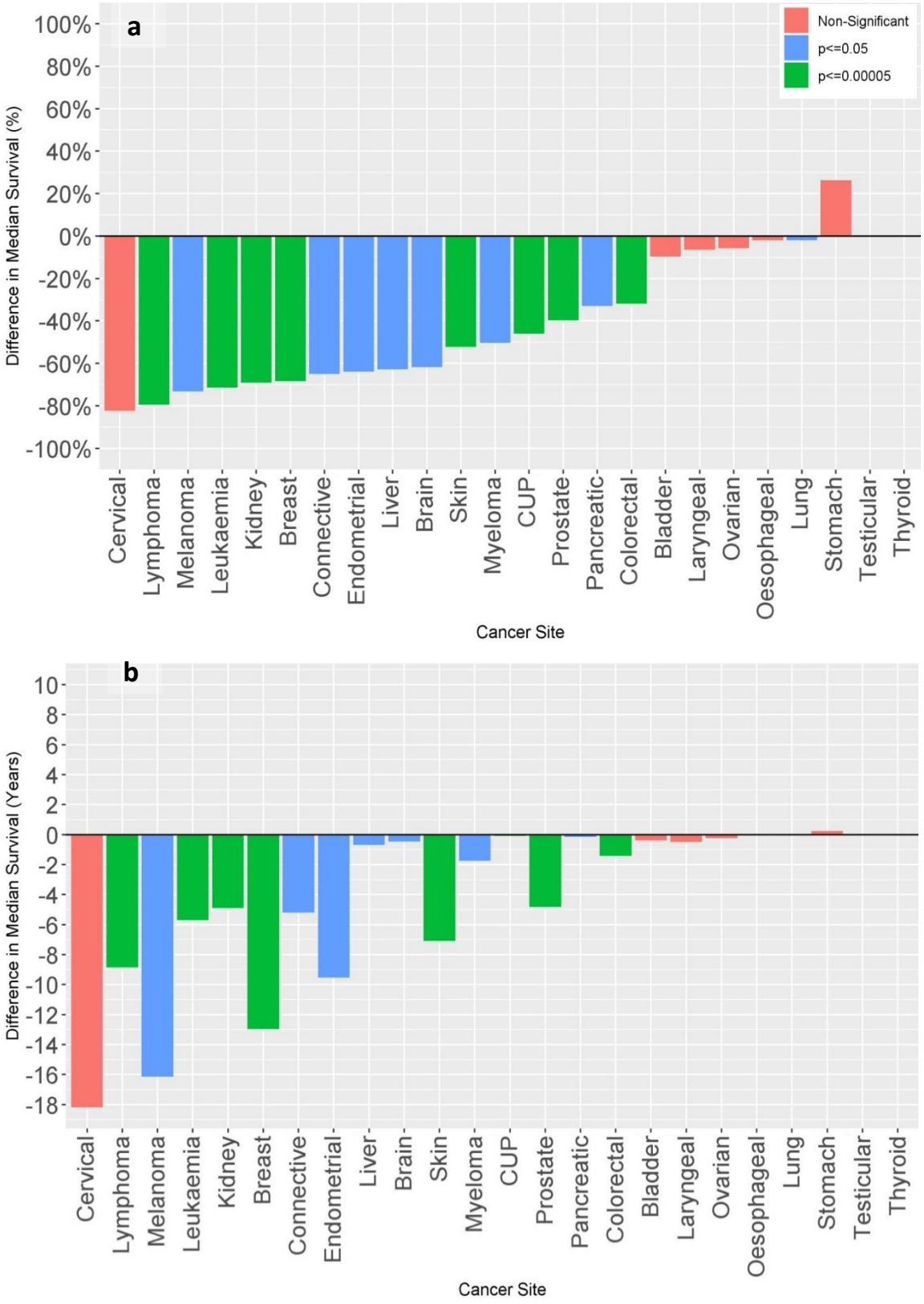


Figure 31: Association of Myocardial Infarction (MI) to Median Survival in Site Specific Cohorts *– a) Difference in the median survival in years between patients with evidence of MI prior to cancer diagnosis and those without evidence of prior MI in each of the Cancer Site Specific Cohorts b)Difference in the median survival in percentage between patients with evidence of MI prior to cancer diagnosis and those without evidence of prior MI in each of the Cancer Site Specific Cohorts Significance levels at 0.05 and corrected for multiple comparisons are denoted by colour*

The association between myocardial infarction and site specific outcomes shows only decreased survival where results are statistically significant (**Figure 31**). Gastric cancer was associated with improved survival, however this result was non-significant. Several other sites were associated with non-significant survival differences including cervical, bladder, laryngeal, ovarian and testicular cancer. As with the other comorbidity analyses, the most impactful sites by percentage and time vary, with lymphoma showing a large decrease in survival by percentage of -79.5% ($p = 9.55 \times 10^{-13}$) and an 8.5 year decrease by time. Although the percentage change is lower in several other sites the time differences seen are larger, as in the case of melanoma (-16.2 years), breast (-13.0 years) and endometrial cancer (-9.5 years). In other sites such as pancreatic cancer the large percentage difference of -32.9% ($p = 0.042$) represents a more modest difference of 54 days decrease in median survival.

4.2.3 - Identifying additional Comorbidities and Sites of Interest

	Percentage p<=0.05	Percentage p<=0.00005
Congestive Cardiac Failure	96%	83%
Arrhythmia	96%	71%
Coronary Artery Disease	96%	67%
COPD	92%	75%
Dementia	92%	71%
Renal Dysfunction	88%	67%
Stroke	88%	54%
Peripheral Arterial Disease	83%	50%
Hypertension	79%	50%
Diabetes	75%	42%

Table 8: Ten Most Impactful Comorbidities *– A stratified Kaplan Meier estimate was created for each comorbidity in each cancer site specific cohort. A log rank test was applied to each and the p value extracted. The percentage of cancer sites where each comorbidity had a p value less than the stated cut offs were calculated. The table includes the ten comorbidities with the highest percentage of meeting the p value cut offs. Values in green met the predefined threshold for further analysis those in red failed to meet the threshold.*

Using our predefined significance threshold, the comorbidity summary results indicated a significant level of impact across most cancers in the case of congestive cardiac failure (CCF) and COPD (**Table 8**). As a result these comorbidities were added to the list of comorbidities for detailed analysis. None of the cancer sites met the required threshold for further analysis. The results of these additional comorbidities are presented below.

CCF is shown to be associated with a reduced survival across all cancers with only laryngeal cancer showing a non-significant association (**Figure 32**). CCF is particularly notable compared to the other comorbidities assessed, as all of the estimated changes in median survival were of at least -50% with the exception of gastric cancr. Additionally, 6 cancer sites show an 80% or more decrease in median survival. This translates to clinically meaningful changes in survival time even in cancers with a poorer prognosis. The -74.6% ($p = 8.15 \times 10^{-18}$) change in pancreatic cancer for example represents a reduction in median survival of 126 days. Cervical cancer shows the largest change both by time and percentage with a -96.2% ($p= 9.57 \times 10^{-11}$) or 21.2 years decrease in survival. In total 8 of the cancer sites show reduced survival of over 5 years in those patients with pre-cancer CCF. This on average makes CCF the comorbidity most closely associated with poorer outcomes.

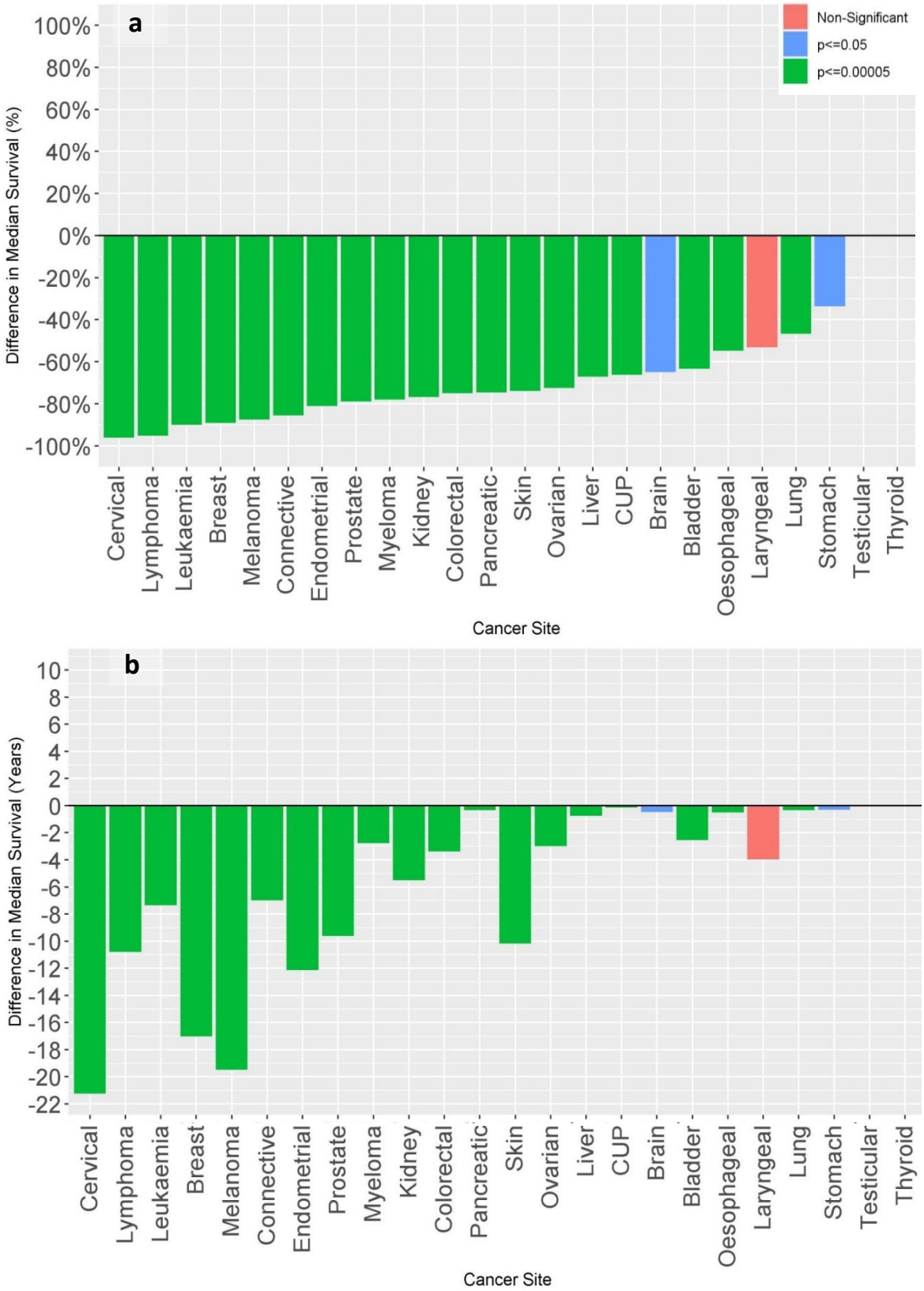


Figure 32: Association of Congestive Cardiac Failure (CCF) to Median Survival in Site Specific Cohorts *– a) Difference in the median survival in years between patients with evidence of CCF prior to cancer diagnosis and those without evidence of prior CCF in each of the Cancer Site Specific Cohorts b)Difference in the median survival in percentage between patients with evidence of CCF prior to cancer diagnosis and those without evidence of prior CCF in each of the Cancer Site Specific Cohorts Significance levels at 0.05 and corrected for multiple comparisons are denoted by colour*

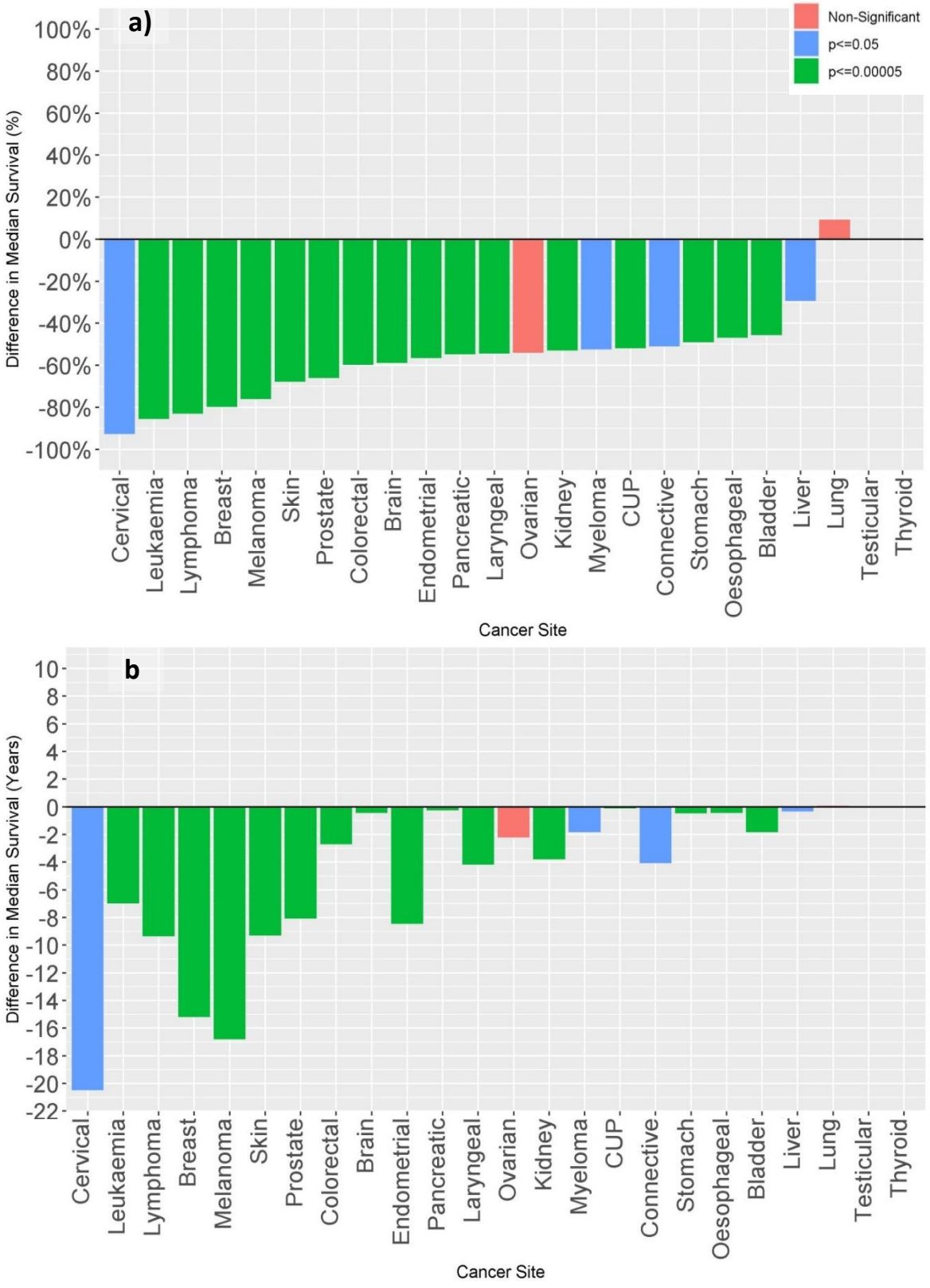


Figure 33: Association of Chronic Obstructive Pulmonary Disease (COPD) to Median Survival in Site Specific Cohorts *– a) Difference in the median survival in years between patients with evidence of COPD prior to cancer diagnosis and those without evidence of prior COPD in each of the Cancer Site Specific Cohorts b)Difference in the median survival in percentage between patients with evidence of COPD prior to cancer diagnosis and those without evidence of prior COPD in each of the Cancer Site Specific Cohorts Significance levels at 0.05 and corrected for multiple comparisons are denoted by colour*

COPD is demonstrated both in terms of time and percentage to have clinically and statistically significant associations with survival differences (**Figure 33**). Ovarian and lung cancer were associated with non-significant differences, however all other cancer sites had statistically significant decreases in survival in the comorbid group. As with CCF, COPD is associated with large scale survival differences with 19 cancers sites having a statistically significant reduction in survival of over 50% and 8 cancer sites showing a 5 year or more reduction in medial survival for those with COPD. Cervical cancer again shows the largest effect size with a -92.7% (p=0.0006) or -20.5 year change in median survival.

4.2.4 - Relationship between stratified Survival and Overall Survival

To assess the relationship between the median survival for a given cancer site and the degree of effect that a comorbidity has, we calculated the correlation of these two metrics in those comorbidity and cancer site combinations that were shown to be statistically significant. This resulted in highly statistically significant weak negative correlation between these two metrics. When looking at the results of just the diabetes analysis we can see a similar trend with significant correlation, however a number of cancer sites can be seen to vary from the overall trend including liver, primary brain tumours, leukaemia, lymphoma, renal tumours and CUP.

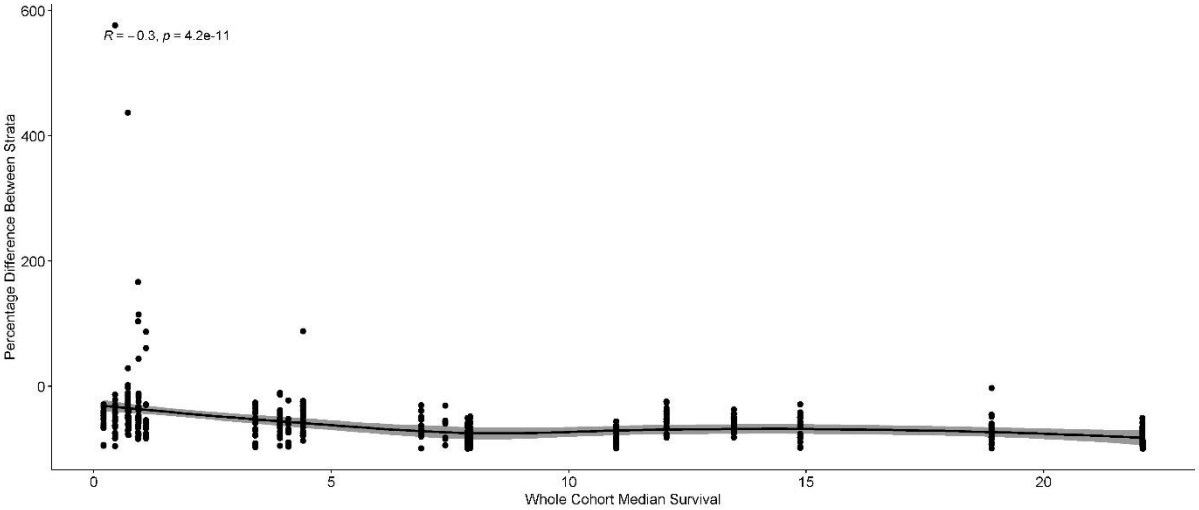


Figure 34: Correlation between Median Survival and Percentage Change in Median Survival in Comorbid Group *- Scatter plot for all the Sites Specific Cohorts and all comorbidities. The median overall survival for each cancer is plotted on the horizontal axis and the percentage difference between the comorbid and non-comorbid group is plotted on the vertical axis. Results are derived from univariate Kaplan Meier estimates. Negative numbers represent a survival disadvantage with the comorbidity. A loess curve has been fitted to demonstrate the overall trend and Pearson's correlation coefficient calculated*

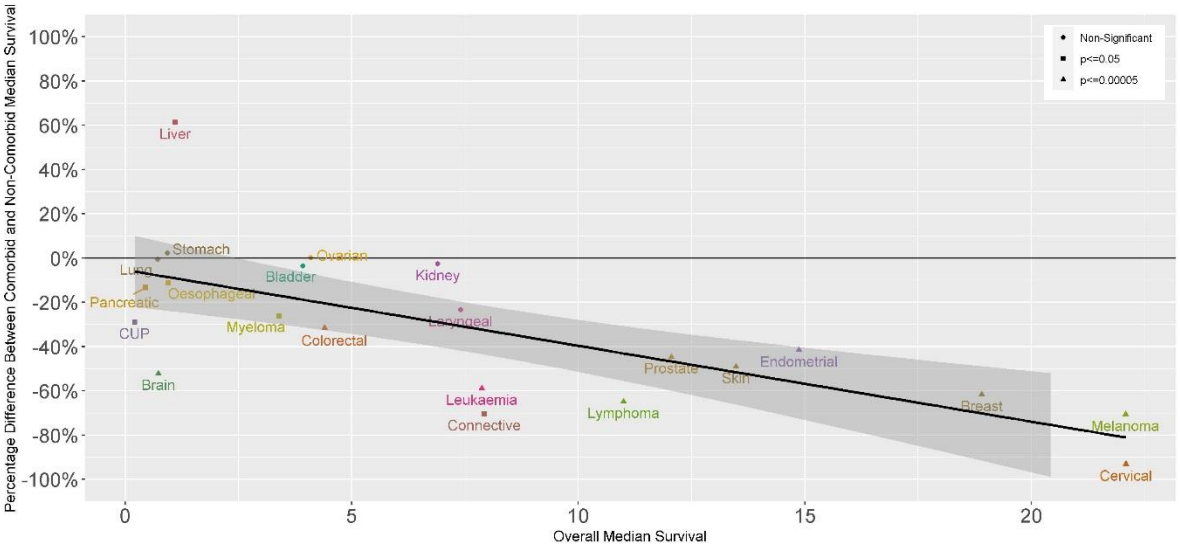


Figure 35: Correlation between Median Survival and Percentage Change in Median Survival in Diabetic Group *- Scatter plot for each Cancer Site Specific PPM cohort showing the overall median survival for that cancer site on the horizontal axis and the percentage difference in median survival between the subgroup with diabetes mellitus and the subgroup without diabetes mellitus. Negative values on the vertical axis represent a negative survival impact of diabetes. Results are derived from univariate Kaplan Meier estimates. A Line of best fit has been plotted to show the overall trend.*

4.3 - Discussion:

4.3.1 - Total Population Survival and Summary Statistics

The non-stratified, site specific analysis demonstrates the wide variation in the median survival when comparing different cancers. When comparing the LTHT site specific cohort data, to population level survival, clear differences exist. In the case of bladder cancer, the median survival locally is 57% reduced relative to the reported national average. The most likely explanation for this is in fundamental differences in the makeup of the cohorts, which introduces a form of bias.[166] The local data is derived from patients excluding in situ diagnoses, whereas, the published information does not specify whether these patients are included or excluded. If the national data does not exclude these patients it will increase the volume of patients with potentially curable disease, resulting in improved survival outcomes.[243]

There may be an additional local effect that creates a bias towards more advanced disease in the local cohort. As LTHT is the regionally centre for the delivery of radiotherapy, all patients requiring this treatment will be included in the PPM cohort. In the case of bladder cancer, radiotherapy is offered in patients not fit for surgery, or patients with more advanced disease.[244] As a result, this may increase the representation of more advanced bladder cancer and more frail patients, as compared to the national average.

Other explanations for regional variation could contribute, such as differences in the health presenting behaviours of patients locally.[245] If this were the case, it would seem unlikely that health seeking behaviour differences would vary substantially between different types of cancer. The wide variation seen in the differences between local and national data, makes this explanation seem less likely. A further explanation could be due to differences in local treatment offerings. Despite this having been highlighted as an issue in other cancer sites previously[246,247], it seems unlikely to be the

cause of the differences seen due to clinicians using approved guidelines as the framework for their clinical decision making.

When comparing other cancer sites regional and national outcomes data, the local data demonstrates favourable survival outcomes. This could be explained by the national data being less recent having been published on data from 2007. This was used as no more recent single source of UK median survival data could be identified. Recent treatment advances may therefore be present in local data but absent in the national data. The introduction of treatments such as immunotherapy over the last 5 years have improved outcomes in a number of cancers including, lung, melanoma and renal cancer.[248] The effect of improvements on summary outcomes data may vary between different cancers, not only due to the effect size in that particular cancer, but due to differences in the overall median survival. Where cancers have long survival times, improvements due to new treatments may take many years before differences in median survival become apparent across the cohort. Conversely where survival times are shorter, it will take less time for the impact of new treatments to become apparent. It would therefore be the case that improvements in treatment would be more likely to be impactful in cancers with shorted survival times and this pattern is not clearly demonstrable in the results data.

As described for the bladder cancer, population differences in the cohort composition could be influential. In some cancers the delivery of specific treatments could cause the opposite effect of that described in the bladder cancer cohort. In many cancers, early diagnosis resulting in curative treatment, is achieved via surgical intervention, radical radiotherapy and SABR.[249] Much of this is specialist care which is centralised in Leeds, thus the PPM cohort may have a bias towards early stage disease in a number of cancers such as pancreatic, lung, renal and prostate cancer.[250,251] This would inflate the local survival statistics compared to national data.

The largest survival improvement seen is in the case of leukaemia. This difference may be due to altered case mix. As LTHT is a paediatric oncology centre there will be a relatively high representation of paediatrics cases of which leukaemia is common. Typically, paediatric leukaemia outcomes are better than those in the adult population.[252] As such the differences seen may well be due to a form of selection bias.

These differences identified in the baseline survival are important to consider, as they demonstrate the large number of ways in which the local data could differ from those found in other geographical areas. The results of the subsequent analyses may therefore not be externally valid if significant differences exist in the basic population makeup of other locations. Interpretation of results beyond the local dataset should therefore be done with caution.

4.3.2 - Stratified Survival

Previous research has suggested almost universally that patients with pre-existing health conditions and cancer have a worse survival outcome than those with cancer alone.[19,136,239,240,253] Our data would mostly support this previous research showing statistically significant negative differences in survival outcomes in most comorbidity defined subgroups, in most cancer sites. Despite this, the direction and size of effect is not universal, with variation across both cancer sites and comorbidities. In the vast majority of cancers and comorbidities the results suggest that having cancer with additional comorbidity is associated with a reduction in median survival. There are additionally a small number of instances where comorbidities are associated with survival improvements, such as diabetes in liver cancer, varicosities in colorectal cancer and obesity in lung cancer. A number of possible explanations exist that could contribute to the associations seen in either direction. Initial

consideration will be given to conditions associated with decreased survival before exploring those associated with improved survival

Previous research has suggested that a significant component of survival differences seen in all-cause mortality in patients with cancer and co-morbidity is due to the independent mortality effect of the pre-existing health condition.[239] In effect, patients have two conditions which have an additive effect on the probability of death and thus they have worsened survival. The survival disadvantage is however not simply the sum of the risk of death for the two conditions due to overlapping effects. This can be understood using a thought experiment where we compare two patients with one condition each to one patient with both conditions. In scenario 1 we have a patient with diabetes who dies 4 years after index date and a second patient with breast cancer who dies 3 years after index date. In scenario 2 we have one patient with both conditions where the mortality effect is the same as for the previous pair of patients. In a population made up exclusively of patients like scenario one the risk of death would simply be the sum of the risk of cancer death plus the risk of diabetes death, plus background risk (assuming that the patient does not have other unmeasured risk). In a population made up of patients with scenario 2 then the risk would simply be the risk of cancer death plus background risk as the patient is dying of cancer before they can die of diabetes. The real world population of multi-morbid cancer patients is however made of those where sometimes they die from comorbidity and sometimes from cancer, they cannot die from both. As such, one would expect that risk of death would be greater than the risk of any one condition alone but less than the risk of both summed.

This is further complicated by the possible interaction between the conditions independent of this overlap. It may be the case that a particular comorbidity may result in direct or indirect risk differences. Many conditions have been found to increase the risk of developing cancer. As such the comorbidity may be associated with the underlying pathogenic process of oncogenic transformation. This may result in biological differences in comorbid patients compared to non-comorbid patients, which fundamentally alter how a given cancer develops, progresses and responds to treatment. Furthermore, comorbidity may impact treatment choices either by necessity or clinical judgement. Patients for example with a high comorbidity burden may not be fit for curative surgery and thus may have worse outcomes.[30,254] Additionally, existing organ damage may further prevent the use of certain other treatment options, such as autoimmune conditions limiting the use of immunotherapy[255], lung fibrosis preventing the use of lung radiotherapy and liver and renal dysfunction limiting systemic anti-cancer treatment dose.[256,257] These same effects may occur indirectly where one condition is associated with another independent of cancer.[28,42] Diabetes for example is associated with renal dysfunction and thus the impact of diabetes could be mediated via its link to renal dysfunction rather than directly. There is also the impact of clinician bias where a clinician or group of clinicians may view certain populations of patients as having different risk and therefore offer nonstandard therapy. This could therefore result in different outcomes for this group. This effect has previously been demonstrated in the elderly breast cancer population and has been used to explain a potential cause for worse breast cancer outcomes in the UK compared to the rest of Europe.[18,258]

The effect of comorbidity could also be caused by differences in the route to diagnosis. Patients with other health conditions may dismiss new symptoms as being related to their known health conditions and thus delay presenting to a medical professional.[259,260] Additionally, health professionals may assume new symptoms are related to known conditions and thus delay investigations until they have persisted for a long period of time. In these cases, presentation may be delayed with patients being diagnoses at a more advanced stage of disease. Previous research

supports this argument having shown patients with several health conditions typically present later and with more advanced disease.

These explanations however assume a worsened outcome which is not the case universally. Our data highlights three examples where comorbidity is associated with improved survival outcomes. One example of this is in liver cancer where diabetic patients have a 61.4% increase in median survival equating to 7.6 months, when compared to non-diabetic liver cancer patients. This difference raises important questions as it creates a survival paradox. There is a recognised association between diabetes and malignancies of the liver, both hepatocellular carcinoma (HCC)[261–263] and cholangiocarcinoma.[264] Patients with diabetes have twice the risk of developing HCC compared to the non-diabetic population. Diabetic patients are therefore at greater risk of liver malignancies and greater risk of dying from liver malignancy compared to the background population. Despite this, our results suggest that within the liver cancer population, diabetic patients have a reduced risk of death, thus creating a survival paradox. These paradoxical findings are seen in a number of other areas of medical research and have a number of potential explanations.[265]

One possible explanation is the effect of treatments for the comorbidity impacting on the cancer. A number of medications used in chronic disease such as aspirin[266–269] and oral hypoglycaemics[270–274] have been shown to impact on cancer cell growth and the probability of metastasis. Thus the effects of other treatments in chronic illness may provide a survival advantage.

Patients with chronic illness also engage with services differently compared to the background population. In many conditions there is a requirement for interval follow up for medication reviews, blood test screening and physical examination. This has historically formed part of GP payments via QOF.[156] As such, patient with certain long term health conditions such as diabetes may have health problems in some instances picked up at an earlier stage than the background population. This could lead to a true survival benefit as cases may be diagnosed at an earlier and therefore more treatable time yielding improved remission rates. Alternatively, it may be the case that the advantage is in fact due to lead time bias[275] where the outcomes are the same and identified earlier giving the appearance of a survival advantage.

The underlying physiology of patients with comorbidities is also commonly altered. This is often part of the link between the co-development of multiple conditions in the same patient.[42] This altered physiology may result in different forms of cancer developing or behaving differently once developed, providing an advantage in some instances.

A further possibility is that these differences are entirely artefactual due to bias in the data. If by assessing a particular cancer site, selection bias[166] is introduced, such that the non-comorbid group contains a large representation of patients with some particularly harmful characteristic, such as extreme age, then when comparing the outcomes, the comorbid group may appear to have an advantage. If however this was overcome with adjustment or stratification it may identify that the comorbidity is still disadvantageous when accounting for this confounding or selection bias.[165]

As discussed above, with the univariable approach employed through the KM estimator, the analysis describes the pattern of survival in the two groups. As no adjustment is made for other forms of bias and confounding, detailed interpretation of whether the pattern is a fair reflection of an inferred association in the wider population is not possible. Further analysis is therefore needed to identify if these survival patterns continue to be present, once appropriate conditioning on multiple variables has taken place. A condition by condition description of the literature and possible biological

mechanisms of the patterns of association will be provided in later chapters, where these associations are shown to persist after adjustment. These can be found in chapter 5.

Another potential bias, reverse causation, might explain the patterns identified within the stratified survival curves. This occurs when an end point or intermediary is related to the exposure of interest in a causal way but where the direction of causation is the opposite of that assumed by the analysis. An example of this could be in the case of the results presented for obesity which in some instances are shown to be associated with better outcomes. Here it is plausible that the most aggressive cancers are causing significant weight loss and cachexia. If this were true then the presence of an aggressive cancer would be causally associated with an increased probability of being in the non-obese category. Analysis results would therefore suggest obesity is protective when in fact it is simply a marker of a less severe cancer which is increasing the likelihood of patients remaining in the obese category. This theory could be investigated in future by focussing analysis on patients with serial weight measurements and looking at as a variable for specific investigation. This approach has been successfully employed previously when investigating the effect of blood pressure lowering treatments on survival.[276] Previous literature focussing specifically on reverse causality and illness related weight loss has however suggested that this form of bias is minimal[277] and has therefore not been an area focus for the work presented in the later chapters. It could however represent an area of future study and research.

4.2.3 - Identifying Additional Comorbidities and Sites of Interest

Comparisons between median survival outcomes for two of the cancer groups, namely Thyroid and Testicular were not possible due to the high survival rates preventing accurate estimation of median survival. Despite this, it was possible to demonstrate statistically significant differences between survival outcomes when conditioning on individual comorbidities in these cancer populations. When assessing the identified survival patterns of comorbidity across the cancer sites that had not been pre-selected, none of them met the pre-specified thresholds for further analysis. This may be due to the smaller numbers of cases yielding higher p values particularly in the context of rarer comorbidities. The presence of low patient numbers in several of the pre-existing conditions impairs the ability for drawing meaningful conclusions and warrants further investigation using cohorts with greater co-morbid representation.

When assessing sites by comorbidity, two conditions were found to meet our pre-specified thresholds for further investigation, namely, COPD and CCF. The groups defined by these conditions were demonstrated to have not only consistent effects, but also large scale survival disadvantages in most cases. As described in section 4.2.2, although a number of possible explanations could underlie the survival patterns observed, more meaningful discussion will be possible if these relationships continue to persist after adjustment. Thus, further detailed discussions can be found in chapter 5 where relevant.

4.2.4 - Relationship between Stratified Survival and Overall Survival

Several previous studies when presented with the results showing an increased risk of death in patients with cancer plus comorbidity, have attempted to compare the relationship between survival time for a given cancer overall and the impact of a given comorbidity.[278–282] The rationale applied, is that in cancers where patients live longer there is a greater amount of time for the mortality effect of the comorbidity to exert its influence and thus one would expect to see negative correlation between overall population survival and comorbid population survival. The above analysis attempted to mimic this by comparing the site specific cohort median survival and percentage

change in median survival. Our results agree with those found in the previous literature, which suggest the arguments outline above seem more plausible. Although the level of correlation found was modest at -0.3 the significance levels show a very small p value. There is however a significant methodological flaw in both our analysis and those previously published which compromise their utility.

The methodological issue arises as a result of mathematical coupling.[283,284] In both of the estimates being plotted the values are a product of one another. The total population survival is made up of the non-comorbid and comorbid populations. The comparison of the median survival is the division of one subset of the same population divided by the other. In effect the difference and the overall survival are different mathematical representations of the same information. As a result, they are mathematically coupled. This leads to the appearance of correlation even if none exists in what is termed spurious correlation.[285] The linear regression applied to the data therefore likely represents the pattern of coupling, rather than a true association. This does however offer the opportunity to derive insight not from the association, but those that deviate from it. In this case the regression line can be thought of as the null, that is, the expected pattern. Those that do not conform to this may be demonstrating effects outside of the mathematical coupling. **Figure 35** identifies a number of examples which fall outside of the 95% confidence interval for this null association. These cancer diagnoses may be important targets for further investigation as there is some effect beyond the association's inherent to the methodology applied.

4.2.5 - Limitations:

In order to understand and interpret the findings presented above it is important to consider the limitations of them stemming from the underlying data, methods applied and context of the analysis. Some of these have be briefly mentioned in the above sections however are covered in more detail below.

Data Limitations:

The identification of comorbidity is based on the clinical coding data captured by the hospital trust and additionally the HbA1c blood test results in the case of diabetes. As detailed in chapter 2 in order for a patient to be coded as having a co-morbidity the patient has to have had an inpatient admission within Leeds Teaching Hospitals NHS Trust. Patients who are admitted should be assigned both an ICD-10 code for the cause of their admission and ICD-10 codes for all their pre-existing health conditions. This coding is not conducted by the clinicians, but by a dedicated team of clinical coders. Patients erroneously stating they have a diagnosis or clinicians incorrectly or unclearly documenting information, may result in a condition being added to a patient's coding data that they do not have. Clinical coders, although trained, are not healthcare workers and so may misclassify or code conditions. As clinical coders assessing subsequent admissions will automatically refer to previous clinical coding data, or the same previous clinical records, errors will be propagated forwards such that once an incorrect diagnostic label has been applied, it is likely to be repeated on subsequent admissions. Errors of omission may occur as patients who have an existing health condition will only have this coded if it is recorded within their clinical records. Where clinicians fail to capture this, then so too will the resulting clinical coding. This has been demonstrated in previous research.[219,221,286,287]

As demonstrated by our DM coding analysis in chapter 3, missing diagnostic codes may occur if a patient is not admitted to hospital. As a result, patients that have conditions managed in the community who never attend hospital with an acute or chronic issue requiring admission, will not have their diagnostic information captured by the hospital. Patients may attend other hospitals

either through their geographical location, such as those close to another hospital, or through personal choice such as those using private care. Patient data from these alternative care providers will be unavailable and thus may lead to the omission of further diagnostic codes. These factors may introduce systematic misclassification bias to the analysis, such that those within a given geographical area may be more or less likely to have clinical codes in their hospital records or those from higher socio-economic groups may be less likely to have clinical codes due to the use of private care. Additionally, those patients with overall poorer health or worse controlled comorbidity are more likely to be coded as they are more likely to require admission.

A further limitation stems from the clinical coding data capturing only the presence of a diagnostic label at a specified time point[127]. Thus it is impossible to distinguish between longstanding and recent diagnoses from this coding. This becomes a particular issue when combined with clinical coding only taking place for admissions to hospital and not outpatient contact. Cancer patients with pre-existing health conditions managed as an outpatient may only have their diagnosis coded at an admission after their cancer diagnosis event. These patients will be erroneously labelled as having no evidence of their pre-existing health condition prior to cancer, despite the fact that it is long standing.

These various limitations to the coding data result in the potential for the introduction of misclassification bias and selection bias.[166,218] The coded population within the hospital record may represent a greater severity of a particular condition, lower socio-economic status or greater level of frailty, and results may not be externally valid when applied to the wider population with cancer and the comorbidity of interest.

Methodological Limitations

The Kaplan Meier method relies on a univariate approach when stratifying survival outcomes. In situations where there is underlying confounding then a difference may be identified as being associated with a particular condition, when this is in fact not the true cause.[165] If for example patients with diabetes are more likely to be older, then an identified association between diabetes and worse cancer outcome might simply be a reflection that older patients have worse outcomes and diabetes is a surrogate marker for age. This issue is more significant if trying to draw a causal link between and exposure (a prior-health condition) and an outcome (survival). Within the context of trying to identify at risk groups based on identifiable baseline characteristics this issue may be less important. If we however want to quantify the effect of an association a multivariate approach that adjusts for other potential confounders would potentially result in a more accurate approximation of the true effect size.

The use of censoring within any analysis approach introduces some assumptions.[175,176] First it assumes that those patients who are censored are equally at risk as those that are not censored. This assumption cannot be easily assessed. This issue is more likely to be impact full when an event is used as a censoring date. An example would include an analysis of survival in liver failure where a transplant is used as a censoring date. This bias is less likely to be an issue within this analysis as censoring is applied to those with no further clinical records or those who have yet to complete the 15 year analysis period. It seems unlikely that loss to follow up is a surrogate marker for a variable associated with altered survival risk and this assumption is therefore likely to be valid in this instance. A second assumption is that patients diagnosed at all time points have an equal risk. This could be assessed by comparing survival curves at various time-points to identify differences. The literature relating to survival outcomes over the past 15 years have for the most part demonstrated improvements in survival outcomes over time.[241] It is therefore likely that this assumption has been

violated and alternative methods of adjusting for year of diagnosis with alternative methods may help mitigate this issue.

Context Based Limitations

The use of overall survival as the end point of interest is potentially problematic. Each of the pre-existing health conditions assessed will have a mortality effect independent of cancer. Thus patients with a two conditions with a mortality effect, having worse survival outcomes than those with only one, is not overly surprising. Differences between particular conditions and particular cancers may still be of clinical and academic interest. The introduction of a control cohort with the prior-diagnosis but no cancer would improve the information yielded by the analysis.[216] As the cancer and non-cancer related outcomes work as a competing risk in those patients with both a cancer and a comorbidity, combining the independent survival risk from the cancer only cohort to the survival risk from the prior-diagnosis may not represent a valid baseline risk comparator. Further analysis that estimates the impact on the cancer related outcomes and non-cancer related outcome may however be used to overcome this issue and will be address is subsequent chapters.[179]

Thresholds for Comorbidities of Interest and Cancers of Interest

Within our analysis we have highlighted three comorbidities and four cancers in which to focus our analysis. The choice of comorbidity was based on the results of the clinical coding accuracy analysis that was undertaken in chapter three. This highlighted potential issues around misclassification error in clinical coding data which may introduce significant bias into the results of analyses. In the case of diabetes mellitus we demonstrated how the inclusion of HbA1c blood results can at least partially mitigate this issue. Diabetes mellitus was therefore selected as a key comorbidity of interest. In order to select other comorbidities, clinical domain expertise was applied to identify conditions, that when they occur, typically result in a hospital admission and require hospital testing in order to obtain the diagnosis. Two conditions were felt to meet these criteria and they were myocardial infarction and stroke. By selecting these, the nature of the conditions diagnostic pathway increases the likelihood of hospital clinical coding and therefore reduces the chances of misclassification bias. In each of these conditions all cancer sites were analysed for their association with survival outcomes. Despite the domain expertise guided rationale applied here there is still an inherent risk of misclassification bias still persisting even if it is to a lesser extent.

In order to determine the impact of a broad range of conditions the analysis of each comorbidity was undertaken in each of the top four cancer sites of interest. This was conducted as due to their common nature, the information on the impact of comorbidity in these cancers has the broadest potential clinical use. In addition, the relatively large number of cases of each cancer increased the likelihood of obtaining sufficient patient numbers with less common comorbidities. Beyond simply allow for analysis to take place, it maximises the likelihood of higher precision estimates being obtained from any modelling strategies applied.

Beyond our preselected cancer and comorbidities we specified some pre-analysis thresholds to identify other comorbidities or cancers that may be of particular interest. The rule thresholds were however entirely arbitrary and not meeting the threshold does not suggest a lack of important findings in other comorbidities or cancer sites. Additionally the focus on p value thresholds creates a potential bias towards conditions or cancer sites with larger numbers as with increasing numbers of patients there is a higher likelihood of small p values even in the context of minimal differences between the compared groups. The thresholds applied were as a result of clinical opinions with the decision taken to focus additional groups beyond those that were preselected to those with the

most significant effect sizes on average. Many other cancers and comorbidities did however show evidence of statistically significant differences and thus future research is needed to explore these in greater detail in future.

4.4 - Summary

These exploratory survival analyses have identified a number of ways in which the outcomes of the PPM dataset are different from those quoted nationally. This suggests possible differences in case mix and treatment delivery, which may limit the applicability of the research findings to other settings. Our analysis highlights large numbers of cancers and comorbidities where statistically and clinically significant survival differences occur. The majority of these identify that patients with comorbidity are estimated to have worse outcomes than those without comorbidity. Despite this general trend, three examples of comorbid patient groups having improved survival estimates are demonstrated. Our assessment of all comorbidities and all cancers has highlighted particularly marked survival differences on average in COPD and CCF, thus these conditions have been selected for further more detailed analysis within both this and subsequent chapters. The analysis of the relationship between the prognosis of a given cancer and the effect size of comorbidity highlights issues with previous conclusions drawn on the basis of spurious correlation. The results do however present an alternative approach to using this data, with utility for identifying conditions with survival associations that appear to be particularly impactful. Despite the wide range of findings presented, a number of limitations have been discussed including ones based on the methods employed. In order to overcome some of these, subsequent analysis using a multivariable approach aiming to deal with potential cofounding will be employed and discussed in the next chapter. Additionally, the principles set out in chapter two for inferential analysis will be applied, moving away from significance tests and placing a greater emphasis of precision of estimates, consistency of effect direction and clinical importance.

Chapter 5: Multivariable Approach to Analysing the Relationship between Comorbidity and Cancer Survival Outcomes

5.0 - Introduction

Thus far the focus has been on descriptive and exploratory analysis of the data. In doing so, no adjustment within the analysis has been undertaken to deal with potential confounding. This has prevented the ability for analysis output and the relationships identified, to be used to make inferences beyond the study population. Work in chapter 3 highlights population level differences in gender, deprivation and age, and discusses how this may be important in interpreting the findings of the results seen in chapter 4. This chapter builds on the previous analysis, by employing multivariable survival modelling, using the Cox proportional hazard (Cox PH) method for time to event analysis.[108] This allows for specific adjustment for potential confounders, allowing a greater level of inference beyond the PPM cancer cohort. Despite the multivariable approach used, this modelling does still not represent an attempt at causal modelling.

In view of the inferential nature of this analysis, a stricter framework for interpretation has been employed to consider on-going bias, effect direction, effect size and precision. More details on this can be found in chapter 2. Further discussion is also provided to describe not only previous research findings in comorbidity and cancer, but also in research that might be relevant to potential mechanisms that might underlie any differences identified. The output of this chapter aims to provide insight, which might be more generally applicable in cancer populations outside of those recorded in PPM.

5.0.1 – Aims and Objectives

Aims

1. Quantify the association between comorbidity and all-cause mortality in cancer patients using a multivariable approach.

Objectives

a) Develop Cox PH models for each combination of cancer and comorbidity whilst adjusting for age, gender and deprivation as appropriate.
b) Test for violations of underlying model assumptions.
c) Identify outlier or influential patients within the dataset and quantify their impact on model estimates.
d) Quantify hazard associated with comorbidity.
e) Assess the confidence interval for precision.
f) Assess confidence interval for high or low probability of unidirectional hazard.

5.1 - Methods

5.1.1 - Building Survival Models

Survival models were built using the Cox proportional hazard approach to time to event analysis.[108] Survival was estimated as a function of the individual comorbidity of interest, age (in 10 year bands), gender and deprivation. Data on grade and stage was not included due to the high proportion of missing data as outlined in chapter 3 and due to their potential as mediators of effect. Histology was not included as per the confounder principles outlined in chapter 2. Models were built for all comorbidities of interest in the breast, lung, prostate and colorectal cancer site-specific cohorts. All site specific cohorts were assessed for each of the key comorbidities of interest namely, diabetes, MI, stroke, CCF and COPD. The focus of analysis is on the impact of comorbidity and thus results are predominantly presented based on the coefficients of each individual comorbidity. The inclusion of age, gender and deprivation was to control for confounding rather than for direct assessment of effect. The covariate model specification for these analyses are detailed in **Table 25** of the appendix. Models were built using the R "survival" package. Two model versions were built one without stratification and one with stratification by comorbidity, to enable both the extraction of hazard ratios and also the plotting of stratified survival curves.

5.2.2 - Analysis of Model Assumptions

In order to assess for potential violations of underlying Cox proportional hazards model assumptions, a number of approaches were applied. Analysis was undertaken using Schoenfeld residuals to identify evidence of violations of the proportional hazards assumption.[288] Due to the large cohort numbers of over 10,000 patients this was done visually, plotting residuals over time for each covariate rather than relying on p values (more details can be found in chapter 2 and the discussion section below).

To assess for violations of the linear assumptions Martingale residuals were estimated and plotted against the value of the covariate of interest. This was done using the raw values of the covariate along with this value after a logarithmic, square root transformation and quadratic transformation. Results were assessed visually to identify patterns of nonlinearity and whether scale transformations could overcome this.

Assessment of the effect of outliers on estimates was conducted through the numerical assessment and visualisation of calculated dfbeta values for each observation. A cohort size adjusted threshold was applied using a cut off of 2/sqrt(n) to identify important and influential outliers.[289] To assess the impact of these Cox models were generated with and without outliers with the resulting hazard ratio estimates for the comorbidity of interest compared. Due to the large numbers of models being analysed this outlier exclusion approach was implemented in only the five key comorbidities across the top four cancers as example cases.

5.2.3 - Extraction of Summary Statistics

In order to assess the relationship between comorbidity and survival, hazard ratios were extracted from the completed Cox models by taking the exponent of the coefficients. As the analysis is reliant on observational data and large case numbers, p values were not analysed.[59,61] Instead the confidence intervals were assessed and reported for precision. Results were then categorised as having a high probability of unidirectional hazard where the confidence interval did not include 1 or a low probability of unidirectional hazard, where it did include one. This is based on the concept that where a confidence interval does not include 1 then in 95% of realisations, the population value of the hazard ratio is the same direction of effect as the point estimate, that is, it is unidirectional.

Where the confidence interval includes 1, or includes values both above and below 1, 95% of realisations have a population hazard estimate that suggests the same direction of effect, opposite direction of effect or no effect, thus reducing the utility of the estimate, due to the lower probability of having the same direction of effect as the point estimate.

The span of the confidence interval for each estimate was assessed in terms of absolute span and the percentage of point estimates. Results were filtered to show the most precise comorbidity models by limiting results to those where the confidence interval span was less than or equal to 0.25 and where the confidence interval span represented less than or equal to 25% of the point estimate. This allows for the identification of results which are not only unidirectional but also have high levels of precision.

5.3 - Results

5.3.1 - Survival Models

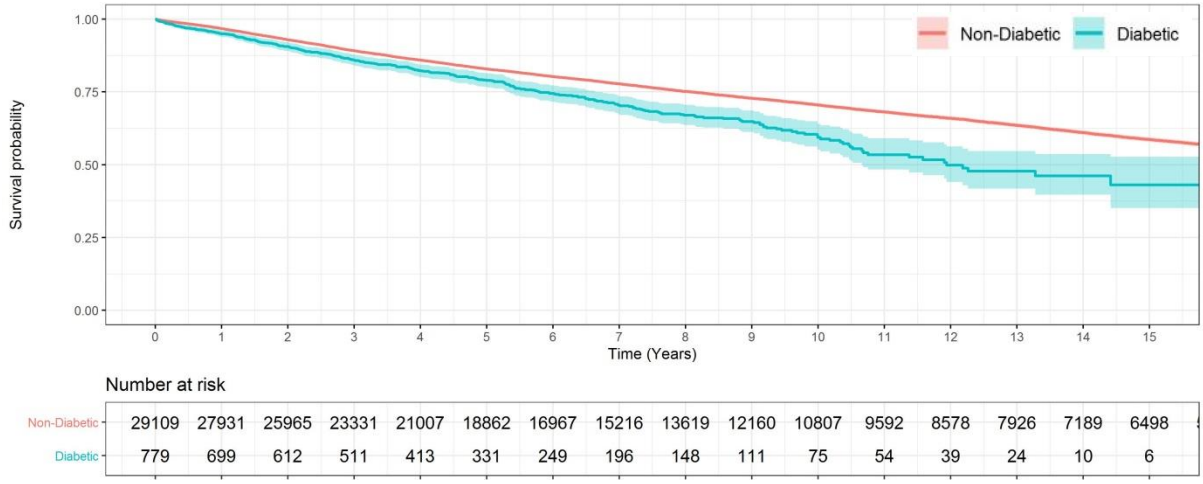


Figure 36: Association between Diabetes and Survival in Breast Cancer *- Cox Proportional Hazard Model for Breast Cancer Site Specific Cohort Stratified by Diabetic Status. Model includes adjustment for age, gender and deprivation.*

As in chapter 4 for the KM approach, individual stratified survival curves were produced for each of the comorbidity and site comparisons undertaken. This equates to 295 comparisons and an equal number of stratified curves. This approach to reviewing the data would prove inefficient and was therefore used only as a sense check in order to identify potential issues with data pre-processing. The detailed analysis and results presented are based on the summary statistics found in the later section 5.3.3

5.3.2 - Cox Model Diagnostics

Proportional Hazards

Individual plots for each of the models produced were created showing Schoenfeld residuals against time for each covariate. On manually review, none of the models were identified as having evidence of a meaningful violation of the proportional hazards assumption. Although statistically significant assessments were found according to p values, these were disregarded due to the context of high population numbers. The absence of a violation of assumptions is demonstrated by the overall trend being a straight line at or close to zero. Although some fluctuation is seen over time this is only small in size and therefore not deemed to be important for the interpretation of model results. An example of these diagnostics plots is included below for reference (**Figure 37**).

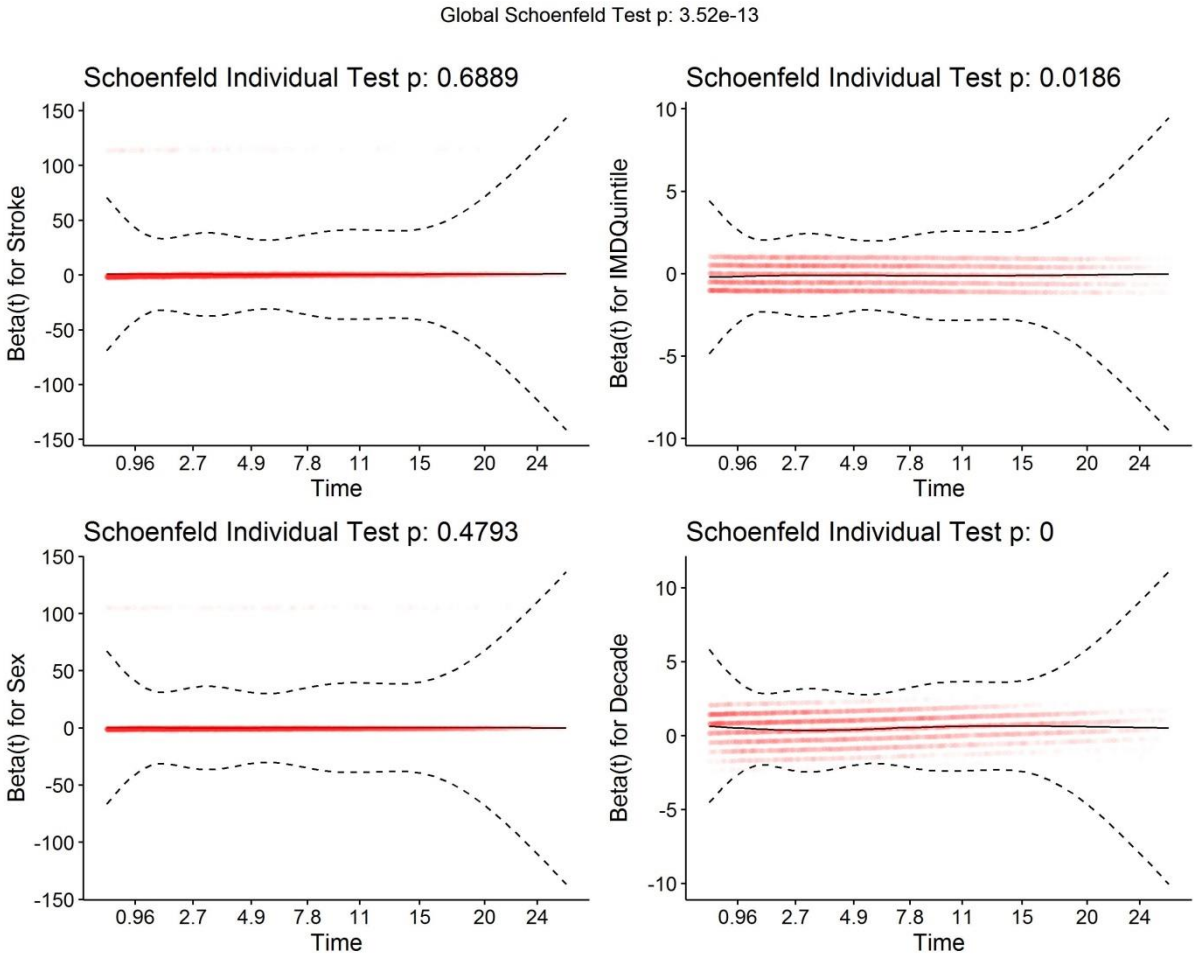


Figure 37: Shoenfeld Residual Plot for Covariates in the Stroke in Breast Cancer Cox Model – *The Shoenfeld residuals for each variable are plotted on the vertical axis and time is plotted on the horizontal axis. Each red dot represents one patient at that time within the Breast Cancer site specific cohort. The black line is the line of best fit and the dotted lines the confidence interval. When the black line is horizontal and almost entirely straight there is no evidence for a violation of the proportional hazard assumption. If the line is angled or showing multiple areas where the line is curved this provides evidence for a breach of the proportional hazard assumption.*

Linear Assumptions

The assessment of the linear assumptions of non-categorical variables highlighted patterns suggestive of nonlinearity in multiple models. This was most marked in terms of age in decades particularly in breast cancer models, but also in prostate and colorectal cancer as well. IMD quintile also showed non-linear relationships, although to a lesser extent. Applying the alternative square root and logarithmic scale transformation did not resolve the nonlinearity in any of the cases. A summary of all models can be found in the **Table 9**. **Figure 38** shows an example of the appearance of non-linear relationships using this approach. Sequential quadratic transformation of age was attempted to overcome the issue of non-linearity up to 3 degrees. Repeated visual assessment of Martingale residual suggested ongoing non-linearity. Increasing numbers of covariates with each degree of quadratic adjustment also negatively impacted on the precision of the obtained effect estimates.

Site	Age	IMD
All	0	0
Bladder	0	0
Breast	1	0
Cervical	0	0
Colorectal	1	0
Connective	1	0
CUO	0	0
Endometrial	0	1
Intracranial	0	1
Kidney	0	0
Renal	0	0
Laryngeal	0	0
Leukaemia	0	0
Liver	0	1
Lung	0	0
Lymphoma	0	1
Melanoma	0	0
Myeloma	0	0
Oesophageal	1	1
Ovarian	0	1
Pancreatic	0	0
Prostate	0	0
Skin	0	0
Stomach	0	0
Testicular	1	1
Thyroid	1	1

Table 9: Summary of Linear Assumptions Assessment – *The plots for the Martingale residuals for each continuous variable in each site specific cohort was reviewed for evidence of a violation of the linear assumptions. The results are recorded win the table with 0 = No evidence of Violation, 1 = Evidence of Violation*

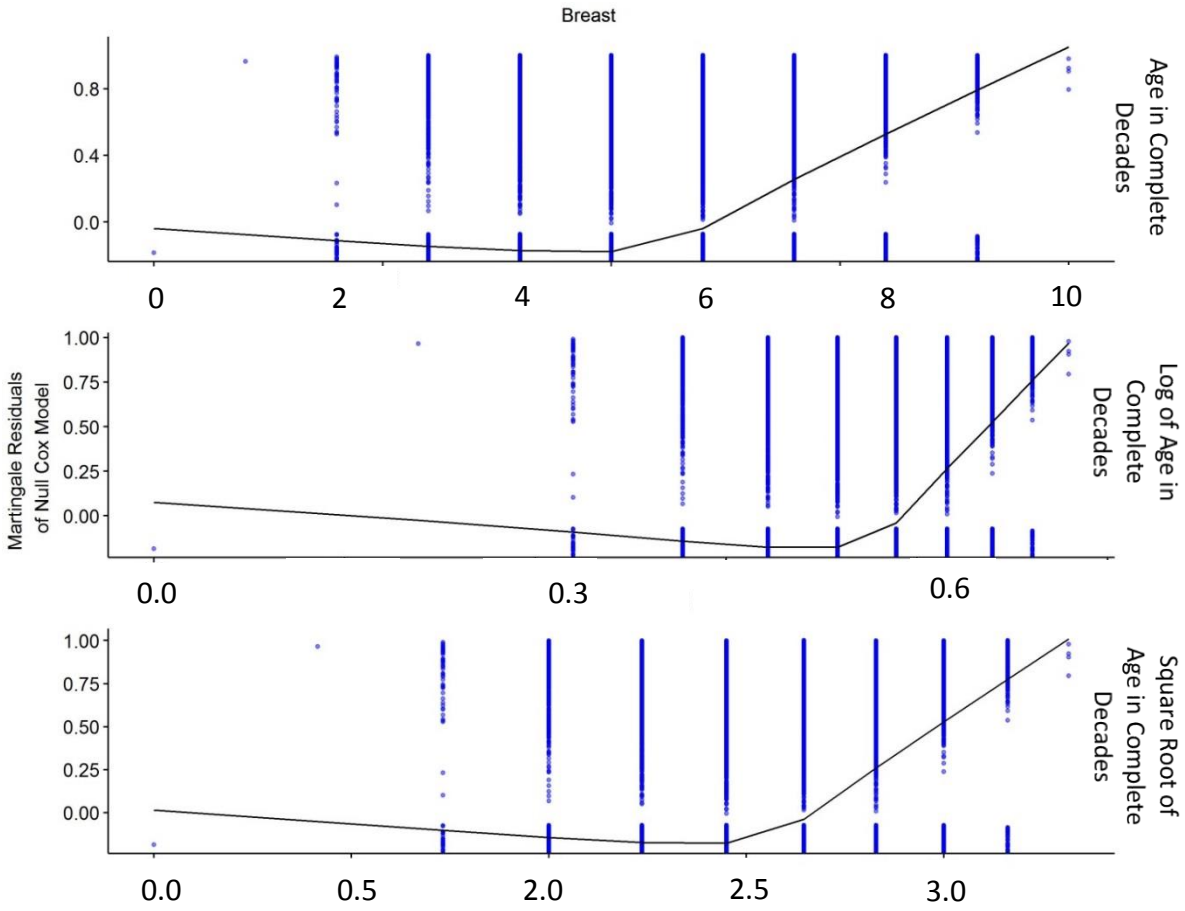


Figure 38: Assessment of Linear Assumptions of Age in Breast Cancer Site Specific Cohort*. = Martingale residuals are plotted on the vertical axis against age on the horizontal axis. Three plots represent age on different scale, the top plot represents age as completed decades in effect representing ten year age bands. The second plot represents age as the log of age in complete decades and the third plot represents age as the square root of age in complete decades. Where linear assumptions hold the line should be a continuous straight line. Curves or changes in the trajectory of the line suggest non-linearity.*

Influential Outliers

Analysis of the dfbeta values was conducted both visually and numerically. Numerical assessment identified that there were no examples of influential outliers for IMD and age. Influential outliers for gender were only identified in the context of breast cancer. The majority of influential outlier values were for individual comorbidity assessment. This was most commonly the case for comorbidities with low numbers within the cohort. In many of these instances the influential outliers were the majority of the cases of the comorbid patients. In such a situation it suggests that the variation between comorbid cases was high. When assessing the 4 key cancers and the 5 key comorbidities, no influential outliers were found for lung cancer or colorectal cancer models. In breast cancer and prostate cancer, the diabetes models were also identified as having no influential outliers. The results of the models with and without outlier values for the remaining 4 comorbidities in breast and prostate cancer are shown in **Figure 39**. In some cases, such as CCF and MI in breast cancer and CCF, and stroke in prostate cancer, the point estimate without outliers shifts beyond the confidence interval of the full cohort estimate. In none of the cases does the direction of effect change or does the hazard ratio move to a low probability of unidirectional hazard.

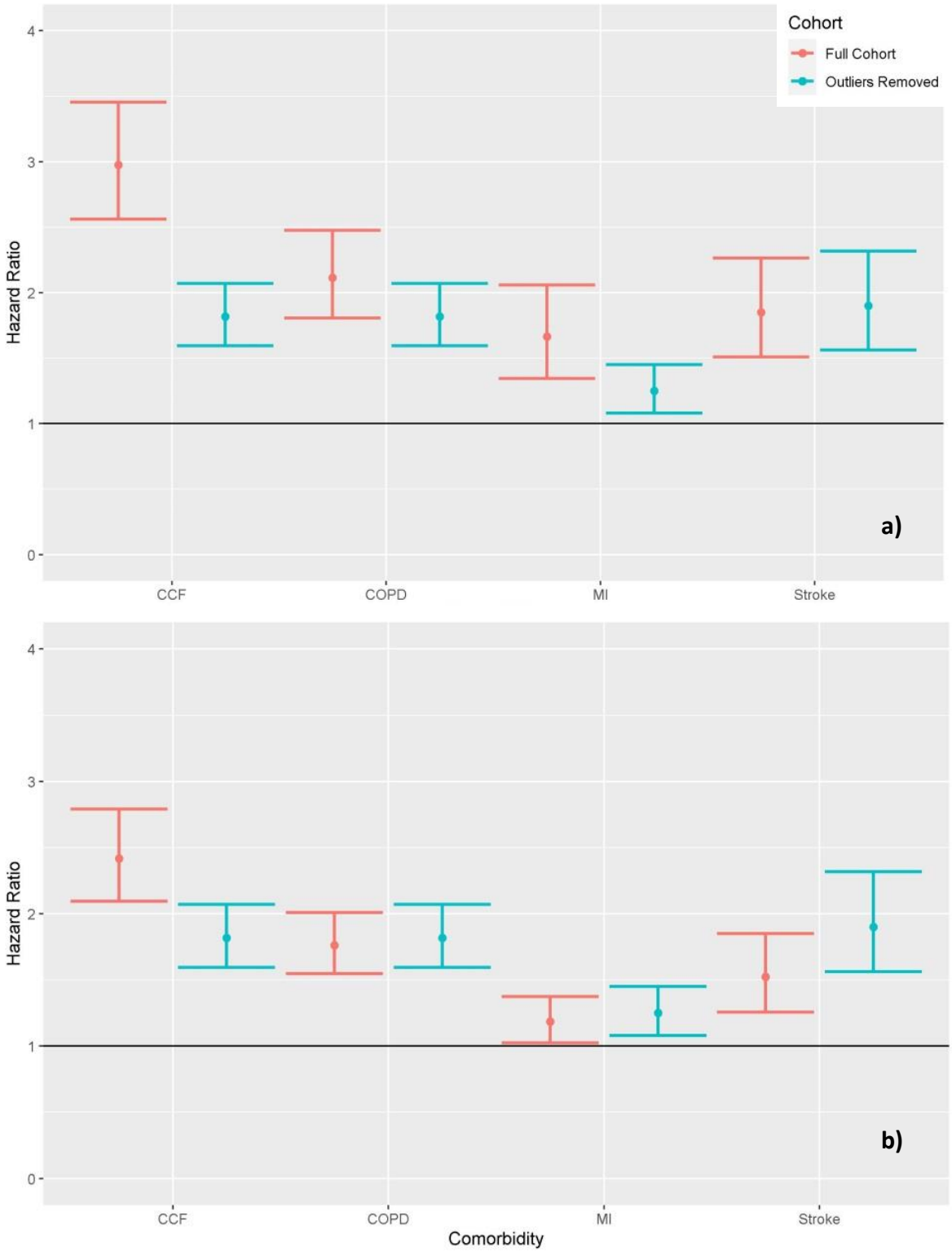


Figure 39: Effect of Influential Outliers – *Cox derived comorbidity hazard ratios for comorbidities with influential outliers identified in a) Breast Cancer and b) Prostate Cancer Site Specific Cohorts. Results in red are the hazard ratios obtained when the full site specific cohort was used to generate a Cox model those in blue are where the model was generated after the removal of influential outliers.*

5.3.2 - Extraction of Summary Statistics

40 models were developed per site specific cohort investigated along with 24 models per comorbidity of interest. The results of these are demonstrated in summary plots of the hazard ratio grouped by site and comorbidity below. Where a hazard ratio is quoted below it is presented as a point estimate with its 95% confidence interval following within parenthesis.

Breast

A total of 28 comorbidities were found to have a high probability of unidirectional hazard (**Figure 40**). In all but one case these associations were with reductions in the survival of these groups. Varicosities where however associated with improved survival with a hazard ratio of 0.69 (0.49-0.97). The largest effect size seen was in ankylosing spondylitis with a hazard ratio of 6.10 (1.96-18.92). It is however important to note that the precision of this estimate is low with wide confidence intervals. Dementia and congestive cardiac failure both had lower point estimates but greater levels of precision, with a hazard ratio of 3.08 (2.53-3.75) and 2.97 (2.56 -3.45) respectively. Of those estimates with a high probability of unidirectional hazard, the lowest effect size was seen in peptic ulcer disease which identified a hazard ratio of 1.35 (1.06-1.71).

Lung

Within the lung cancer cohort a total of 22 comorbidities were associated with a high probability of unidirectional hazard (**Figure 41**). Despite this, the effect sizes seen in this cohort were on average smaller than those found in the breast cancer population. The direction of effect was more varied, with 7 of the 22 comorbidities suggesting a reduction in hazard. The largest reduction was seen in obesity with a hazard ratio of 0.44 (0.39-0.48) and showing a high level of precision. The other conditions associated with decreased hazard were more modest in terms of effect size, with analysis suggesting a hazard ratio of 0.84 (0.78-0.91) for asthma, 0.90 (0.85-0.95) for hyperlipidaemia, 0.91 (0.87-0.94) in COPD, 0.92 (0.88-0.98) for diabetes and 0.87 (0.84-0.91) in hypertension.

Amongst those associated with increased hazard, the largest effect size was seen in motor neuron disease with a hazard ratio of 3.61 (1.16-11.19). Due to small case numbers, the precision of the estimate was low. Demyelination and HIV have the next highest point estimates, however as with MND the precision is low with estimates extending beyond 100% variation of hazard in both directions. The remaining comorbidities with a high probability of unidirectional effect have higher levels of precision but demonstrate more modest effect sizes. Examples include a hazard ratio of 1.47 (1.29-1.67)for thromboembolic disease, 1.42 (1.27-1.60) for dementia, 1.11 (1.01 – 1.22) for CKD and 1.27 (1.01-1.52) for liver dysfunction.

Prostate

A total of 26 conditions were found to have a high probability of unidirectional hazard in the prostate cancer cohort, all of which suggested increased hazard (**Figure 42**). As with the previous cohorts the largest point estimates were associated with low levels of precisions such as liver disease and other respiratory diseases, both of which have confidence intervals spanning more than 200% risk difference. Of those with a higher degree of precision (CI band smaller than 100% risk difference) the largest effect size was seen in CCF with a hazard ratio of 2.42 (2.09-2.79). The smallest effect size was in asthma with a hazard ratio of 1.20 (1.01-1.43).

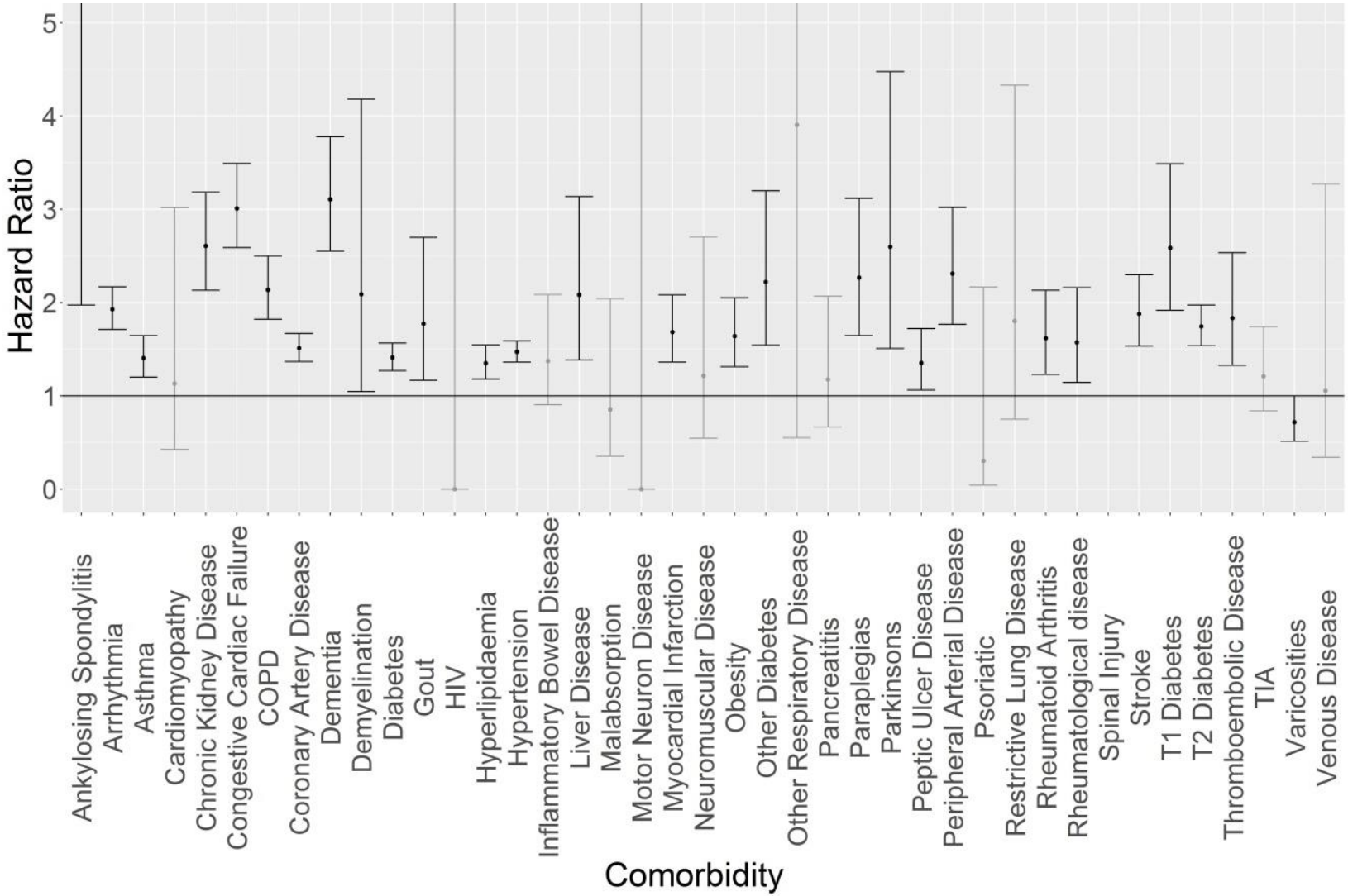


Figure 40: Cox Derived Hazard Ratios for Comorbidity in Breast Cancer *– Cox derived hazard ratios with 95% confidence interval associated with each comorbidity of interest in the Breast Cancer Site Specific Cohort. Each estimate is derived from a standalone model adjusting for age, gender and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

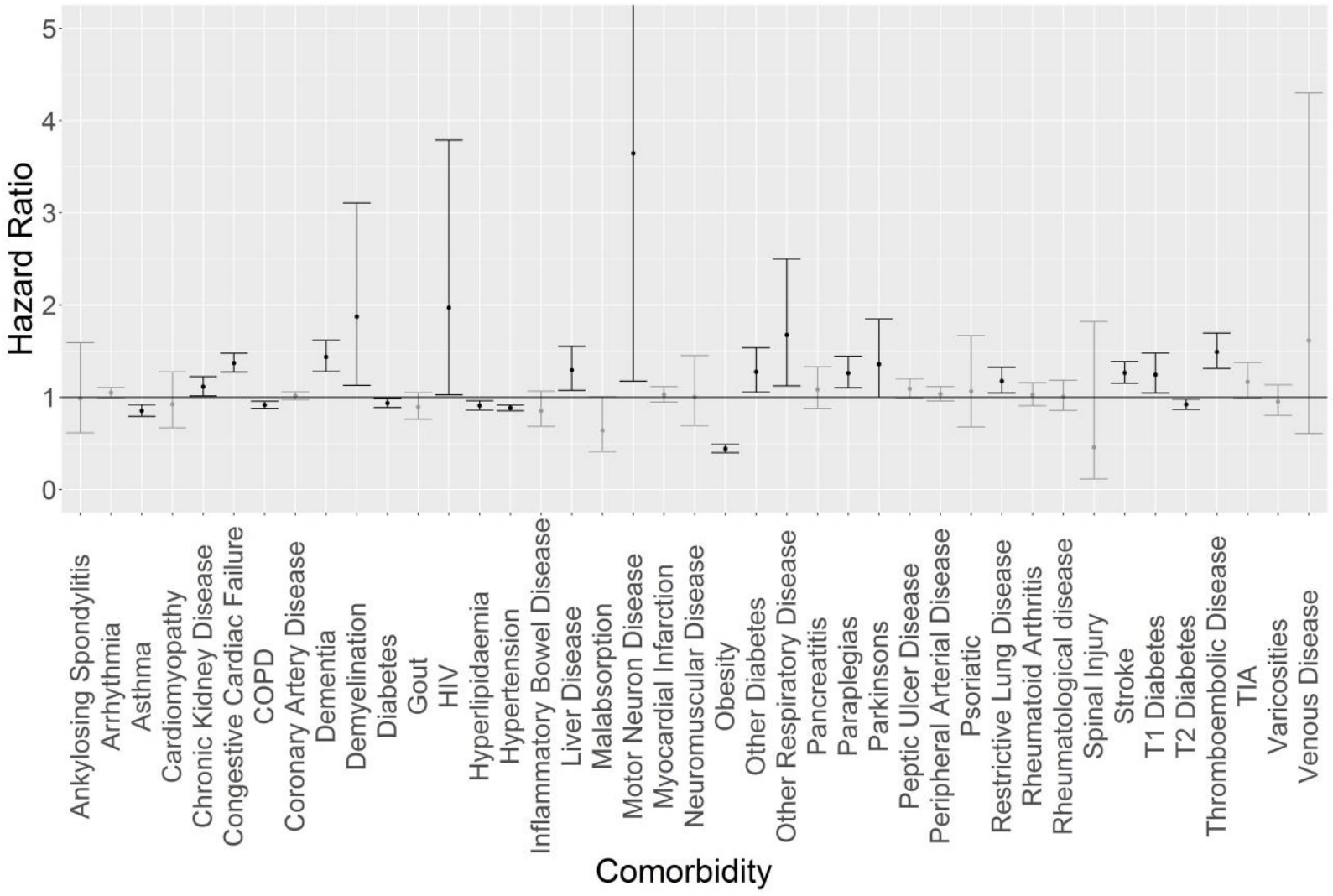


Figure 41: Cox Derived Hazard Ratios for Comorbidity in Lung Cancer *– Cox derived hazard ratios with 95% confidence interval associated with each comorbidity of interest in the Lung Cancer Site Specific Cohort. Each estimate is derived from a standalone model adjusting for age, gender and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

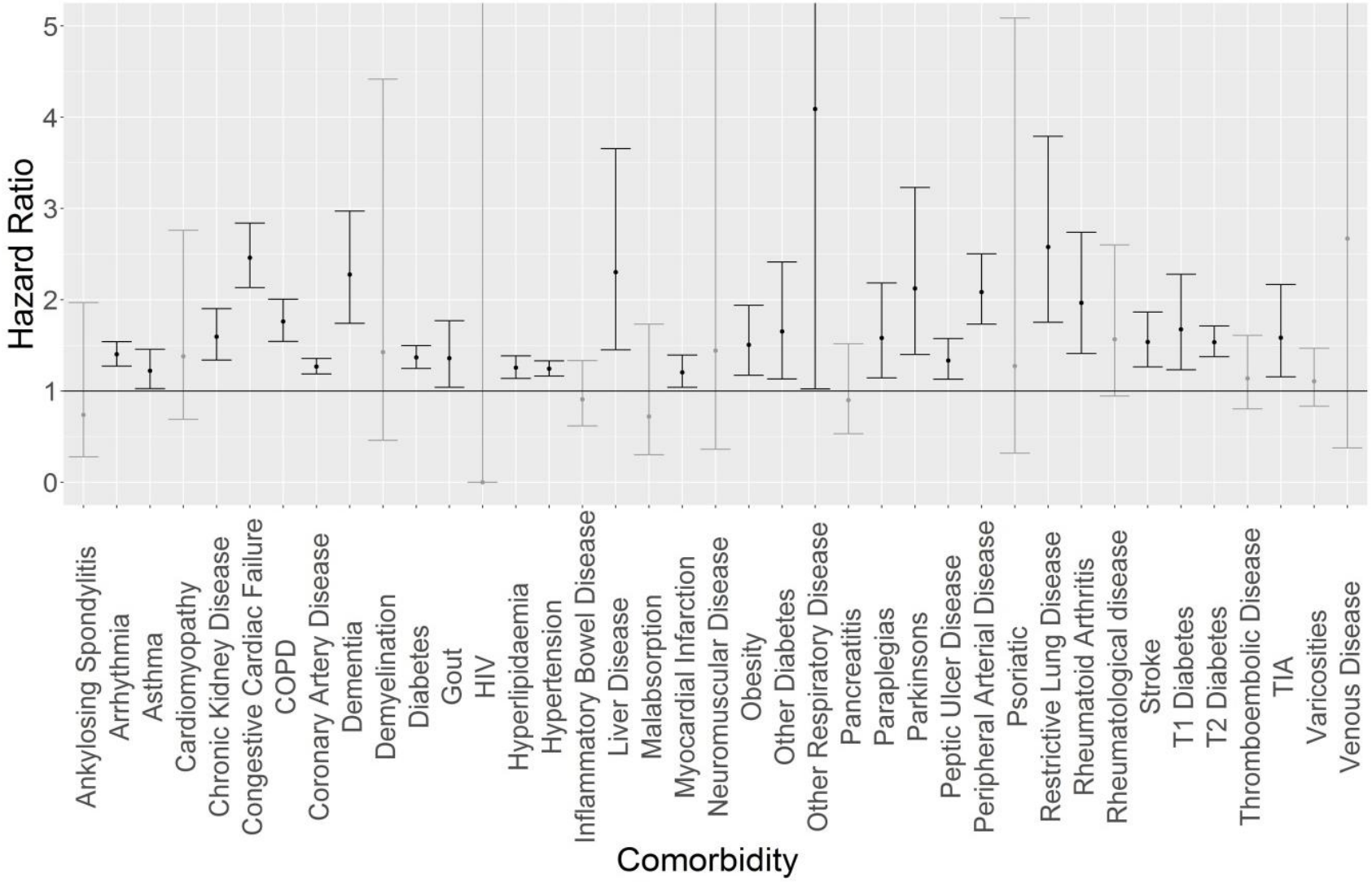


Figure 42 : Cox Derived Hazard Ratios for Comorbidity in Prostate Cancer *– Cox derived hazard ratios with 95% confidence interval associated with each comorbidity of interest in the Prostate Cancer Site Specific Cohort. Each estimate is derived from a standalone model adjusting for age and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

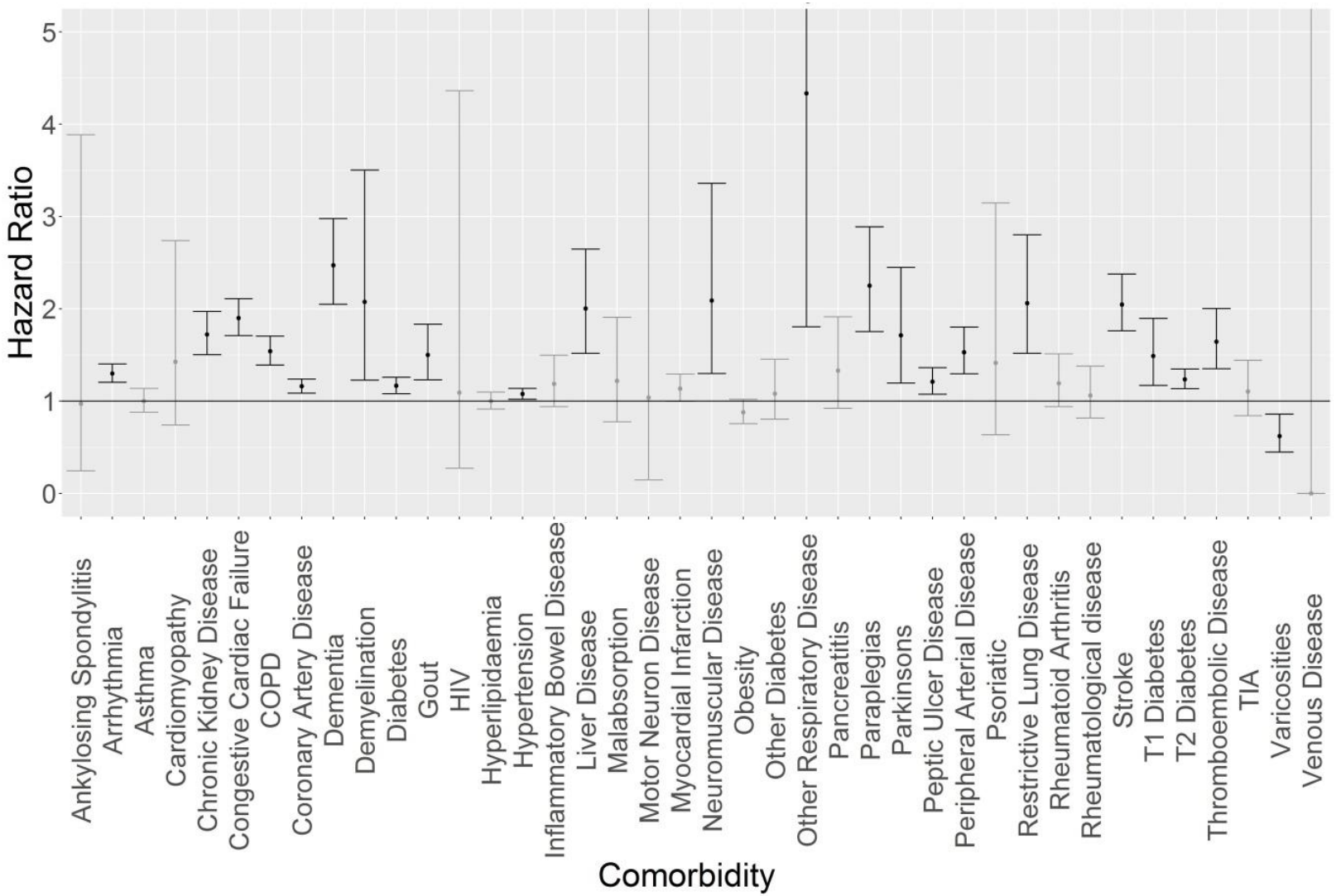


Figure 43: Cox Derived Hazard Ratios for Comorbidity in Colorectal Cancer *– Cox derived hazard ratios with 95% confidence interval associated with each comorbidity of interest in the Colorectal Cancer Site Specific Cohort. Each estimate is derived from a standalone model adjusting for age, gender and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

Colorectal

23 comorbidities in the colorectal cancer cohort were found to have a high probability of unidirectional hazard (**Figure 43**). Varicosities were the only condition associated with a reduction in hazard, with a hazard ratio of 0.63 (0.45-0.86). All other conditions were found to be associated with a deleterious impact on survival. Of those with a level of precision showing a CI band of less than 100 the largest effect size was in dementia with a hazard ratio of 2.46 (2.04-2.96) and the smallest effect size in hypertension with a hazard ratio of 1.07 (1.01-1.13).

Diabetes

When comparing the association of diabetes with survival across all site specific cohorts, 11 sites are associated with a high probability of unidirectional hazard (**Figure 44**). Of these two are associated with survival advantages and nine with survival disadvantages. The greatest survival advantage was seen in primary liver malignancy with a hazard ratio of 0.77 (0.68-0.87). As mentioned previously, lung cancer had a hazard ratio of 0.92 (0.88-0.98) Of those with increased hazard and a confidence interval narrower than 100% risk span, the largest effect was seen in skin cancers with a hazard ratio of 1.52 (1.42-1.63). The smallest effect was seen in colorectal cancer with a hazard ratio of 1.15 (1.07-1.24).

MI

A smaller number of sites were associated with unidirectional hazard (8 sites) when assessing the impact of MI (**Figure 45**). The level of precision for most estimates was also low with only prostate cancer and skin cancer showing precision of better than a 100% hazard span. In skin cancer, this was associated with a hazard ratio of 1. 35 (1.19-1.52) and 1.19 (1.03-1.37) in prostate cancer

Stroke

12 cancer sites demonstrated a unidirectional hazard relationship with stroke. In all cases, stroke was associated with worse outcomes (**Figure 46**). The largest risk difference seen was in lymphoma with a hazard ratio of 2.25 (1.72-2.95). As with the other analyses this largest effect size was also accompanied by low precision, with a confidence interval span exceeding 100% risk difference. The majority of unidirectional effects seen, were accompanied by low levels of precision with only lung cancer showing a CI span of less than 100%. This was also the lowest effect size seen a hazard ratio of 1.25 (1.14-1.37).

CCF

The average effect size seen across all sites was largest in CCF compared to all other conditions analysed (**Figure 47**). All but 4 sites were associated with a high probability of unidirectional reductions in survival outcomes in the comorbid group. The largest effect size was seen in testicular cancer, with a hazard ratio of 8.21 (1.14-59.1), , despite this, the precision is extremely low with the lower bound of the CI suggesting a risk increase of just 14%. CCF was also associated with large hazard differences in the breast cancer population with a hazard ratio of 2.97 (2.56-3.45). By contrast to testicular cancer the precision of this estimate was far greater leading to higher confidence that the true effect is greater than 156% increase in hazard. The lowest effect size was seen in lung cancer which showed a hazard ratio of 1.36 (1.26-1.46).

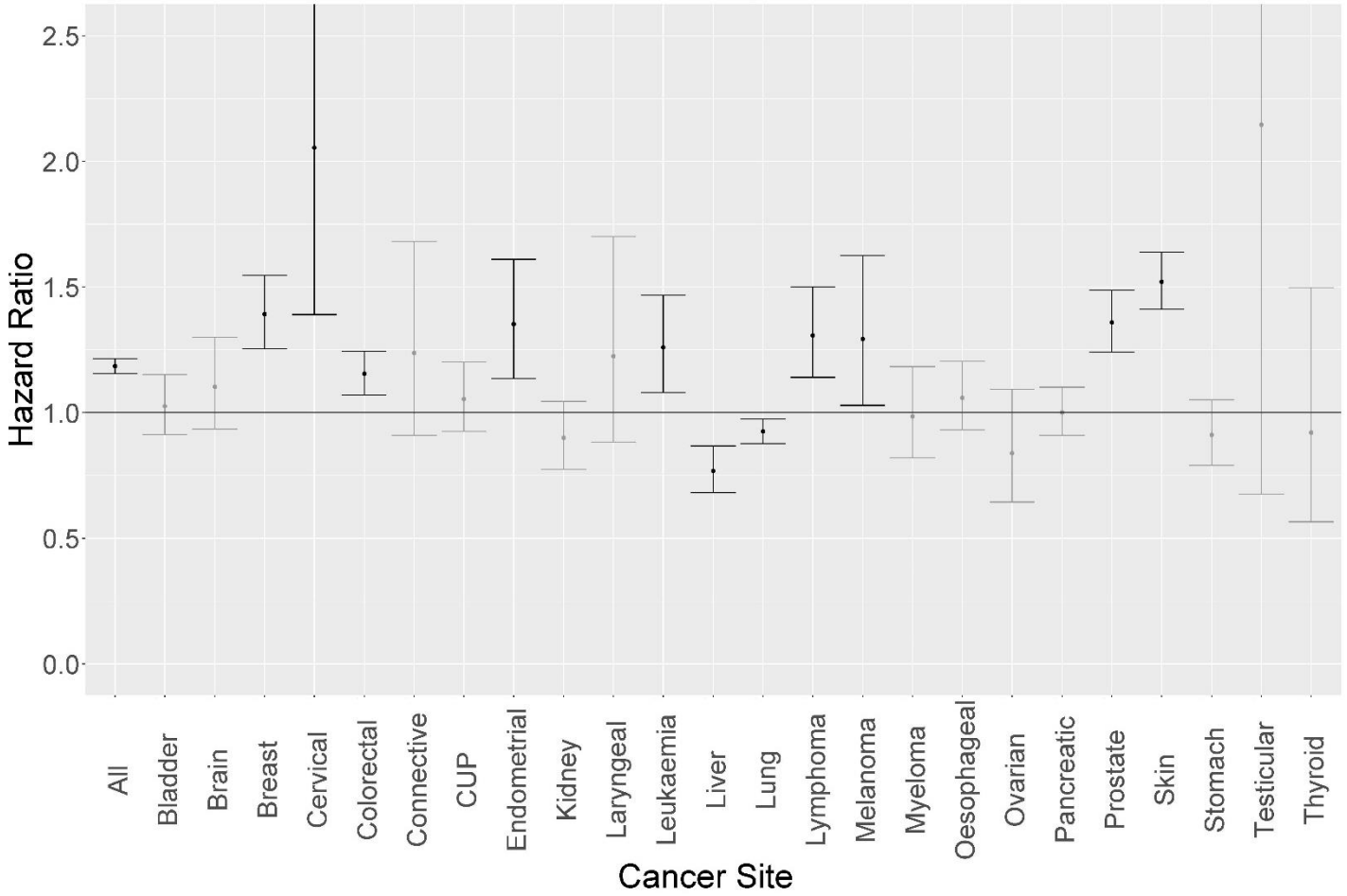


Figure 44: Cox Derived Hazard Ratios for Diabetes Mellitus in All Cancer and Site Specific Cancer Cohorts *– Cox derived hazard ratios with 95% confidence interval associated with Diabetes Mellitus in All Cancer Cohort and Cancer Site Specific Cohorts. Each estimate is derived from a standalone model adjusting for age, gender (where appropriate) and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

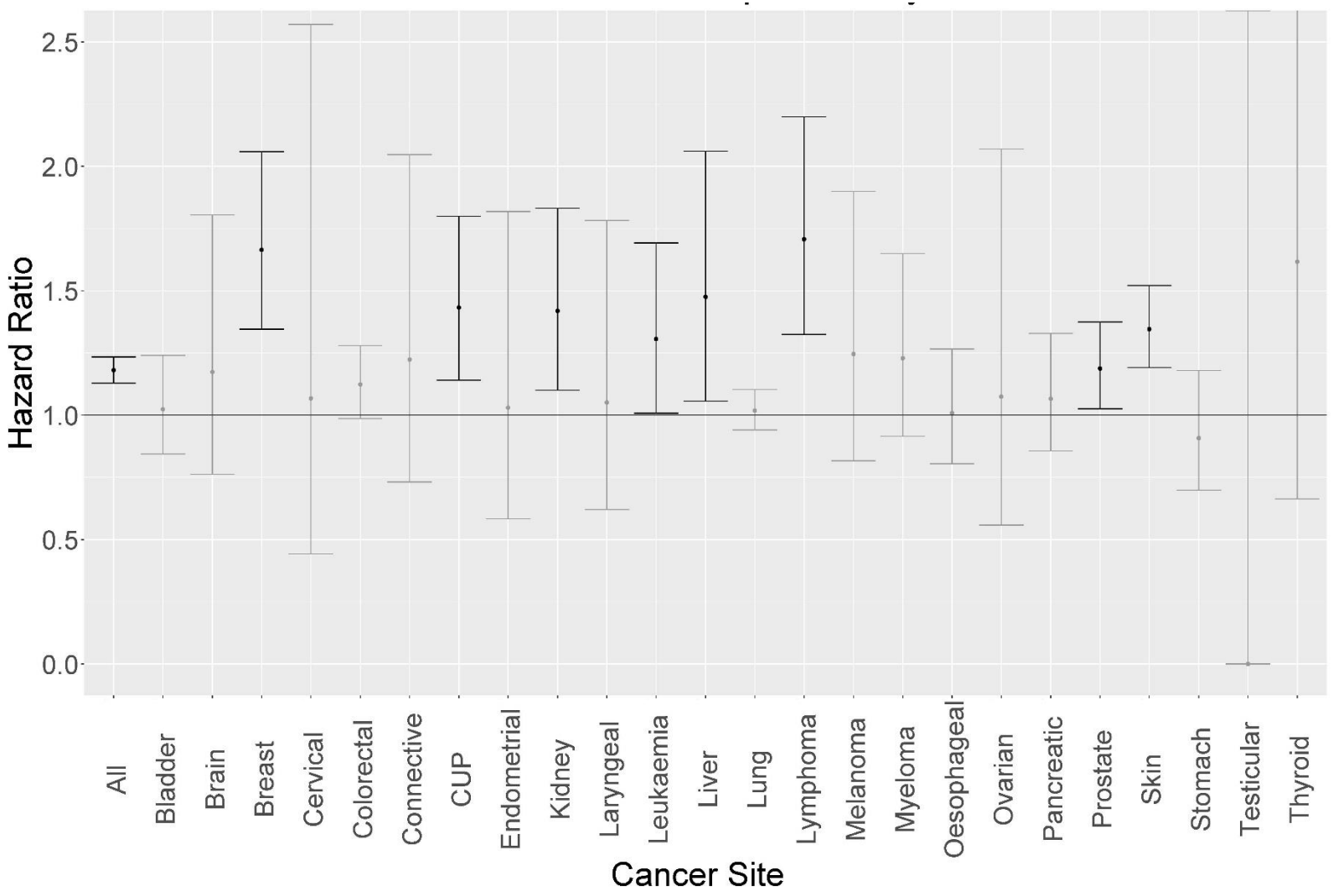


Figure 45: Cox Derived Hazard Ratios for MI in All Cancer and Site Specific Cancer Cohorts *– Cox derived hazard ratios with 95% confidence interval associated with MI in All Cancer Cohort and Cancer Site Specific Cohorts. Each estimate is derived from a standalone model adjusting for age, gender (where appropriate) and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

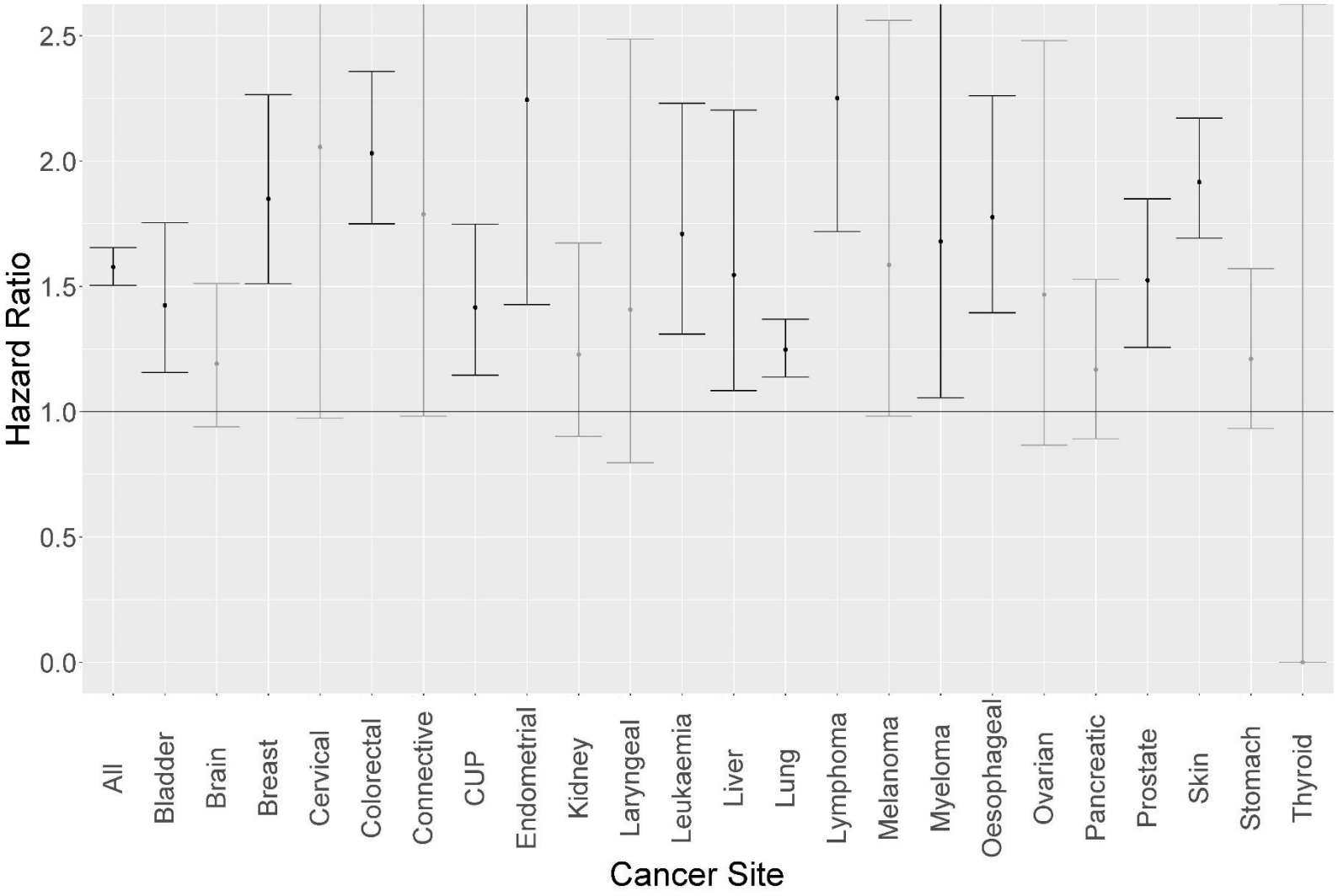


Figure 46: Cox Derived Hazard Ratios for Stroke in All Cancer and Site Specific Cancer Cohorts *– Cox derived hazard ratios with 95% confidence interval associated with Stroke in All Cancer Cohort and Cancer Site Specific Cohorts. Each estimate is derived from a standalone model adjusting for age, gender (where appropriate) and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

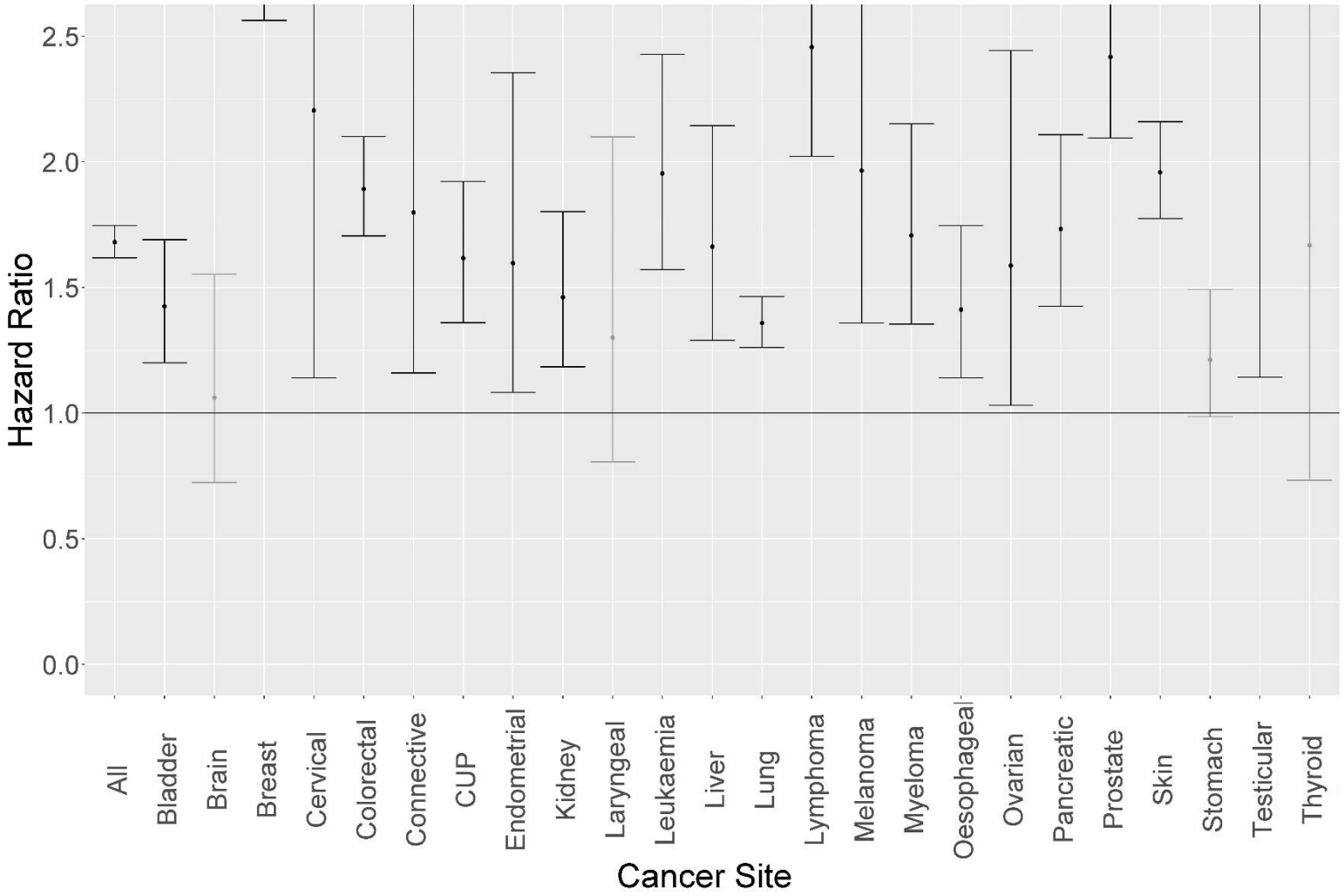


Figure 47: Cox Derived Hazard Ratios for CCF in All Cancer and Site Specific Cancer Cohorts *– Cox derived hazard ratios with 95% confidence interval associated with CCF in All Cancer Cohort and Cancer Site Specific Cohorts. Each estimate is derived from a standalone model adjusting for age, gender (where appropriate) and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

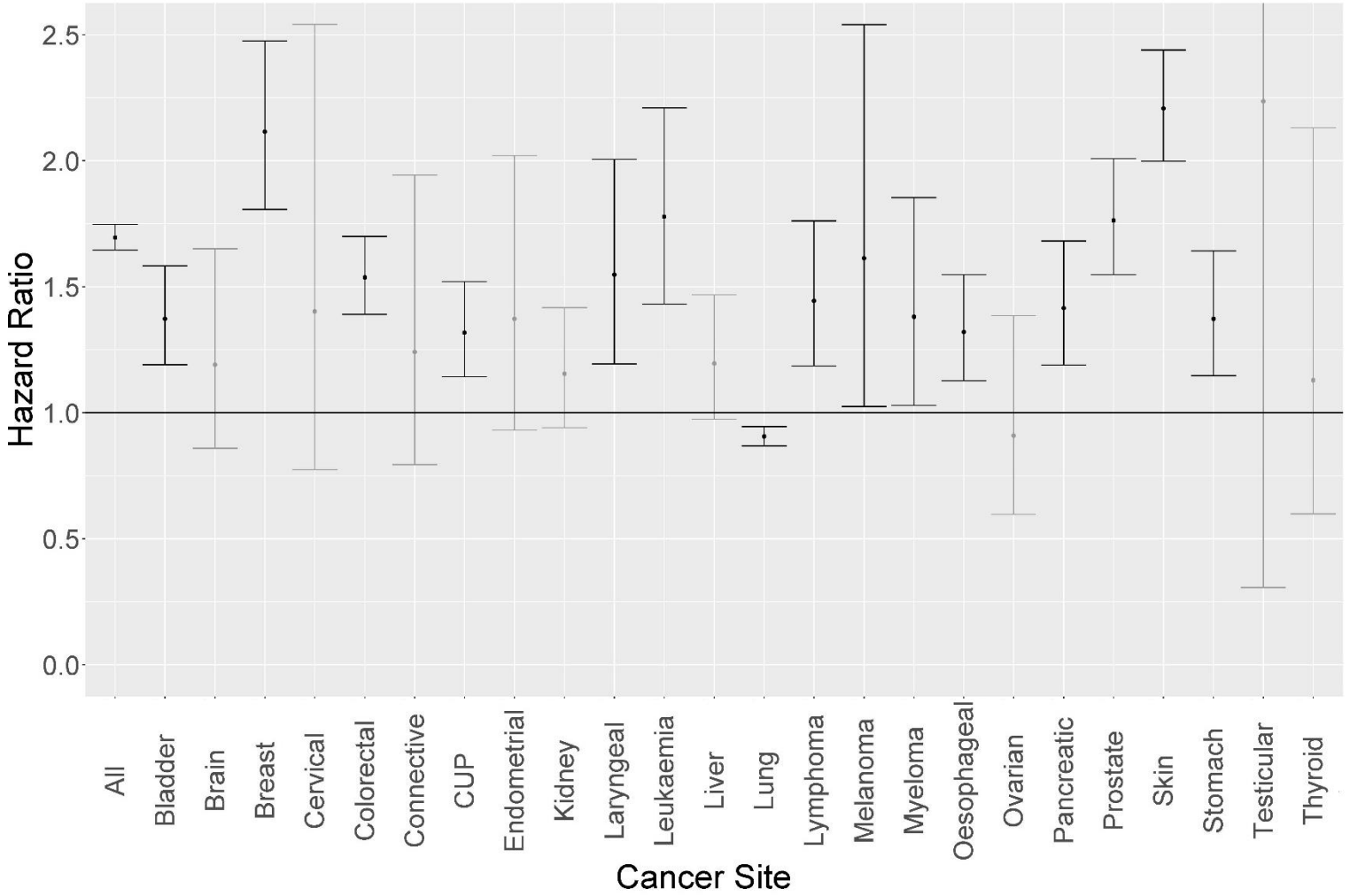


Figure 48: Cox Derived Hazard Ratios for COPD in All Cancer and Site Specific Cancer Cohorts *– Cox derived hazard ratios with 95% confidence interval associated with COPD in All Cancer Cohort and Cancer Site Specific Cohorts. Each estimate is derived from a standalone model adjusting for age, gender (where appropriate) and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

COPD

15 cancer sites were associated with a high probability of unidirectional hazard when assessing COPD (**Figure 48**). Lung cancer was associated with a modest reduction in hazard with a hazard ratio 0.91 (0.87-0.94). All other unidirectional results were associated with worse survival outcomes and increased hazard. Skin cancer demonstrated the largest effect size with a hazard ratio of 2.21 (2.00-2.44). This was accompanied by reasonable precision with a CI span of less 50%.

Highest precision results

Table 10 shows the hazard ratios that met our pre-specified precision cut offs described in the methods section. This highlights a number of examples with both a high probability of unidirectional effect and high precision for the point estimates. When assessed in the all cancer cohort 20 comorbidities meet the precision threshold. By contrast in breast cancer only one comorbidity meets the threshold, four in prostate cancer, nine in lung cancer and five in colorectal cancer. Hypertension features as precise and unidirectional across all of the top four cancers with an increased hazard in all but lung cancer, which shows a 10.1% decrease. Diabetes meets the precision thresholds in lung, colorectal and prostate cancer. Of note hyperlipidaemia, COPD, asthma and diabetes which all show an associated reduction in hazard in lung cancer also meet the precision threshold.

	Label	Hazard Ratio	Lower CI	Upper CI	CI Span	CI Span as Percentage
Breast	Hypertension	1.45	1.34	1.56	0.22	15.43%
Prostate	Diabetes	1.36	1.24	1.49	0.25	18.20%
	Coronary Artery Disease	1.25	1.17	1.34	0.17	13.36%
	Hypertension	1.23	1.16	1.32	0.16	13.31%
	Hyperlipidaemia	1.25	1.13	1.38	0.24	19.58%
Lung	Diabetes	0.92	0.88	0.98	0.10	10.66%
	T2 Diabetes	0.91	0.86	0.97	0.11	12.13%
	Congestive Cardiac Failure	1.36	1.26	1.46	0.20	14.96%
	Hypertension	0.87	0.84	0.91	0.06	7.29%
	Asthma	0.84	0.78	0.91	0.13	14.91%
	COPD	0.91	0.87	0.94	0.08	8.47%
	Chronic Kidney Disease	1.11	1.01	1.22	0.21	18.84%
	Stroke	1.25	1.14	1.37	0.23	18.50%
	Hyperlipidaemia	0.90	0.85	0.95	0.10	11.13%
Colorectal	Diabetes	1.15	1.07	1.24	0.17	15.17%
	T2 Diabetes	1.22	1.12	1.33	0.21	17.16%
	Coronary Artery Disease	1.15	1.08	1.22	0.15	12.96%
	Hypertension	1.07	1.01	1.13	0.11	10.73%
	Arrhythmia	1.29	1.19	1.39	0.20	15.25%
Skin	Diabetes	1.52	1.41	1.64	0.23	14.89%
All	Diabetes	1.18	1.16	1.22	0.06	5.07%
	T1 Diabetes	1.54	1.42	1.67	0.24	15.92%
	T2 Diabetes	1.30	1.27	1.34	0.08	5.85%
	Myocardial Infarction	1.18	1.13	1.23	0.10	8.89%
	Congestive Cardiac Failure	1.68	1.62	1.75	0.13	7.61%
	Coronary Artery Disease	1.15	1.13	1.18	0.05	4.21%
	Hypertension	1.11	1.09	1.13	0.04	3.70%
	Arrhythmia	1.29	1.26	1.33	0.07	5.46%
	Varicosities	0.81	0.74	0.88	0.14	17.14%
	Thromboembolic Disease	1.78	1.66	1.90	0.24	13.38%
	Asthma	1.11	1.06	1.16	0.09	8.34%
	COPD	1.70	1.65	1.75	0.10	6.02%
	Inflammatory Bowel Disease	1.20	1.09	1.32	0.23	19.33%
	Peptic Ulcer Disease	1.25	1.20	1.31	0.11	8.80%
	Chronic Kidney Disease	1.47	1.40	1.54	0.14	9.39%
	TIA	1.19	1.09	1.30	0.21	17.71%
	Stroke	1.58	1.50	1.66	0.15	9.62%
	Dementia	1.87	1.76	1.98	0.22	11.88%
	Rheumatological disease	1.20	1.10	1.31	0.21	17.73%
	Rheumatoid Arthritis	1.32	1.23	1.41	0.19	14.07%
	Hyperlipidaemia	1.04	1.01	1.08	0.06	5.96%
	Peripheral Arterial Disease	1.55	1.48	1.62	0.15	9.48%

Table 10: High Precision Comorbidity Hazard Ratio Results *- Hazard ratios associated with comorbidity extracted from each Cox proportional hazard model. The confidence interval for the hazard ratio was extracted and is shown as the lower confidence interval, upper confidence interval and the total span of the confidence interval. The span is further expressed as a percentage of the hazard ratio point estimate. The results shown are limited to those that have a high probability of unidirectional hazard, a confidence interval span of less than or equal to 0.25 and where the span was less than or equal to 25% of the point estimate for the hazard ratio*

5.4 - Discussion

5.4.1 - Assessment of Model Assumptions

As per our analysis framework detailed in chapter 2, first consideration is given to limitations and bias within the analysis. A key step in this process is the interpretation of the model assumptions applied, and whether these hold. The results of the model diagnostics identify a number of potential issues with the models developed. Many of these issues stem from methodological limitations or assumptions inherent to the Cox proportional hazards approach.[178,288] The assessment of Schoenfeld residuals to look for evidence of variables breaching the proportional hazards assumption is in this instance, based on a subjective assessment. In most cases a statistically significant p value would be an indication that the assumption has been breached. This could either be globally across the model, or in terms of individual covariates. The issue with this approach, in this context, is that it is recognised that p values are of limited utility in large cohorts, as very small differences in absolute values can result in highly significant p values due to the sample size. Previous literature has suggested that samples at or above 10,000 should be used as a hard cut off for the utility of p values[290]. Although where researchers choose to draw their arbitrary cut off is a matter for debate, it is certainly the case that this is an issue for the high prevalence cancers in PPM that have large numbers often exceeding 20,000 patients. This requires instead the assessment of the visual trends and the application of a subjective judgement. This approach is inherently more open to variation and assessor bias. It could be argued that in the cases of less common cancers the numerical approach could have been used due to the smaller sample sizes, however this would have meant applying our analysis inconsistently across models.

When assessing the models visually, it is also important to state that this subjective review might be influenced by the scale used within the plots. If plots had been assessed on a smaller scale the level of variation from a straight horizontal line would become more apparent. Thus how the results are represented could have influenced the interpretation of the proportional hazards assumptions. It is also important to note, that in some cases where models were based on smaller samples sizes there was disagreement between the numeric assessment produced via p values and the visual assessment. This introduces the question of robustness of the visual inspection approach. This issue could be entirely avoided through the use of other modelling approaches that do not in rely on the proportional hazards assumption.

The assessment of the linear assumptions showed that both of the continuous variables used demonstrated some degree of non-linearity across several of models. As with the interpretation of the proportional hazards assumption, the interpretation is subjective and how much deviation away from a straight line relationship was regarded as being a violation of the assumption, may not have been entirely consistent as a result. The identified of non-linear associations were most apparent when assessing age in breast cancer, which is consistent with previous research findings.[291] This issue is an inherit problem in most methods reliant on linear regression to model a time to event. A variety of approaches can be used to overcome this, including dividing up the cohort. Here individual models are built using subsets of the population and later plots can be “stitched” together to identify the non-linear patterns. This is however only useful if you are attempting to model this relationship with age as opposed to trying to develop an overall prediction or explanatory model. Where multiple continuous variables are demonstrating nonlinearity, then further subdivisions must be made, i.e. each age band in each IMD band. This then produces issues due to having small sample sizes with much lower precision of estimates, whilst also requiring a significant increase in the number of models being built, impacting on the practicality of this approach. Further, it prevents inclusion of interaction terms between these, if this were felt to be required and appropriate. In

order to overcome this, alternative flexible non-parametric approaches could be applied which do not make linear assumptions. However, many of these approaches are time consuming and also may be less easy to interpret the resulting model outputs.

The assessment of outliers found that in the majority of models there were no individuals that were contributing to large scale shifts in the effect size estimates for each of the variables. Most of the examples of influential outliers were in the context of the comorbidity coefficient estimation. This was particularly apparent in those comorbidities with low numbers. In many of these cases, the majority of comorbid patients were influential, suggesting variance in the outcomes of these patients. This variance could be due to the sample size being small and thus insufficient numbers are available to accurately estimate the average effect. It may also be the case that the association between that comorbidity and outcome is varied in the population and that large sample sizes would show the same high variance. As such, further research should be undertaken using larger sample sizes to identify whether this solves the variance issue or not.

The results from analyses with and without influential outliers demonstrates that there is variation in both the point estimates and confidence intervals derived from the models. In all cases the direction of effect was constant but the precision and effect size was altered. Although the effect of outliers was not assessed in terms of hazard ratio in all models, it is reasonable to assume that a similar pattern of effect would be seen in other cases. The models more likely to be impacted, are those with smaller numbers of comorbid patients. These models have lower precision and should therefore be viewed with a higher degree of scepticism as a result.

Although not explicitly assessed within this chapter, the issue of collinearity, as identified in chapter 3, is of relevance. Where variables are correlated or collinear there is a risk of inappropriately attributing the effect of one covariate to its correlated counterpart. As a result, there is a risk that due to the low level, but significant correlation seen between several covariates, this may be inflating or deflating the estimates of effect size. It would be possible to introduce interaction terms, however as almost all covariates were correlated to some extent, and that the analysis aims to estimate the total effect of comorbidity, doing so would risk splitting the effects between multiple interaction terms, as well as reducing precision of estimates due to an increasing number of covariates on a polynomial scale.

5.4.2 - Identifying Potential Bias Using Directed Acyclic Graphs (DAGs)

Within the data exploration outlined in chapter 3 a number of limitations of the underlying dataset have been identified and described. These include; missing data, collinearity, limitations of clinical coding and patient demographics that are unrepresentative of the wider UK population. Within section 5.4.1 a number of methodological assumptions are further identified and assessed, all of which may introduce inaccuracy and bias into analysis results. Beyond this, there may be elements of the observational study design, which may introduce additional bias and misestimation that must be considered.

The fundamental question at the heart of the analyses undertaken in this chapter and chapter 4, is "what is the impact of individual comorbidities on cancer outcomes"? Although the results presented are clearly described as associations, descriptive or inferential analyses, the question and its ideal answer are causal. The phrase "correlation does not equal causation" is commonly found within the literature and popular science writing [292] but despite this, correlated variables are often presented as being a finding that is meaningful and potentially clinically relevant or useful. In effect, correlation is used to suggest a hint of possible causation. Causal inference research has however shown the absence of correlation does not mean the absence of causation, and may be an artefact

of the study design and the analysis implemented.[293] In view of this, it is important to objectively assess the analyses undertaken, to determine if the results are biased, if so how and whether they are interpretable enough to draw meaningful conclusions about associations, as potential indicators of causal relationships, which require further investigation.

Although our analysis is not attempting to make causal conclusions, the application of causal thinking can still be utilised to suggest additional sources of bias not previously considered or detailed. A formal assessment of the analysis can be built around current knowledge of the likely temporal ordering of relevant factors. Drawing this knowledge in the form of a directed acyclic graph (DAG) allows us to assess potential biases that are introduce throughout study design and the analysis approach.[294,295]

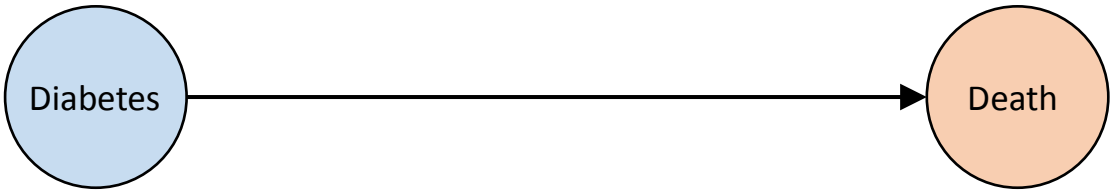


Figure 49: Diabetes and Death Directed Acyclic Graph (DAG) *- Representation of simplified cause and effect of diabetes on death. Light blue denotes exposure of interest and red denotes outcome of interst.*

The key desire from any causal analysis is to identify the causal effect of an exposure on an outcome of interest. By taking a single example of the impact of diabetes on death in cancer, it is possible to start to identify potential pitfalls in the methods applied. As our analysis is concerned with pre-cancer diabetes, then our exposure is diabetes and our outcome of interest is death (**Figure 49**). The research question concerns the impact of diabetes in a specific population, and thus cancer must be added as a mediator (**Figure 50**). If this simple causal diagram were the true reflection of reality, and our estimand (the quantity that is being estimated) was the total causal effect, then the causal relationships of interest would be represented by **Figure 51**.

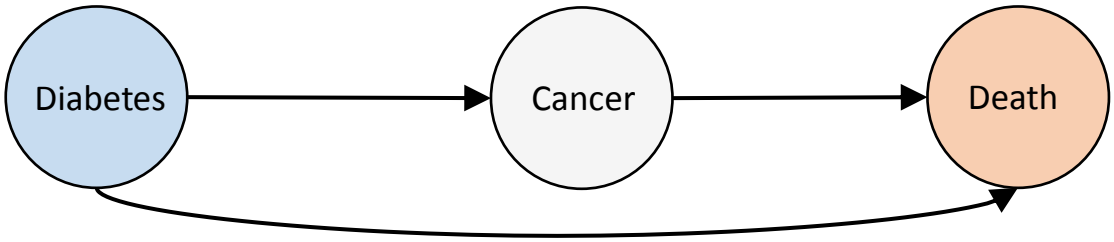


Figure 50: Diabetes and Death Mediated via Cancer Directed Acyclic Graph (DAG) *- Representation of simplified cause and effect of diabetes on death mediated via cancer. Light blue variables denote the exposure of interest, grey denotes mediators of effect and red denotes the outcome of interst.*

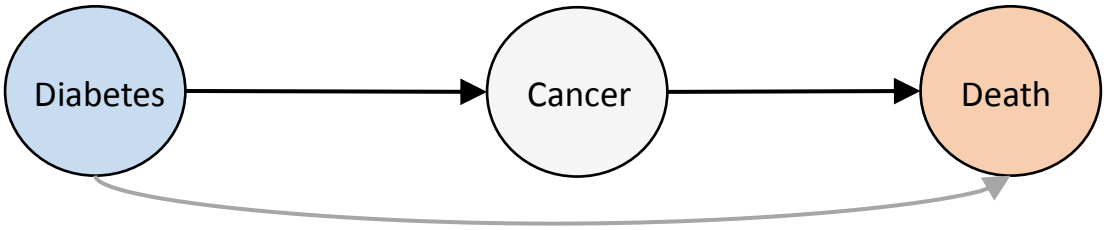


Figure 51: : Total Causal Effect of Diabetes and Death Mediated via Cancer Directed Acyclic Graph (DAG) *- Representation of simplified cause and effect of diabetes on death mediated via cancer. Light blue variables denote the exposure of interest, grey denotes mediators of effect and red denotes the outcome of interst. Grey lines represent causal relationships that should not be included within the estimand in order to obtain the desired total causal effect estimate.*

The true causal relationships are however more complex, and this must be reflected in the DAG. It is necessary therefore to consider the mediators of death caused by diabetes via cancer and non-cancer related processes. There are other additional factors beyond these that alter a patient's risk of cancer and also affect death, but are independent of cancer. If these were included as grouped entities then the DAG would be represented by **Figure 52** . With this particular DAG, the total causal effect would be represented by **Figure 53**.

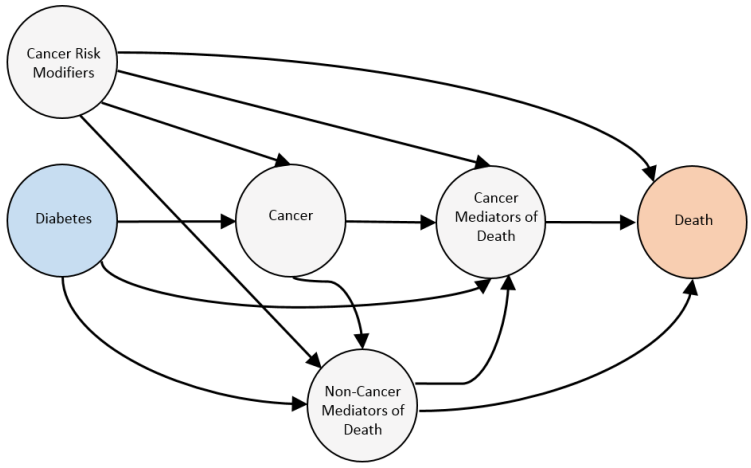


Figure 52: Detailed Diabetes and Death Mediated via Cancer Directed Acyclic Graph (DAG) *- Representation of cause and effect of diabetes on death mediated via cancer with the addition of cancer risk modifiers and mediators of death both cancer and non-cancer related. Light blue variables denote the exposure of interest, grey denotes mediators of effect and red denotes the outcome of interst.Despite being more detailled this DAG still represents a simplification of the known causal relationships.*

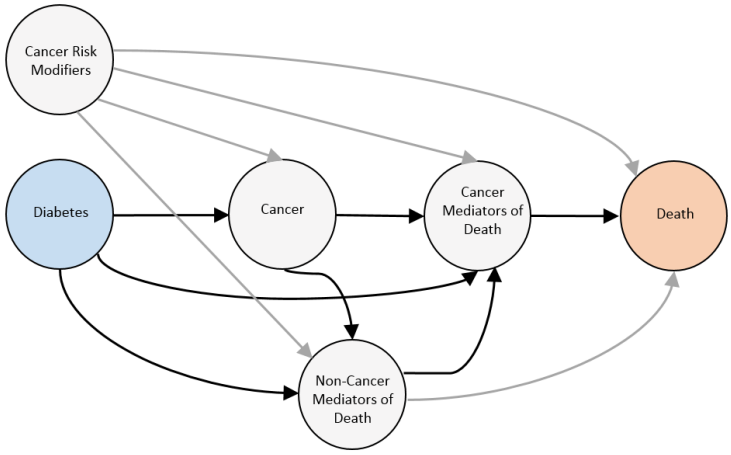


Figure 53: Detailed Total Causal Effect of Diabetes and Death Mediated via Cancer Directed Acyclic Graph (DAG) *- Representation of causal pathway of diabetes on death mediated via cancer including the addition of cancer risk modifiers and mediators of death both cancer and non-cancer related. Light blue variables denote the exposure of interest, grey denotes mediators of effect and red denotes the outcome of interst. Despite being more detailled this DAG still represents a simplification of the known causal relationships. Grey lines represent causal relationships that should not be included within the estimand in order to obtain the desired total causal effect estimate.*

Although this appears at first simple to achieve, our study design is limited to cancer patients and therefore involves stratifying patients by the presence or absence of cancer, which conditions on cancer as a variable.[296] This blocks downstream causal effects which are removed, at least in part, from any estimates and introduces collider bias, such that all conditions that impact the risk of cancer and are associated with death, are added to the estimates obtained. The open causal pathways are represented in **Figure 54**.

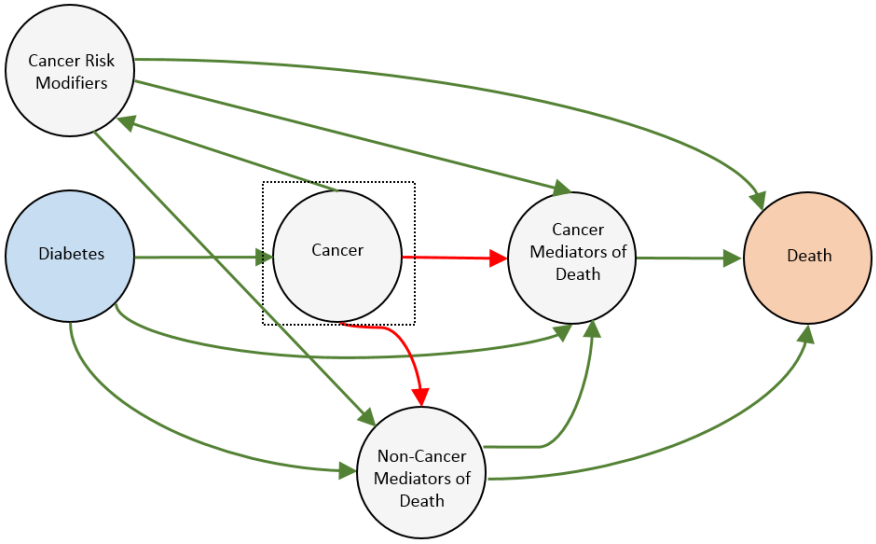


Figure 54: DAG of Open and Closed Causal Pathways due to the Study Design *– Simplified DAG representing the causal pathways due to the implemented study design for assessing the effect of diabetes on death via cancer. Green lines represent open pathways, red represent cosed pathways. Mediators within a dotted box represent those that have been conditioned upon. . Light blue variables denote the exposure of interest, grey denotes mediators of effect and red denotes the outcome of interst.*

In order to overcome this issue it would be possible to condition on the cancer risk factors as shown in **Figure 55**. Although this sounds straightforward, this represents a significant challenge. It assumes that all cancer risk factors are known, which is unlikely to be true. Even if all risk factors were known, they would all have to be measured and it is probable that some or many would be missing from any dataset. Even if they were known and measured, the pattern of relationship between the risk factor and death is unknown. Selecting an appropriate model to adjust for this would therefore be extremely difficult. If for example the relationship was non-linear, then linear models will not adequately remove bias. If we assume that this too was not an issue, and all risk factors were known, measured and pattern of association known, even then, conditioning on a variable does not entirely remove bias due to natural variation and misclassification, thus residual confounding remains.[297] In the case of non-cancer death mediators, similar issues are present, however it would also need to be known which were cancer related, and which were not.

A further issue arises via the temporal ordering of variables. If we again take diabetes as an example, in many patients renal dysfunction arises after long standing diabetes.[298] In this scenario renal dysfunction is a mediator of effect and thus should not be adjusted for in the analysis. In other cases however, a patient may develop autoimmune renal dysfunction, which is managed in part with steroids, which can cause diabetes to develop.[299] In this scenario renal dysfunction is a confounder and should be adjusted for. This highlights how the ordering of events impacts on the model specification[300]. If this was put into practice, it would require an exponential increase in the number of models specifications needed, with the increasing number of variables required to account for

confounding. The greater the number of models, the smaller the cohort size for each and thus lower levels of precision would be obtained.

These factors are hugely important when considering how we can and should interpret the results of the analysis presented above. The approach undertaken leaves a large number of unknowns about the accuracy of the estimates obtained and in some cases the underlying methodology of the analysis may introduce false associations. It is therefore important to avoid translating the results seen into causal type conclusions. Within the cancer population these associations can be seen, but may in fact be due to factors entirely separate to the variable the hazard is attributed to by the model. It may be that comorbidity in many instances is simply acting as a surrogate marker for another critical causal factor or due to selection bias introduced by conditioning on cancer. Additionally, the lack of an identified association does not indicate a lack of a causal relationship, which may have been masked due to the analysis employed.

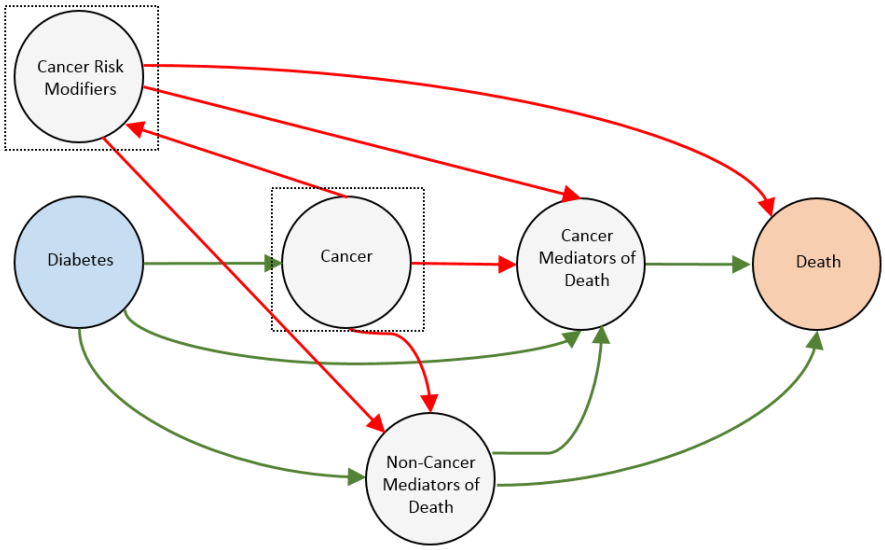


Figure 55: DAG of Open and Closed Causal Pathways Due to Study Design After Adjustment *– Simplified DAG representing the causal pathways due to the implemented study design for assessing the effect of diabetes on death via cancer after adjusting for cancer risk modifiers. Green lines represent open pathways, red represent cosed pathways. Mediators within a dotted box represent those that have been conditioned upon. . Light blue variables denote the exposure of interest, grey denotes mediators of effect and red denotes the outcome of interst.*

This information could also help to explain the number of survival paradoxes that have been identified in our analyses thus far. As outlined above, limiting our study to the cancer population has in effect conditioned on cancer. This is a collider, which is temporally downstream of our exposure which in this case is comorbidity. This has therefore introduced bias which is in most cases unmeasured and unaccounted for within the model. The differences seen, in for example improved survival with COPD in lung cancer, may be entirely artefactual or be independent of COPD entirely. This is similar to other survival paradoxes which has previously been described in the literature such as the birth weight paradox and the obesity cardiovascular paradox.[265,301]

5.4.2 - Associations between Comorbidity and Survival

As with the previous KM analysis, the effect size and direction has been shown to differ across comorbidities and cancer sites. Across all comorbidities and all cancers, numerous associations have been demonstrated to have a high probability of unidirectional hazard in terms of survival outcomes. The majority of these associations suggest that having comorbidity is associated with worse survival

outcomes. Despite this, a number of comorbidities are associated with improved outcomes. This includes the three conditions identified with the KM analysis, along with five additional conditions once adjustment for age, gender and deprivation had been undertaken. As discussed both above and in chapter four, the associations could be as a result of a number forms of bias and confounding. It could be argued that in several cases this is the most likely cause of the associations seen. Despite this, it is important to consider if there are plausible biological mechanisms, which might explain the associations seen. Below, relevant aspects of the literature will be summarised, with an initial focus on comorbidity in general before focussing on the interplay between cancer and each of our key comorbidities, namely; MI, Stroke, CCF, COPD and DM. Potential biological explanations that might account for the highlighted differences in outcome associations will be presented. Beyond these comorbidities, focus will also be given to varicosities as the analysis has demonstrated an association with improved outcomes in certain cancers, which has not been identified previously. It is important to note that these are presented as hypothesis generating, for potential future research, rather than having a basis in the analysis results presented above.

Generic Aspects of Comorbidity

Although the underlying physiology of each comorbidity discussed is different, there are certain aspects of how comorbidity effects individuals, their decision making and their outcomes, which are relevant to many significant health conditions.

Previous research has suggested that patients with comorbidities have increased levels of toxicity from systemic anticancer treatment and surgery[20,136,302,303], however despite this, comorbid patients have still been shown to derive benefit from many of these interventions.[23,302–305] Underrepresentation of comorbid patients within randomised controlled studies[50,53,306], also limits data available for this group of patients, particularly with new and emerging treatments. This can make both patient and clinician decision making extremely challenging and in many cases based on opinion, rather than on objective data. As such, decision making in this context may result in patients deciding to accept shorter survival, for what they perceive to be a better quality of life.[14] For many, this balance of quality versus quantity of life, might be viewed as a preferred outcome. The analysis presented above, focusses only on survival and therefore fails to capture another end point which, to many, is equally, if not more important.

Patients with chronic health conditions engage and interact differently with health care services than those without pre-existing health condition. This can result in differences in the timing of diagnosis. Patients who are regularly reviewed, are more likely to have incidental findings identified on routine testing.[21,307,308] Patients with chronic health conditions, may also receive more encouragement to undergo screening and thus have a higher pickup rate of cancer.[133,309] Conversely, symptoms of cancer may be dismissed as being relevant to the comorbidity, and thus delay the diagnosis of cancer.[43,310,311] These differences may result not only in potential lead time bias, but also in differences in how advanced the cancer is, at the point of diagnosis. This may result in different survival outcomes, either with improved outcomes for early diagnosis, or worsened outcomes for late diagnosis.

As shown in chapter 3, although our analysis focusses on individual comorbidities, the real world is more complex, with patients often having multiple comorbidities. This is often driven through common risk factors such as obesity, smoking, diet and lack of exercise.[28,130] These risk factors are also found within the literature as being drivers for oncogenic change. If the cancer that results via these risk factors in some way differs from those that occur as a result of other processes, then it may follow that the behaviour of the cancer may in turn also be different. This may impact on cancer

growth, spread and treatment response, ultimately resulting in different outcomes for these patients.

The overlap of multiple health conditions may also significantly impact on suitability for treatment. Patients with multiple health conditions may be deemed too frail for treatment, or be treated in a fundamentally different way which impacts on treatment outcomes.[29,30,131,254,312] This difference in treatment may be entirely justified and appropriate, thus the comparison required in this instance might be between those with known health conditions who do not receive treatment, those who receive standard treatment and those who receive non-standard treatment, with survival and quality of life outcomes both considered. This may give a more representative view of which treatment route is most appropriate.

Hereon individual comorbidities are considered along with some of the research in that particular condition which may be relevant to survival outcomes.

Diabetes Mellitus

The analysis of DM across all cancers patients shows that on average, this group is associated with increased hazard. Despite this, when assessing DM on a cancer site basis, the scale, precision and direction of effect estimates differ, with 9 cancers showing associations with increased hazard and two cancers showing associations with improved hazard.

Although DM is a condition characterised by abnormal blood sugar homeostasis, this blood sugar regulation issue causes systemic effects. As such, diabetes is commonly the cause for other conditions driven by microvascular changes, including renal dysfunction, cardiovascular disease, neuropathy and stroke, to name but a few.[313] Thus patients with DM may have numerous physiological changes in other organs which impact on treatment suitability and dosing. This may impact on the management of these patients, such that their outcomes are different.[136]

Previous research has suggested that some of the treatments used in DM alter oncogenic risk. This is particularly relevant to the reduced hazard seen in liver cancer where patients on metformin and thiazolidinediones have a 70% reduction in risk of HCC compared to the background diabetic cohort risk.[314] Metformin use has been demonstrated to lower the cancer risk in a number of other cancer types.[270–272] In vitro and in vivo studies suggest that metformin has an inhibitory effect on cancer cell growth.[273,274] It is therefore possible that the improvements in survival seen in both lung and liver primary tumours, could be as a result of an effect of metformin on cancer cell growth. If this were the case, then it would suggest that lung and liver tumours are particularly affected by metformin's growth inhibitory effects when compared to other cancers, as it is these tumours that show an improvement in outcome.

Other commonly used medications in the diabetic population could also contribute to the effect seen. As diabetes is a risk factor for cardiovascular disease, a large number of patients within the diabetic cohort are likely to be prescribed aspirin in order to reduce the risk of myocardial infarction and cerebrovascular accidents. Research has demonstrated that aspirin may have anti-cancer effects across a range of tumour types.[266–268] The effects both reduced the chances of developing a range of cancers, but also reduce the risk of developing metastatic disease. Thus, the high prevalence of aspirin use in diabetic patients might account for the differences seen. As with metformin, the effect size of aspirin in liver and lung tumours would need to be greater than in other cancers otherwise, one would expect to see similar improvements in multiple cancer sites. The weight of this theory appears to be diminished by virtue of the fact that other conditions where aspirin is extensively used such as MI, stroke and peripheral arterial disease are not associated with decreased hazard, as is the

case in lung and liver cancer. It is however possible that aspirin is dampening down what might otherwise be larger effect sizes in these groups.

The underlying physiology of diabetes might also offer an additional explanation for the differences seen. Patients with type 2 diabetes have impaired tissue uptake of glucose, particularly in the liver, as one the mechanisms that leads to hyperglycaemia.[315] This may alter the glucose taken up within the cancerous area affected by liver tumours. If this were the case, then this might slow the growth of cancer within this area resulting in survival advantages.

In the context of primary liver tumours, cirrhosis may also be of relevance with regards to diabetes. Cirrhosis is a known predictor of HCC outcome with greater cirrhosis being associated with worsened outcomes.[316] Other risk factors for liver tumours such as viral hepatitis and alcohol excess are more strongly associated with cirrhosis, where diabetes often causes non-alcoholic fatty liver disease, but is less commonly associated with cirrhosis.[317] If a large proportion of the non-diabetic population is made up of cirrhotic patients, it may result in the appearance of improved outcomes. If however the group was split into cirrhotic and non-cirrhotic, it may be that the reduction is hazard is no longer seen when comparing diabetics to the non-diabetic, non-cirrhotic population.

Although there are a number of possible explanations for the population level survival advantage seen in diabetic liver cancer patients, further study is needed to identify if this is a true difference and if so, what is the cause of this difference. If the differences seen are due to diabetic pathophysiology or diabetic prescribing, then these could offer potential new avenues to explore in developing new treatment strategies for liver and lung cancer patients.

COPD:

The results from the analysis suggest that after adjustment for age, gender and deprivation, in most cancers COPD is associated with worse outcomes. Of the 16 cancer cohorts shown to have a high probability unidirectional hazard, 15 suggested worse outcomes associated with this comorbid group. The exception to this trend is in lung cancer, which is associated with a modest reduction in hazard of 9% (6-14%). This results is surprising, given that a previous research study using a peripheral vascular disease database, identified that COPD was associated with an increased risk of death from both lung and extra-pulmonary malignancy with a HR of 2.06 and 1.43 respectively.[318] Our results therefore suggest an example of a survival paradox where it appears patients with COPD are more likely to die from cancer and lung cancer in particular, but their risk of death once they develop lung cancer is lower. This could be explained by COPD increasing the risk of developing cancer to a greater extent than any protective effect of it, once the cancer develops. The increased incidence of lung cancer in COPD, is likely to be mediated through the shared exposure of smoking, with previous estimates showing lung cancer is five times more common in patients with objectively recordable airway obstruction.[319] An alternative explanation for the paradox identified is that there is a bias from the use of the vascular database in this previous study. The patients with COPD and vascular disease may in some way differ from the general population of COPD patients, however the increased cancer mortality trend has been previously demonstrated in other more general populations, that would be unaffected by this particular bias.[320]

As outlined above in the diabetes mellitus section, a number of medications have been investigated for their link to changes in cancer outcomes such as oral hypoglycaemics and aspirin. In the context of COPD, statins are associated with improved outcomes in patients with lung cancer[321]. Similar patterns have been seen in patients with COPD taking inhaled corticosteroids where non

corticosteroid users had a HR of 1.3.[322] This could also be a potential explanation for the modest reduction in hazard seen in the lung cancer population.

A number of biological mechanisms linking COPD to cancer have been identified and the risk differences seen, whether advantageous or disadvantageous, might be driven by these. Smoking leading to COPD has been shown to be driven by oxidative stress.[323] This same process forms part of the pathogenic process that links smoking to cancer, as in both instances damage to DNA is introduced.[324] This damage, in the form of point mutations, single strand breaks, double strand breaks and crosslinking, if incorrectly or inadequately repaired, result in somatic mutations which over time may accumulate and lead to cancer. Further, the nitrogen and oxygen radicals from smoking can denature proteins that lead to loss of tumour suppressor function directly.[325] Free radical exposure has also been shown to induce changes in cells that promote differentiation, proliferation and survival, all of which may also impact the behaviour of cancer cells and their growth.[326]

Beyond the link to smoking, COPD has also been shown to be associated with telomere shortening[327], which in turn reduces the time taken to reach cell senescence, increasing the risk of mutation accumulation. A number of genome wide association studies have identified inherited differences which increase the risk of both COPD and lung cancer.[328–330] These may in turn impact the behaviour of tumours that are driven by this mechanism. Epigenetic differences have been found in COPD and lung cancer, such that certain patterns of DNA methylation have been found more commonly in COPD patients with non-small cell lung cancer.[331–333]

Although not directly studied, the biological differences in COPD patients with cancer may in some way contribute to differences in outcomes. Cancer may develop, proliferate and progress in different ways, as well as responding differently to common treatments. As such, these biological differences could be the driver for differential outcomes in patients with COPD. Furthermore, the effects of smoking that relate to COPD and cancer biology are likely to be true also for other conditions where smoking is a significant risk factor such as cardiovascular disease and stroke.

Stroke:

Stroke is a clinical diagnosis where an acute vascular event within the central nervous system, results in loss of neurological function for greater than 24 hours. The precise definition is however different globally, with no agreed upon standard.[334] The term stroke encapsulates a number of distinct events including ischaemic stroke caused by a loss of blood supply and haemorrhagic stroke cause by an acute bleeding event. Our analysis demonstrates that where a high probability of unidirectional hazard is identified, the association was exclusively with increased hazard. Within the context of cancer there are a number of mechanisms by which stroke may be made more likely. Metastatic disease of the brain may result in the development of fragile, unstructured blood supply, which can be prone to bleeding.[335] Patients with cancer are also more at risk haemorrhage if they have abnormal blood counts due to bone marrow invasion or due to the side effects of systemic anticancer treatments.[336] Cancer and its treatment can be pro-thrombotic resulting in ischaemic stroke.[234] In view of this, the identification of stroke at or before cancer diagnosis may be an indicator of patients with more widespread disease. This would explain the adverse outcomes associated with these patients in the majority of cancers.

Where stroke occurs, whether cancer associated or not, it may leave patients with significant neurological deficits.[337] These deficits may impact on a patients overall fitness in a significant way. Patients with a lower performance status, may not be fit enough to undergo the same level of

intervention as patients with no history of stroke. As such, the frequency of interventions such as surgery, systemic anticancer treatment and radical radiotherapy may also be lower, resulting in worse outcomes.

The level of frailty may extend beyond the direct impact of stroke. The most common type of stroke is an ischaemic stroke. Typically, this is associated with risk factors for vascular damage including, hypertension, obesity, smoking, diabetes and hyperlipidaemia.[28,42,130] As such, patients may have features of other conditions that may impact on a patient's fitness for treatment, or level of appropriate dosing, in the case of systemic anticancer treatments. These common risk factors are also meaningful in terms of cancer risk.[338] As detailed above in the case of COPD, the mechanisms such as oxidative stress and DNA damage which can drive the oncogenic process, may result in cancer cells that behave differently from those which develop as a result of other exposures. This could results in variations in terms of growth, spread and treatment response.

To date most research in the area of stroke and cancer has focussed on whether cancer increases the risk of developing stroke after the cancer diagnosis[339,340], and on strokes as a presenting feature of cancer.[338,341,342] Little has been published on how patients with stroke prior to cancer diagnosis, differ from those without, in terms of mortality. Our results show that in majority of cancers, stroke was associated with worse outcomes when compared to those with no prior history of stroke. The literature provides a number of possible explanations for this relationship, however further study is required to understand whether the associations identified are true or artefactual, and if true, which mechanisms underlie the associations seen.

CCF:

The outcome associations seen in heart failure are more consistent than those seen with other comorbidity. All but 4 cancer sites were associated with an increase in hazard that had a high probability of being unidirectional. Previous research has identified links between the oncogenic process and heart failure, with patients known to have heart failure showing an increased risk of developing cancer.[343,344] Although many have advocated that this is due to the high level of monitoring in CCF patients simply picking up an additional common condition, some recent studies have suggested that there is a more fundamental link between them and that cancer may be considered a complication of CCF in some cases.[344]

Two main hypotheses have been put forwards to explain this increase in cancer incidence. The first is that common risk factors exist for both heart failure and cancer.[345,346] As a result, patients are more likely to develop one if they have developed the other already. These factors may include genetic predisposition, other comorbidities such as diabetes and physiological changes such as increased oxidative stress.

Another possibility is that CCF is directly oncogenic. This has been evidenced through studies showing altered neuro-hormonal activation in the renin angiotensin aldosterone system, which in turn has been linked to tumorigenesis. Additionally animal models have demonstrated that post-ischaemic heart failure changes, resulted in the secretion of factors which enhance colon cancer cell growth.[347] If these differences in underlying physiology occur, then it may be that cancer in CCF patients behaves differently as a result, which may account for survival differences.

It is also possible that as with other comorbidities, frailty, multimorbidity and overall fitness impacts on patient management which may in turn alter outcomes. A particular consideration with CCF in this regard is that patients commonly are unable to lie flat without becoming breathless. This can in

severe cases be a barrier to radiotherapy treatment where patients are required to lie flat for several minutes at a time.

Myocardial Infarction (MI)

As with the other comorbidities discussed thus far, previous research has highlighted an increased risk of developing cancer in patients who have had a previous MI.[348] A number of possible explanations for this have been postulated including common risk factors, MI being a presentation of occult cancer and patients being subjected to increased clinical surveillance, increasing the rate of diagnosis. Studies focussing on MI in patients with a known active of previous cancer diagnosis, have shown that outcomes in the short term are the same as those without cancer, however when looking at 1 year post MI the hazard ratio was estimated to be 2.52.[349]

More recently, research in breast cancer has provided evidence to suggest that changes in systemic homeostasis, mediated via the immune system after an MI, alters tumour behaviour.[350] In mouse models the study identified that MI created an acute stress response, which altered the central regulation of the innate immune system. This results in epigenetic changes within the bone marrow which shifts the immune system towards an immunosuppressive state. The immune changes result in increased tumour growth. Studies in patients with cancer who developed later MI identified that these patients in early breast cancer had an increased risk of developing metastatic disease. This provides a potential physiological explanation for the results demonstrated within our analysis. In almost all cancers prior MI was associated with worse survival outcomes. Further study is therefore needed to identify whether the mechanisms seen within the breast cancer setting also affect other tumour types. Additionally further investigation is needed to attempt causal analysis within the context of MI and cancer outcomes.

Varicosities

The degree of effect seen in the hazard reduction for varicosities is sizable with a 37% (14-55%) reduction in hazard in colorectal cancer and a 31% (3-51%) reduction in breast cancer. Although both of these are clinically relevant, the precision of estimates is fairly low. As discussed previously, the association could be explained by a number of methodological or behavioural reasons such as lead time bias[275], collinearity[351] with other variables, confounding[165] and differences in health seeking behaviours.[245,352] There are however some potential biological explanations which might suggest that there is a physiological link between the development of varicosities and improved cancer outcomes. Previous research in patients with varicose veins has attempted to identify changes in the expression and production of vascular growth factors.[353,354] One study looked at levels of Vascular Endothelial Growth Factor (VEGF) production in artificially induced venous stasis, comparing those with and without varicose veins.[353] This identified that in control patients, artificial venous stasis induced via a 90mmHg cuff applied to a lower limb, induced increases in VEGF levels. No such rise was seen in patients with varicosities.[353] This suggests that patients with varicose veins, may have a degree of impairment in their ability to produce VEGF. This difference may be of importance as in a number of cancers VEGF plays a key role.[354–356] Two such examples are breast and colorectal cancer.

In the context of colorectal cancer angiogenesis has been shown to be an important factor in progression. Evidence from both preclinical and clinical studies have highlighted VEGF as the key contributor to angiogenesis in colorectal cancer with expression occurring in approximately 50% of colorectal cancer patients.[356] Studies looking at the associations between VEGF and colorectal cancer outcomes have demonstrated that with increasing disease stage there is also an increase in the

prevalence of VEGF expression. VEGF expression has been found to be an important predictor of disease specific survival at ten years, with better outcomes seen in those patients with lower levels of VEGF expression, along with reduced rates of progression after treatment. The importance of VEGF is also highlighted through the use of anti-VEGF treatments in colorectal cancer.[357]

In the context of breast cancer, the importance of VEGF is less clear. High levels of VEGF expression have been demonstrated in the most aggressive forms of triple negative breast cancer, whilst also contributing to the mechanisms driving metastasis.[358] Additionally, VEGF levels are higher in patients with malignant breast disease compared to benign breast disease.[359] The same study however demonstrated that VEGF was correlated with oestrogen receptors and inversely correlated with disease stage.

If patients who develop primary varicose veins, do so in part because of altered ability to produce VEGF, then the reduction in VEGF production may be having a direct biological effect on the tumours these patients develop. This could result in reduced tumour angiogenesis, resulting in reduced growth rates, lower levels of metastasis and improving outcomes. A literature search identified no previous analyses demonstrating this identified association, however the underlying biology of the two conditions suggests a plausible explanation for the improved survival outcomes seen. Further research is therefore needed to assess potential causal relationships between varicosities and cancer outcomes.

5.4.3 - Precision and Clinical Relevance:

When using observational data it is often regarded as inappropriate to apply statistical significance tests due to the two compared populations not truly being drawn from the same population.[59,61] As a result the focus of results is often placed on the clinical significance of the results.[360] The interpretation of this can often be based on the point estimate of an analysis however in the absence of statistical tests, the confidence interval can be used as a guide as to where there is a high probability that the true point estimate will lie. The span of this confidence interval can be used as a guide to the reliability of the results obtained as a measure of precision. More precise estimates will have narrower confidence intervals, where less precise ones have wider confidence intervals. This therefore raises an important question about what level of precision one would wish to have to determine that the results were meaningful and reliable. Two different approaches could be used. In one it could be based on a fixed raw value for example only those with a confidence interval which spans less than 0.5 hazard ratio is deemed to be precise enough to suggest reliable results. Alternatively the span could be based on the point estimate itself. An example here could be that the span must not exceed 25% of the point estimate. The latter has the advantage of taking into account that precision with small effect sizes is probably more important than where large effect sizes are seen. It does however create the issue that very low precision could still be allowed in the presence of a large effect size. An alternative here would be to combine the two such that it must fulfil both criteria.[60] This allows a minimum precision to be set whilst allowing for some accommodation of a need for greater precision with small effect sizes.

As with a 95% confidence level with statistical significance testing, the threshold to be applied in precision is an entirely arbitrary one. Unlike with confidence tests, the choice of cut of does not allow any estimation or misclassification error. In clinical practice a difference in hazard of 50% would be deemed clinically relevant however getting estimates to this level of precision may be challenging unless extremely large numbers are present. Even when large numbers are present, where a class imbalance occurs, the precision levels may still be low, as one of the two groups still has small numbers. This problem becomes more acute with an increasing precision threshold. Within

our results we have applied a 25% and 0.25 hazard ratio precision cut off for **Table 10**. This has been applied to all analyses, and yielded a small number of results. By applying our filter for unidirectional hazard and precision cut offs we can identify comorbidities that have strong evidence for association in one direction that are also precise. The subsequent task is to determine if those differences are clinically relevant. It is important to note however that shifting the precision threshold has a profound impact, moving from 0.25 hazard ratio span to 0.5 increased the number of results meeting the cut offs by 54. There is an argument for saying that applying a blanket rule is unhelpful as the required precision will differ depending on the question being asked and the precise clinical scenario. Not having a fixed cut off up front however poses the potential risk of allowing results to be significantly impacted by researcher bias, with a post hoc justification applied for allowing one result over another.

The impact of cohort size is highlighted by a number of features in the results. Firstly the greatest numbers of comorbidities that meet the precision thresholds are when assessed in the all cancer cohort. Amongst the site specific cohorts, the only cancer not in the top 4 to have any models meeting the precision threshold is skin cancer which is the next most common cancer. Although the relationship between precision and cohort size is unsurprising, it is important to consider that those failing to meet this threshold do not need to be discounted, however require a greater degree of scepticism when using the point estimate values.

Overall the effect sizes where there is also high precision are large. Defining what is clinically relevant is a somewhat subjective process and there is no commonly accepted threshold. This variation occurs in part because of different perceptions of what a meaningful improvement in survival is. In some cases it would be balanced against cost, as is the case for National Institute of Clinical Excellence assessment. When assessing the site specific cohorts the smallest effect size is seen in hypertension in lung cancer with a between 1 and 13% difference in hazard. At 1 % this is unlikely to meet the threshold of significance by most clinical standards. At 13% however, this would be seen by many as highly relevant. Applying a general rule of thumb of 5% suggests that all other comorbidities meeting the high precision threshold are also clinically relevant, even if using the smallest hazard difference identified by its confidence interval.

5.4.4 - Variable Inclusion

These analyses build upon the results of the initial associations seen in chapter four with the KM approach. The introduction of a multivariable approach allows us to adjust for factors which could be considered to be confounders of the results initially seen. Due to known associations between age, gender, deprivation and population level all cause survival these are potential sources of confounding.[308,361] The identified correlation between many of the comorbidities analysed and these factors, (as seen in chapter two) adds weight to the rationale for their adjustment within the model being required. There are however a number of additional parameters that one could consider adjusting for that were not included in the models developed within the above analyses. These include stage, grade and histology. Clear links have been shown in the literature between increasing grade and stage, and decreased survival in the context of most cancers.[23,43] Additionally different histological types of cancer even in one anatomical site have very different survival trajectories such as small cell lung cancer versus non-small cell lung cancer where the five year survival for the former is 6% and 24% for the latter.[362] It could therefore be argued that not including these potentially leaves unresolved confounding within the generated estimates.

A counter argument to this is however that in the case of most comorbidities studied, links have been found between the presence of comorbidity and differences in grade, stage and histology at

presentation. As our definition of comorbidity is predicated on the comorbidity being temporally precedent to the cancer diagnosis, then any associations between comorbidity and difference in grade, stage and histology are in fact part of the mechanism causing differences in survival outcomes and should be included in the estimate by not adjusting for them. The issues around adjustment for staging data have been highlighted previously in causal inference research which has demonstrated that inclusion of this introduces a significant bias and potentially results in worse estimates of the true effect size seen.[363]

5.5 - Summary

The results presented in this chapter draw similar conclusions to much of the previous research into comorbidity. In general chronic health problems prior to a cancer diagnosis are associated with worse outcomes. Despite this a number of conditions are shown to be associated with improved outcomes. The findings seen, whether suggesting a survival advantage, or survival disadvantage, could be due to various forms of bias and the application of causal inference approaches shows further areas of complexity around how the analysis approach hampers interpretation of findings. The analysis of underlying model assumptions highlights some specific examples where violations have occurred. In such cases further research may be needed to apply alternative methods to overcome these potential limitations. Despite the highlighted analysis limitations, a number of plausible biological mechanisms have been identified in the literature that could offer explanations for the different survival outcomes seen. Across each of these analyses however, no distinction is drawn between cancer and those from non-cancer related deaths. To further explore this concept analyses in the subsequent chapter will apply cause-specific competing risk methods to estimate the cancer-cause specific hazard effects associated with comorbidity.

Chapter 6 – Inferential Analysis of the Association between Cause-Specific Cancer Risk and Comorbidity in Cancer Patients

6.0 - Introduction

The analyses presented thus far that have focussed on survival outcomes, have utilised overall all-cause mortality as their end point. This provides insight into the outcomes for patients from any cause of death. As discussed in chapters 2, 4 and 5, where multiple conditions co-exist, each with an independent mortality effect, it is likely that patients with both conditions will have an overall mortality that exceeds patients with any one of those conditions in isolation. An alternative approach would be to look at the impact of comorbidity on a specific cause of death, such as cancer related deaths.[180,181,364] As discussed in chapter 2, this can be approached in a number of ways, however given the nature of the questions being asked, a cause-specific approach is the most appropriate due to the focus on relative hazards as opposed to cumulative incidence.

This chapter builds on the work presented in the previous chapter, to conduct the same series of analyses, however using the altered end point of cancer cause-specific hazard. As with the previous chapter, models will be generated, assessed for model assumptions and then interpreted in terms of the principles of inferential analysis detailed in chapter 2. Comparisons between the results for all-cause and cause-specific mortality will also be made to identify commonalities and discrepancies between findings. As with chapter 5, the focus will remain on the hazard attributed to comorbidities with other covariates included for the purpose of controlling for confounding, as opposed to for direct interpretation.

Despite the interplay between comorbidity and cause-specific mortality having been researched previously in the literature, this has primarily focussed on composite measures of comorbidity.[365–367] The analysis presented below, therefore represents one of the largest and most comprehensive assessments of the associations between individual comorbidities and cancer-cause specific hazard undertaken to date.

6.0.1 - Aims and Objectives

Aim

1. Quantify the association between comorbidity and cancer cause-specific mortality in cancer patients using a multivariable approach.

Objectives

a. Develop Cox PH models for cancer cause-specific mortality, with each combination of cancer and comorbidity whilst adjusting for age, gender and deprivation as appropriate.
b. Test for violations of underlying model assumptions.
c. Identify outlier or influential patients within the dataset.
d. Quantify the effect of influential outliers on hazard estimates.
e. Quantify hazard associated with comorbidity.
f. Assess the confidence interval for precision.
g. Assess confidence interval for high or low probability of unidirectional hazard.

6.1 - Methods

6.1.1 - Data Pre-processing

To enable analysis of the data some minor pre-processing was required. A cause-specific death label was generated, here any patient with a death was categorised as a cancer related death or a non-cancer related death. Cancer related death was defined as a death with a cancer "C" ICD-10 code in section 1a, 1b or 1c or the patient's death certificate.[183] Mention of cancer in section 2 was not included in the definition. All patients without a "C code ICD-10 code were classified as a non-cancer death. In cases of non-cancer death, a patient's survival status was altered such that they were regarded as censored rather than deceased. Patients who were known to be dead and had no cause of death data were excluded from the analysis. Thus our analysis included patients who were alive or who have died with a documented cause of death.

To quantify and explore the data on those without a cause of death the number and percentage of each cohort that were removed was calculated. Additionally, analysis in the all cancer cohort was undertaken to look at the number and proportion of cases by year of death with missing cause of death data to identify any patterns over time.

6.1.2 - Building of Models, Testing Assumptions and Quantification of Associations

A similar approach was applied to building survival models as those presented in chapter five. Cox proportional hazard models were generated as a function of an individual comorbidity, age, gender (where appropriate) and deprivation. The updated survival status, based on cause of death was utilised to generate a cause-specific quantification. As with the previous chapter, the focus of analysis is on the impact of comorbidity and thus results are predominantly presented based on the coefficients of each individual comorbidity. The inclusion of age, gender and deprivation was to control for confounding rather than for direct assessment of effect. The covariate model specification for these analysis were identical to those outlined for chapter 5 and in the appendix. The difference is that they are expresses as a function of cancer cause-specific survival rather than all-cause survival.

The same approach for assessing proportional hazards, linear assumptions and outliers were employed as those outlined in chapter 5. Quadratic transformation of variables found to have non-linear relationships was not attempted due to its previously demonstrated deleterious effects on precision, which will be amplified due to the higher levels of censoring in this cause-specific approach. Point estimates and confidence intervals for hazard ratios were derived from these models and used to assess, effect size, consistency of direction of effect, precision and clinical relevance.

As with chapter 5 the analysis will focus on all comorbidities in breast, lung, prostate and colorectal cancer. Additionally analysis will cover all cancers in CCF, MI, stroke, COPD and diabetes mellitus.

6.1.3 - Comparison to All-Cause Results

To facilitate interpretation of results, the cause specific mortality results will be compared to those of the all-cause mortality. Results where both the all-cause and cause specific mortality show a high probability of unidirectional hazard and the same direction of effect will be highlighted and compared. This will be conducted for the hazard results and additionally the precision analysis.

6.2 - Results

6.2.1 - Data Pre-processing and Exploration

Figure 56a shows the trend in the number of patients from the All Cancer Cohort with a recorded death, but no recorded cause of death annually. **Figure 56b** represents the same data but as a proportion of the deaths in this cohort each year. This shows no clear pattern of association between having a missing cause of death and year of diagnosis up until 2017 where there is a sudden rise in missing data peaking in 2018 (by number) but remaining at 100% of deaths from 2018 onwards. **Table 11** demonstrates the proportion of each cohort where cause of death data is missing. On average, this represents around 1 in 10 patients (10.8%), however the most missing data is found in liver malignancy and the least in testicular cancer.

Site	Cause of Death Complete (n)	Missing Cause of Death (n)	Total	Missing Cause of Death (%)
Breast	29058	2655	31713	8.37
Prostate	21785	2771	24556	11.28
Lung	26718	3350	30068	11.14
Bowel	20135	3006	23141	12.99
Colorectal	18618	2785	21403	13.01
Melanoma	6308	617	6925	8.91
Skin	21725	3685	25410	14.5
Connective	1589	208	1797	11.57
Lymphoma	7243	819	8062	10.16
Kidney	4396	606	5002	12.12
Laryngeal	1851	246	2097	11.73
Brain	3059	349	3408	10.24
Intracranial	3203	362	3565	10.15
Bladder	6458	1016	7474	13.59
Pancreatic	3823	607	4430	13.7
Leukaemia	4777	548	5325	10.29
Endometrial	5132	634	5766	11
Testicular	2347	61	2408	2.53
Oesophageal	5459	820	6279	13.06
Ovarian	3333	419	3752	11.17
Stomach	4320	391	4711	8.3
Liver	2999	651	3650	17.84
Myeloma	2333	384	2717	14.13
Thyroid	2390	108	2498	4.32
Cervical	2909	185	3094	5.98
CUP	4431	391	4822	8.11
All	198031	24405	222436	10.97

Table 11: Summary of Completeness of Cause of Death Data by Cohort *– Breakdown of each Site Specific Cohort and the All Cancer Cohort by the completeness of cause of death data for patients recorded as being deceased. Results are expressed as; number with cause of death data, number missing cause of death data, the total number of patients within the cohort known to be deceased and the percentage of the cohort where the cause of death data is missing.*

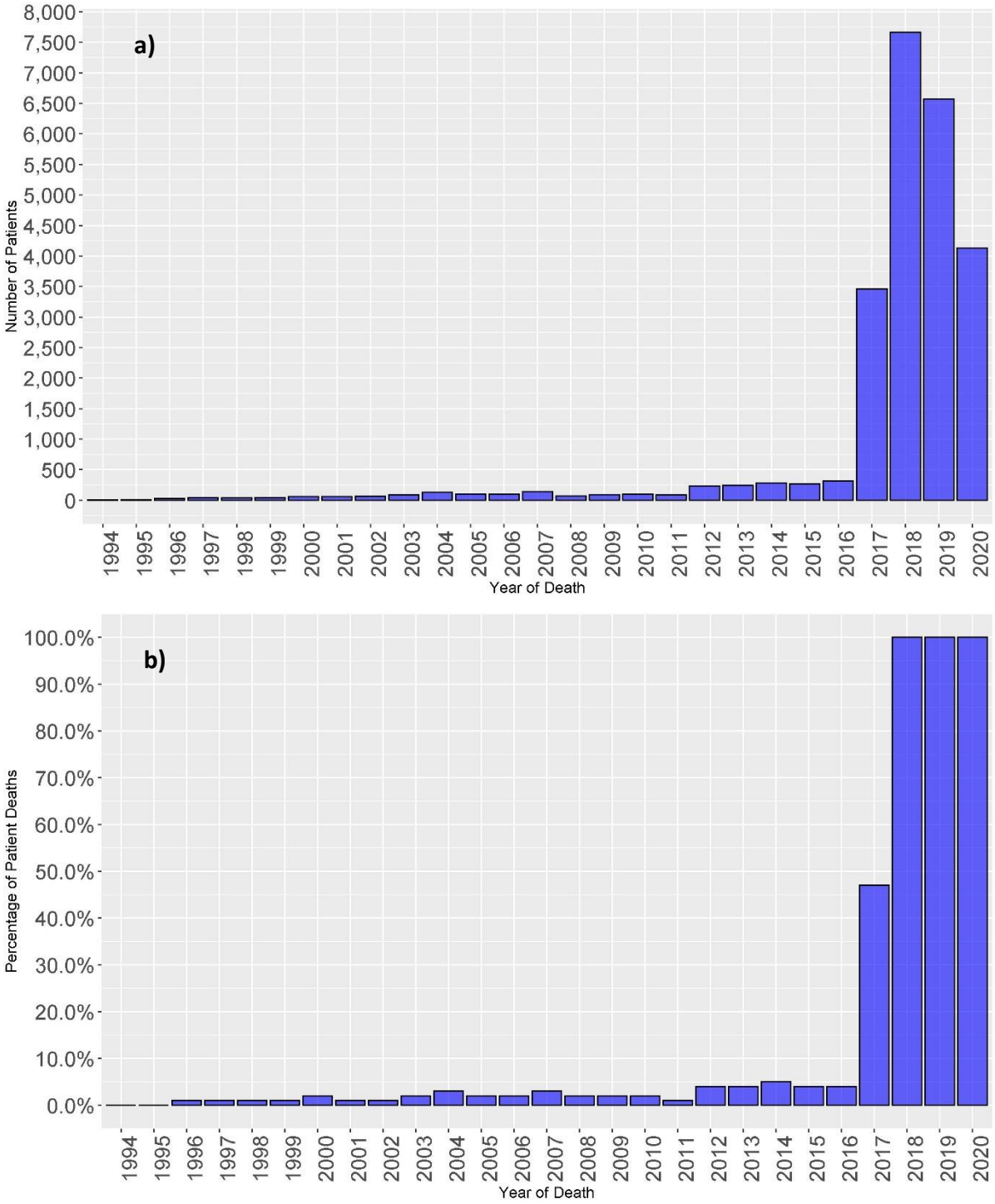


Figure 56: Missing Cause of Death Data in All Cancer Cohort *–Summary of a) Number of patients and b) Percentage of deaths each year where the cause of death data was unavailable for patients within the All Cancer Cohort*

6.2.2 - Model Assumptions

Proportional Hazard Assumptions

All of the models were manually reviewed using their Schoenfeld residual plots. None were found on this graphical review to demonstrate evidence of proportional hazard violations. As with the all-cause overall survival, there were some examples where the cohorts had small numbers, visual review demonstrated no evidence of a proportional hazard violation, but the numerical analysis suggested that a violation had occurred. An example of gender in the CCF and thyroid cancer model is presented below to illustrate this (**Figure 57**).

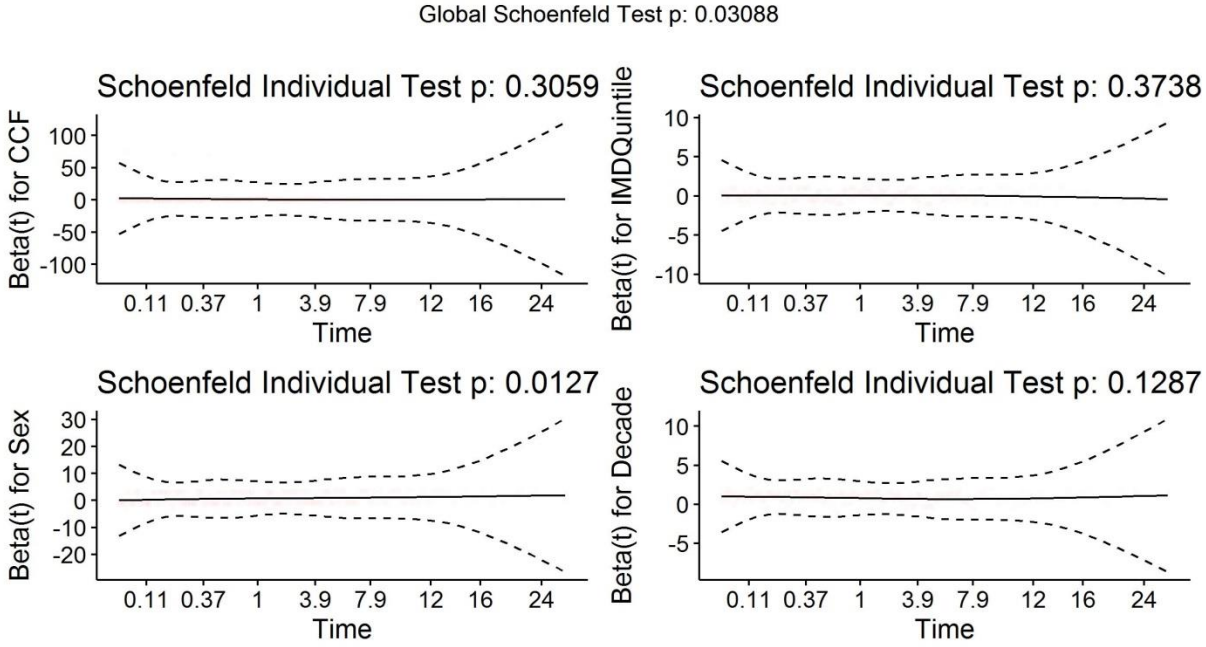


Figure 57: Shoenfeld Residual Plot for Covariates in the CCF in Thyroid Cancer Cause Specific Cox Model – *The Shoenfeld residuals for each variable are plotted on the vertical axis and time is plotted on the horizontal axis. Each faint red dot represents one patient at that time within the Thyroid Cancer Site Specific Cohort. The black line is the line of best fit and the dotted lines the confidence interval. When the black line is horizontal and almost entirely straight there is no evidence for a violation of the proportional hazard assumption if the line is angled or showing multiple areas where the line is curved this provides evidence for a breach of the proportional hazard assumption.*

Linear Assumptions

Visual analysis of the Martingale residuals for continuous variables showed evidence of multiple models with non-linear relationships for both age and deprivation quintile. Leukaemia was the only cohort to show evidence of non-linearity in both age and deprivation. In several cases the deviation from a linear relationship was substantial, as shown in **Figure 58**. In all cases of nonlinearity some non-linearity continued when assessing the variable on different scales. The summary results can be found in **Table 12.**

Site	Age	IMD
All	1	0
Bladder	0	0
Breast	1	0
Cervical	0	0
Colorectal	1	0
Connective	1	0
CUO	0	0
Endometrial	0	1
Intracranial	0	1
Kidney	0	0
Renal	0	0
Laryngeal	0	0
Leukaemia	1	1
Liver	0	1
Lung	0	0
Lymphoma	0	1
Melanoma	0	0
Myeloma	0	0
Oesophageal	0	1
Ovarian	0	1
Pancreatic	0	0
Prostate	0	0
Skin	1	0
Stomach	0	1
Testicular	1	0
Thyroid	1	1

Table 12: Summary of Cause-Specific Linear Assumptions Assessment *– The plots for the Martingale residuals for each continuous variable in each cause-specific Site Specific Cohort and All Cancer Cohort was reviewed for evidence of a violation of the linear assumptions. The results are recorded win the table with 0 = No evidence of Violation, 1 = Evidence of Violation*

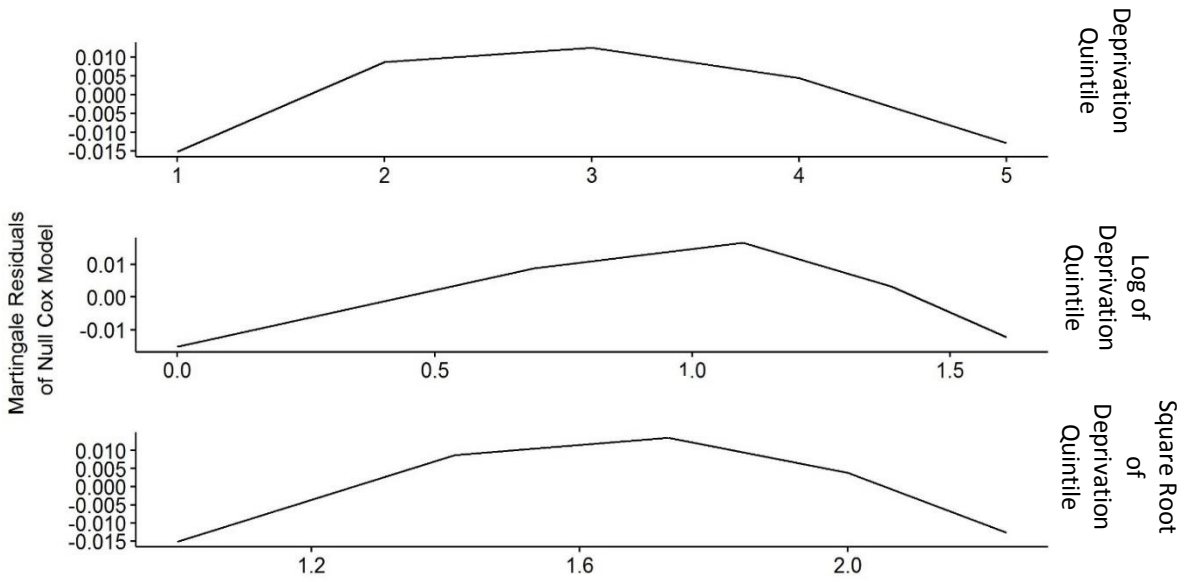


Figure 58: Assessment of Linear Assumptions of Deprivation Quintile in Thyroid Cancer Site Specific Cohort for Cancer Cause-Specific Analysis *- Martingale residuals are plotted on the vertical axis against age on the horizontal axis. Three plots represent deprivation quintiles on different scales, the top plot represents raw deprivation quintile. The second plot represents deprivation as the log of deprivation quintile and the third plot represents age as the square root of deprivation. Where linear assumptions hold the line should be a continuous straight line. Curves or changes in the trajectory of the line suggest non-linearity.*

Influential Outliers

Influential outlier values were only identified in relation to gender and specific comorbidities. No influential outliers were found for deprivation or age. The results identified that diabetes did not have any influential outliers in any of the four common cancer sites. Stroke, CCF, COPD and MI had influential outliers in both breast and prostate cancer cohorts. No influential outliers for any comorbidity were found in lung cancer patients and only stroke was found to have influential outliers in colorectal cancer. The results for the hazard ratios with and without the outlier cases can be found in **Figure 59** and **Figure 60**. The breast cancer results highlight that in the case of COPD and MI the removal of outliers resulted in hazard ratios moving from a low probability of unidirectional effect to a high probability of unidirectional effect. The analysis within the prostate cancer cohort identified that CCF, COPD and MI moved from a low probability of unidirectional hazard to a high probability with the removal of influential outlier. The remaining comorbidities demonstrated a consistent direction and probability of unidirectional effect.

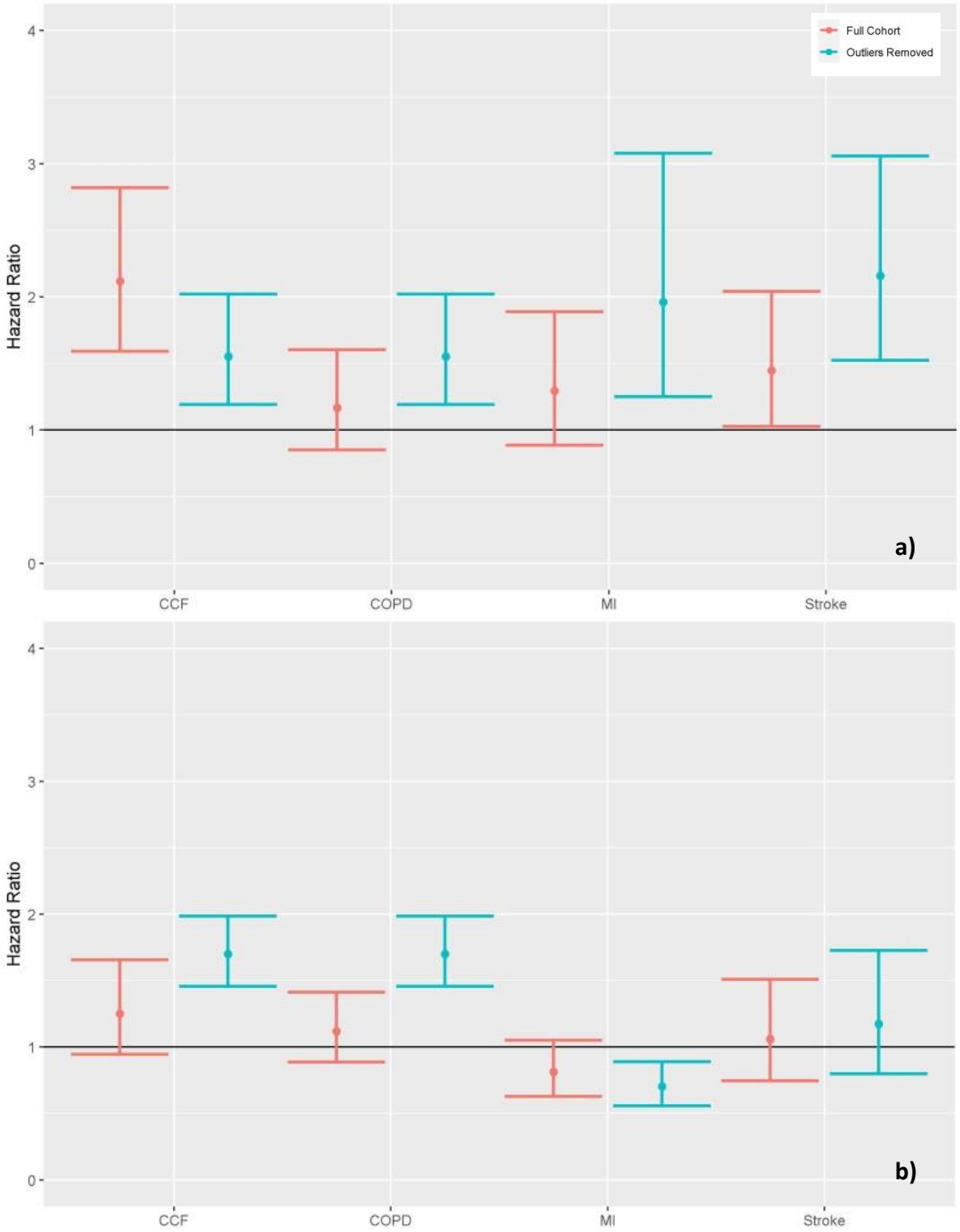


Figure 59: Effect of Influential Outliers on Breast and Prostate Cancer-Cause Specific Hazard – *Cancer Cause-Specific Cox derived comorbidity hazard ratios for comorbidities with influential outliers identified in a) Breast Cancer and b) Prostate Cancer Site Specific Cohorts. Results in red are the hazard ratios obtained when the full site specific cohort was used to generate a Cox model those in blue are where the model was generated after the removal of influential outliers.*

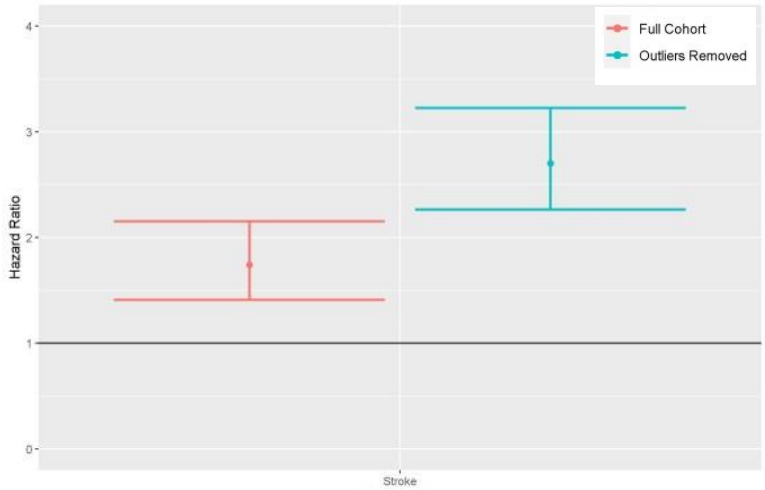


Figure 60: Effect of Influential Outliers Colorectal Cancer-Cause Specific Hazard – *Cancer Cause-Specific Cox derived comorbidity hazard ratios for comorbidities with influential outliers identified in Colorectal Cancer Site Specific Cohorts. Results in red are the hazard ratios obtained when the full site specific cohort was used to generate a Cox model those in blue are where the model was generated after the removal of influential outliers.*

6.2.3 – Cancer Cause-Specific Hazards

Breast

Of the 40 comorbidities assessed only 9 were identified as having a high probability of unidirectional hazard. Of these 8 were associated with an increased hazard and one with decreased hazard. The decrease in hazard was in varicosities with a hazard ratio of 0.52 (0.31-0.71). Dementia was associated with the largest increase in hazard with a hazard ratio of 3.02 (2.14-4.27%) increase. CCF was also associated with a more than doubling of hazard with a hazard ratio of 2.18(1.59-2.82). Arrhythmia shows a more modest effect size, but better precision with a hazard ratio of 1.43 (1.16-1.76%) increase in risk. Chronic kidney disease was associated with a 97% (37-184%) increase in risk.

Colorectal

In the colorectal cancer cohort 15 comorbidities were associated with a high probability of unidirectional hazard. Of note 6 were associated with decreased hazard including asthma, hyperlipidaemia, hypertension, MI, obesity and varicosities. Of these, the largest reduction in hazard was for varicosities with a hazard ratio of 1.53 (1.24-1.71). The largest increase seen is in the case of dementia with a hazard ratio of 2.46 (1.93-3.15%) increase in hazard. CKD, CCF, COPD, MI, paraplegia, stroke and venous thromboembolic disease were also associated with increased hazard.

Lung

18 comorbidities showed a high probability of unidirectional hazard. Of these 8 were a reduction in hazard and 10 were an increase. The largest decrease in hazard was for obesity with a hazard ratio of 0.33(0.29-0.38). Malabsorption also showed a large decrease with a hazard ratio of 0.66 (0.33-0.96) however the precision was low. The largest increases were seen with HIV and demyelination however due to low patient numbers the precision of these estimates was low. Of those with more precise estimates thromboembolic disease was the largest increase with a hazard ratio of 1.47 (1.27-1.71%) increase in hazard.

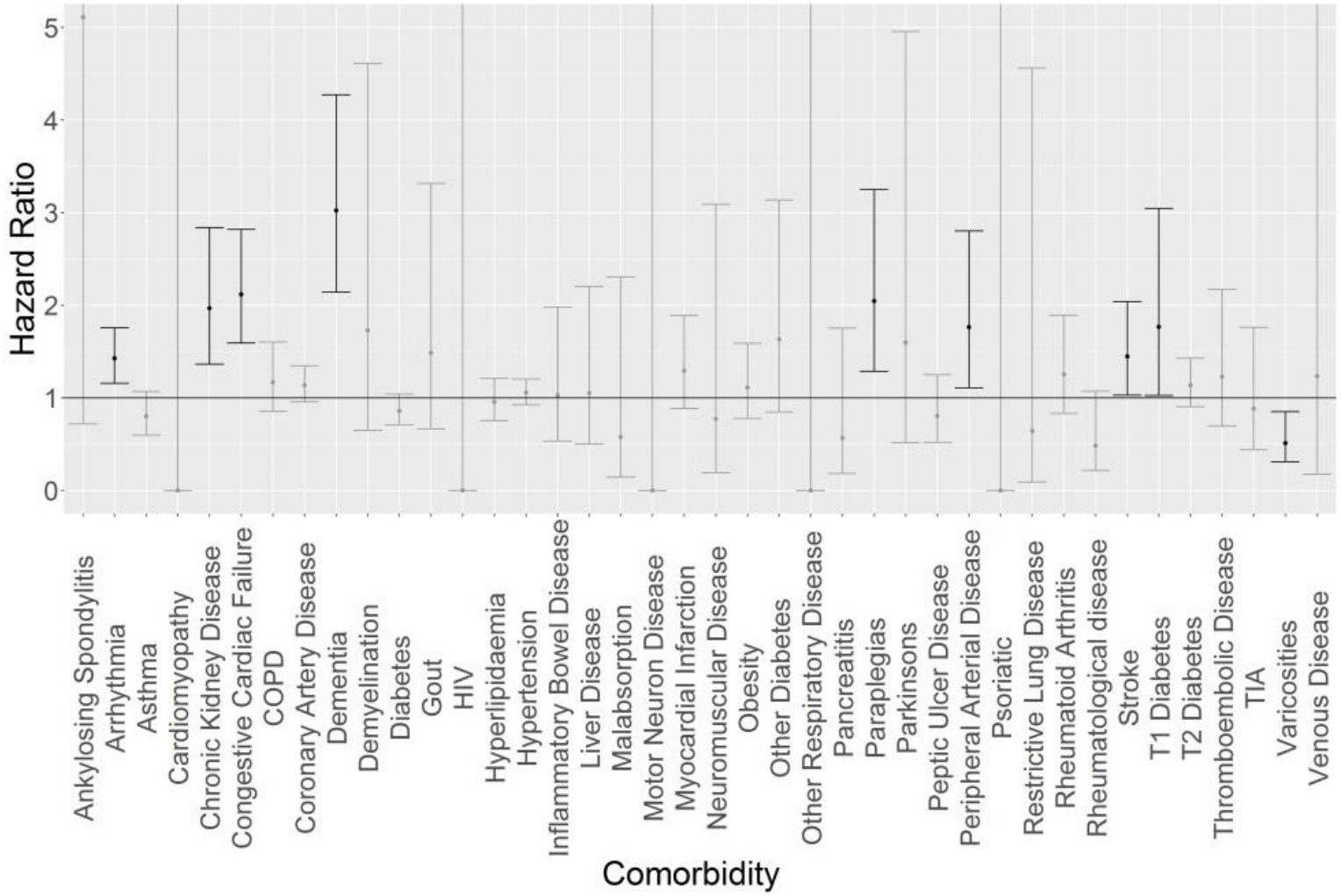


***Figure 61*: Cox Derived Cancer Cause-Specific Hazard Ratios for Comorbidity in Breast Cancer** – *Cox derived cancer cause-specific hazard ratios with 95% confidence interval associated with each comorbidity of interest in the Breast Cancer Site Specific Cohort. Each estimate is derived from a standalone model adjusting for age, gender and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

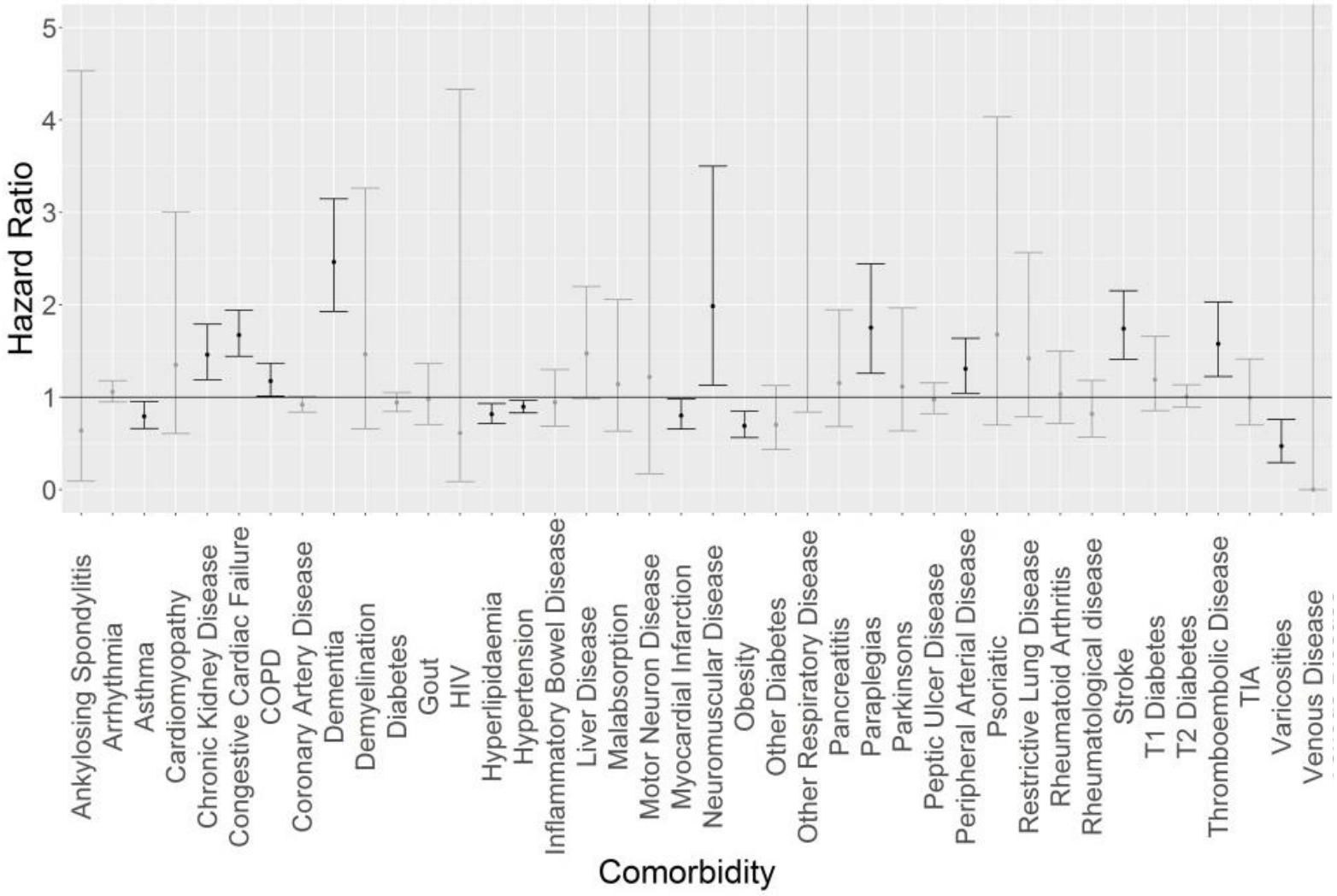


Figure 62: Cox Derived Cancer Cause-Specific Hazard Ratios for Comorbidity in Colorectal Cancer *– Cox derived cancer cause-specific hazard ratios with 95% confidence interval associated with each comorbidity of interest in the Colorectal Cancer Site Specific Cohort. Each estimate is derived from a standalone model adjusting for age, gender and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

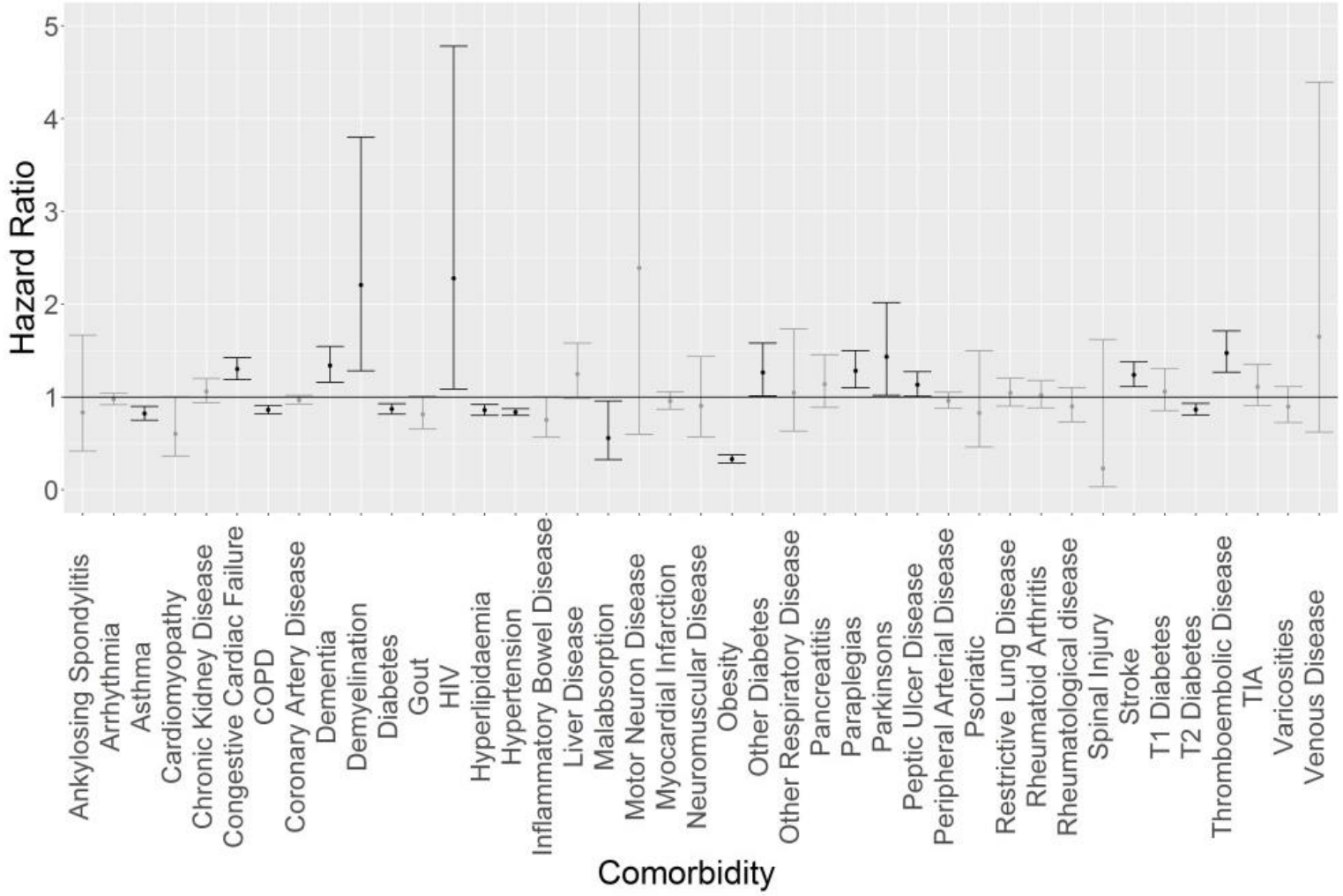


Figure 63: Cox Derived Cancer Cause-Specific Hazard Ratios for Comorbidity Lung Cancer *– Cox derived cancer cause-specific hazard ratios with 95% confidence interval associated with each comorbidity of interest in the Lung Cancer Site Specific Cohort. Each estimate is derived from a standalone model adjusting for age, gender and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

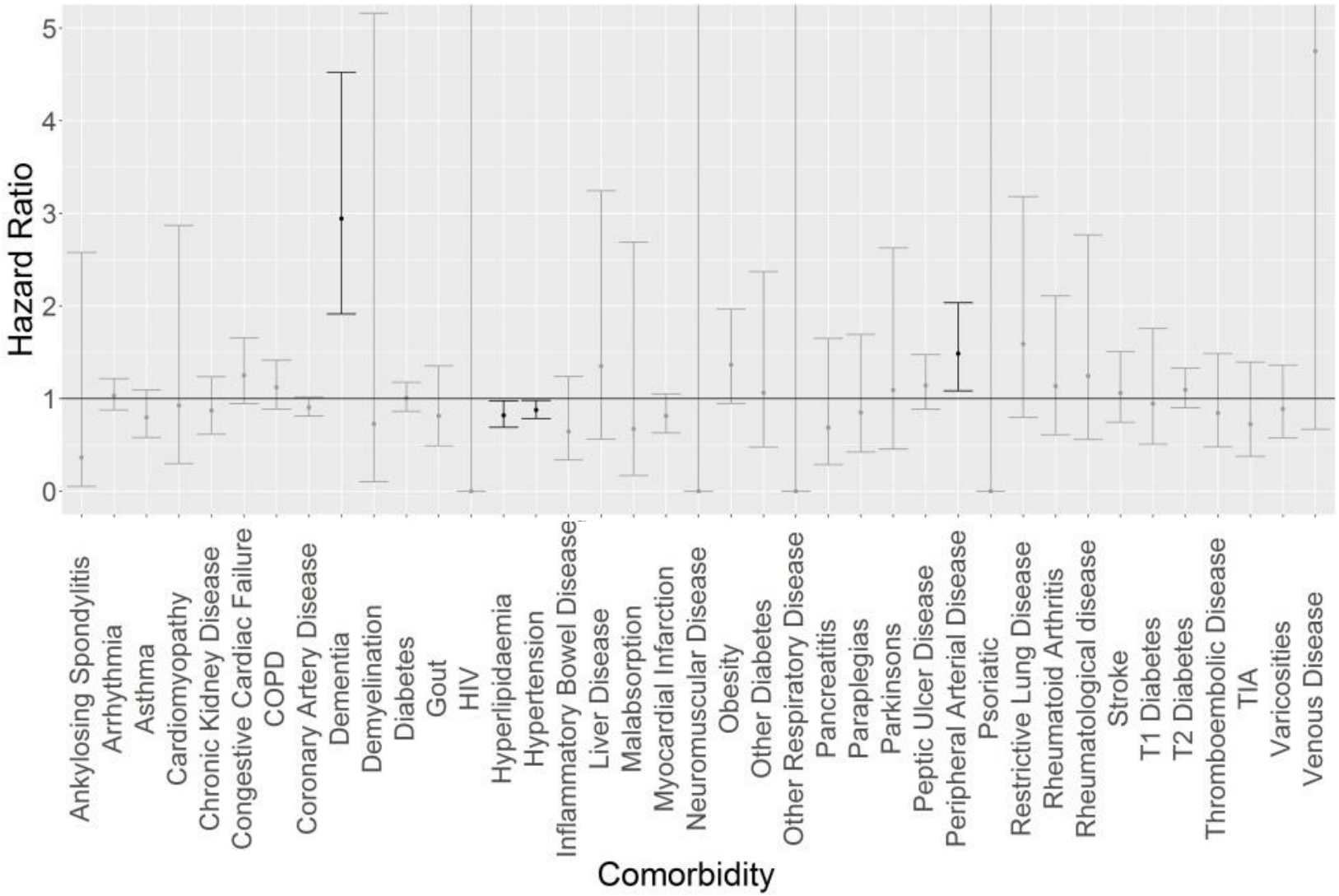


Figure 64: Cox Derived Cancer Cause-Specific Hazard Ratios for Comorbidity Prostate Cancer *– Cox derived cancer cause-specific hazard ratios with 95% confidence interval associated with each comorbidity of interest in the Prostate Cancer Site Specific Cohort. Each estimate is derived from a standalone model adjusting for age, gender and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

Prostate

Only 4 comorbidities were associated with a high probability of unidirectional hazard. Dementia and peripheral arterial disease are associated with an increase in hazard with a hazard ratio of 2.94 (1.91-4.52) and 1.48 (1.08-2.03) respectively. Hypertension and hyperlipidaemia were associated with modest decreases in hazard with a hazard ratio of 1.12 (1.02-1.22%) and 1.28 (1.02-1.32) respectively.

CCF

CCF was associated with a high probability of unidirectional hazard in 9 cancer sites. All cases were increased hazard risk although results had low precision in all but lung cancer. Here it was associated with a 1.30 (1.19-1.42) hazard ratio.

COPD

Eight cancer sites were associated with a high probability of unidirectional hazard including colorectal, CUP, leukaemia, lung, oesophageal, pancreatic, skin and stomach cancer. Of these, lung was associated with a reduced hazard and the others an increased hazard. All of the associations with increased hazard also had moderate to low precision. The largest effect size was in skin cancer with 1.92 (1.50-2.46) hazard ratio. COPD in lung cancer had a high precision with a 1.14 (1.09-1.18) hazard ratio.

Diabetes

Seven cancer sites were associated with a high probability of unidirectional hazard when assessing diabetes. Six of these were reductions in hazard which was the case for bladder, renal, liver, lung, ovarian and stomach cancer. Of these the largest decrease is in ovarian cancer with a hazard ratio of 0.69 (0.49-0.96). Cervical cancer was associated with an increased hazard with a hazard ratio of 2.22 (1.34-3.63) although the precision of this estimate was low.

MI

Three cancer sites were associated with a cause specific high probability of unidirectional hazard. Increased hazard was seen in CUP and liver cancer with a hazard ratio of 1.42 (1.09-1.84) and 2.25 (1.47-3.38) respectively. Higher precision is seen with the reduction in hazard identified as being associated with colorectal cancer with a hazard ratio of 0.53(0.22-0.88).

Stroke

Eight cancer sites showed an association between stroke and an increased risk of hazard. In most cases, precision was low with the exception of lung cancer which showed a hazard ratio of 1.24 (1.11-1.38) increase in hazard. The largest effect size was in lymphoma with a hazard ratio of 2.17 (1.49-3.15%) increase in hazard

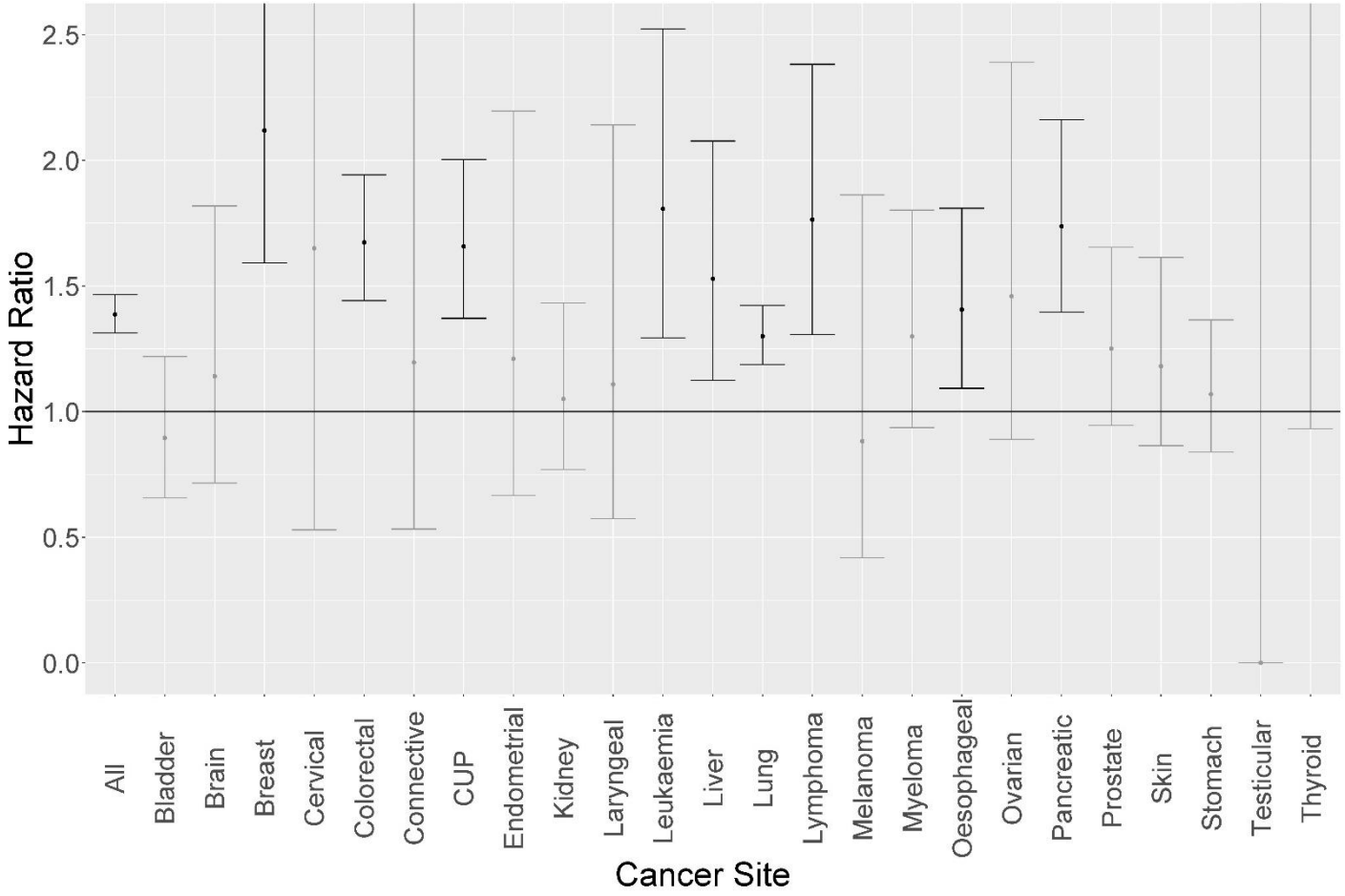


Figure 65: Cox Derived Cancer Cause-Specific Hazard Ratios for CCF in All Cancer and Site Specific Cancer Cohorts *– Cox derived cancer cause-specific hazard ratios with 95% confidence interval associated with CCF in the All Cancer and Cancer Site Specific Cohorts. Each estimate is derived from a standalone model adjusting for age, gender and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

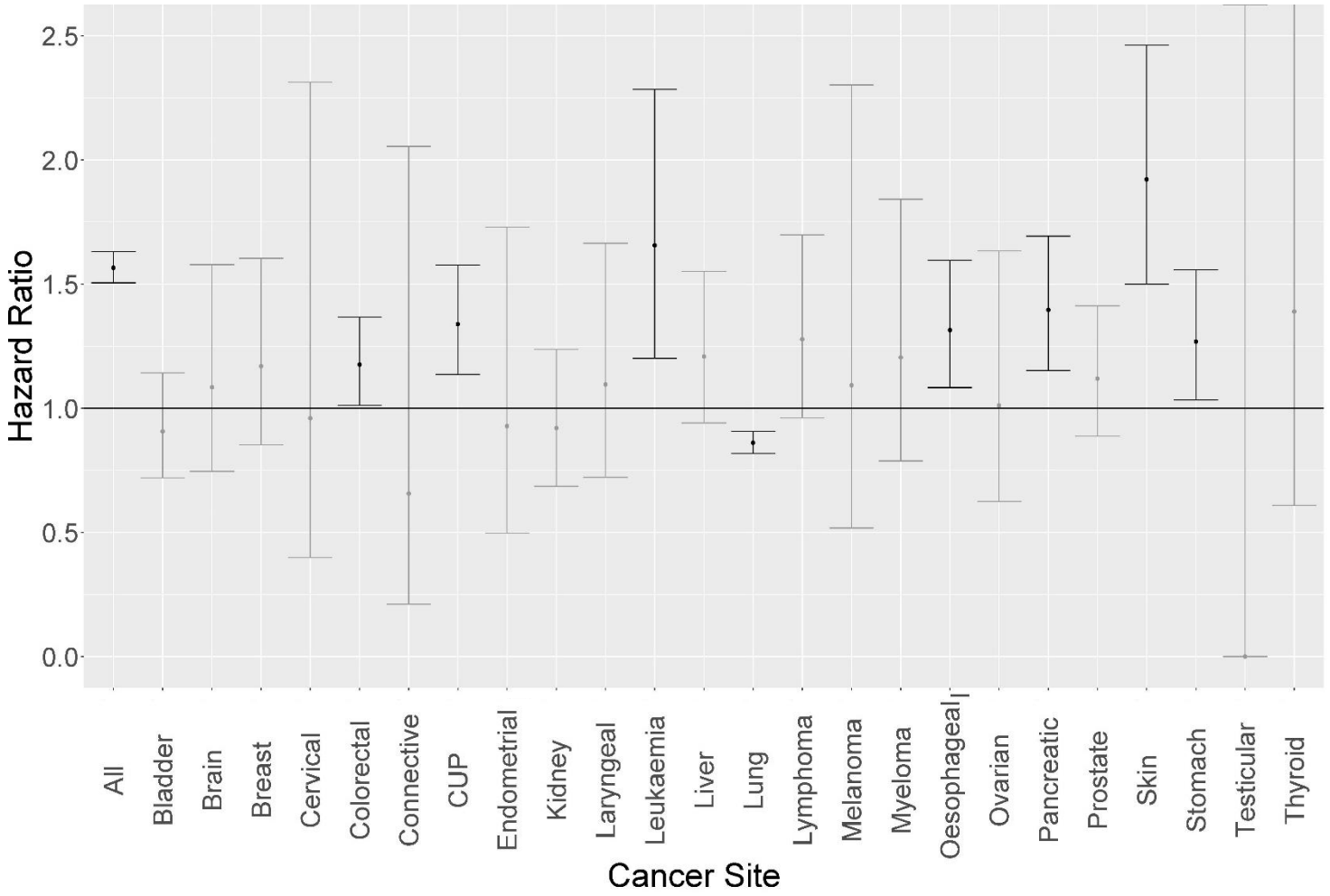


Figure 66: Cox Derived Cancer Cause-Specific Hazard Ratios for COPD in All Cancer and Site Specific Cancer Cohorts *– Cox derived cancer cause-specific hazard ratios with 95% confidence interval associated with COPD in the All Cancer and Cancer Site Specific Cohorts. Each estimate is derived from a standalone model adjusting for age, gender and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

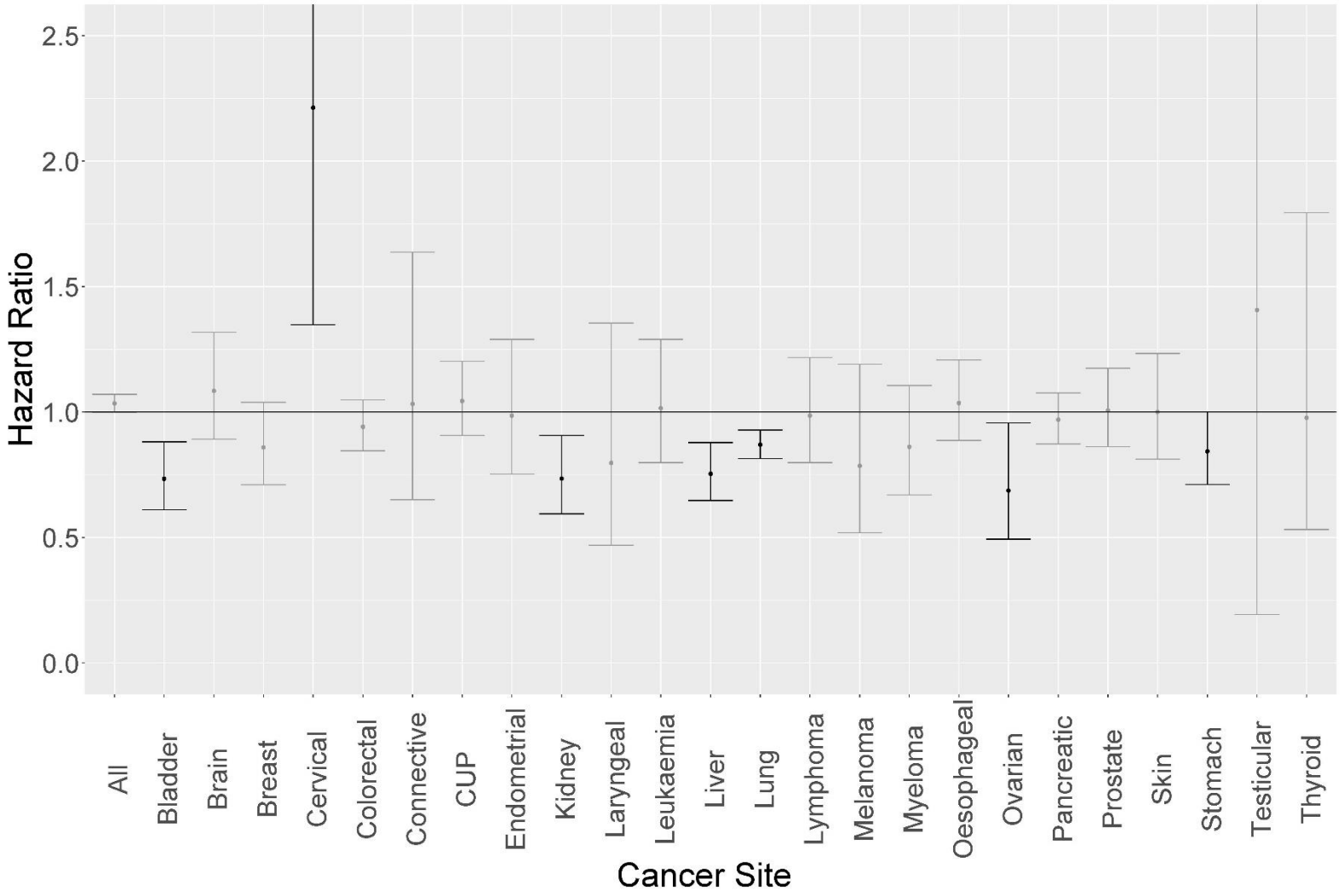


Figure 67: Cox Derived Cancer Cause-Specific Hazard Ratios for Diabetes Mellitus in All Cancer and Site Specific Cancer Cohorts *– Cox derived cancer cause-specific hazard ratios with 95% confidence interval associated with Diabetes Mellitus in the All Cancer and Cancer Site Specific Cohorts. Each estimate is derived from a standalone model adjusting for age, gender and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

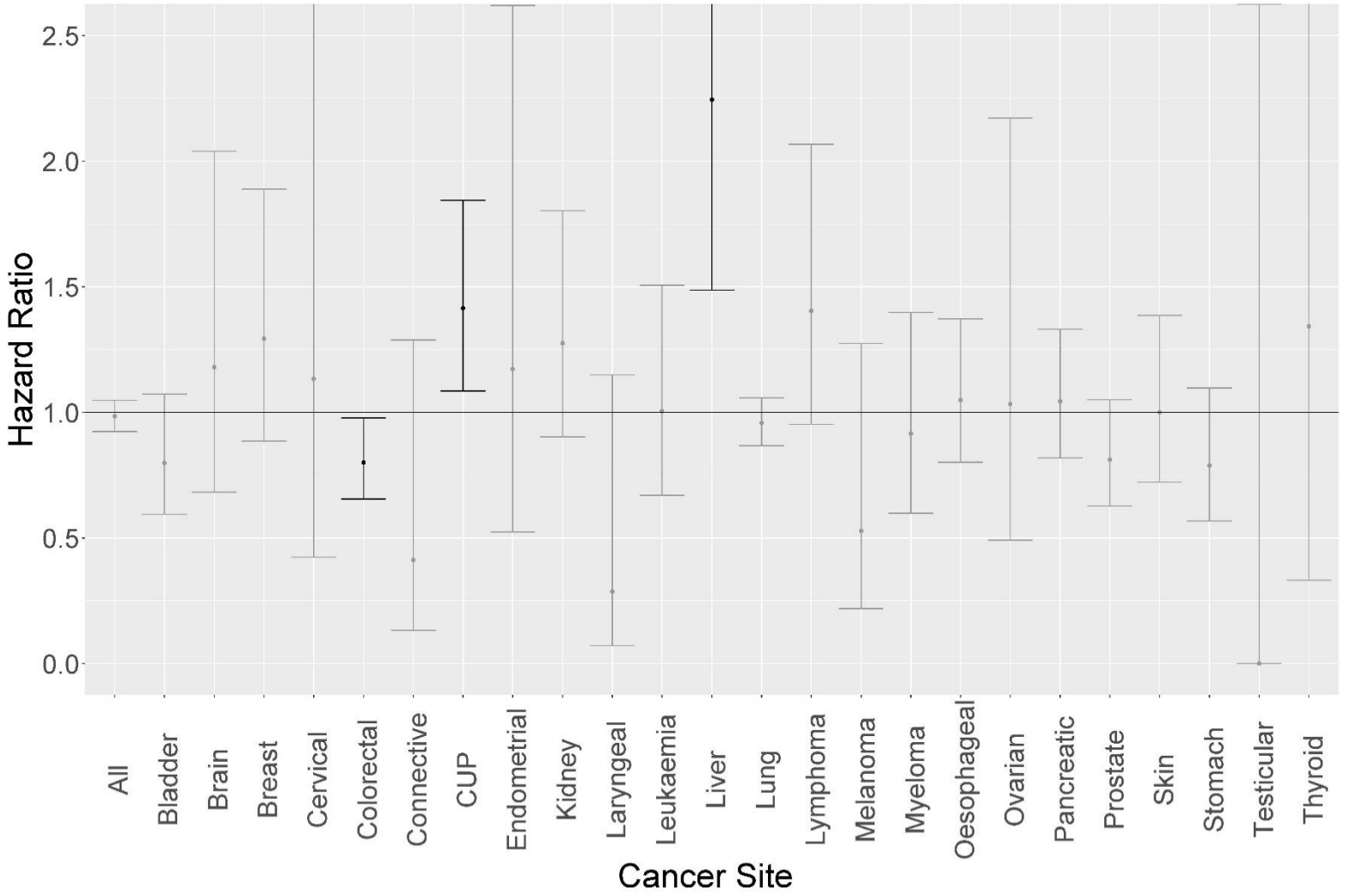


Figure 68: Cox Derived Cancer Cause-Specific Hazard Ratios for MI in All Cancer and Site Specific Cancer Cohorts *– Cox derived cancer cause-specific hazard ratios with 95% confidence interval associated with MI in the All Cancer and Cancer Site Specific Cohorts. Each estimate is derived from a standalone model adjusting for age, gender and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

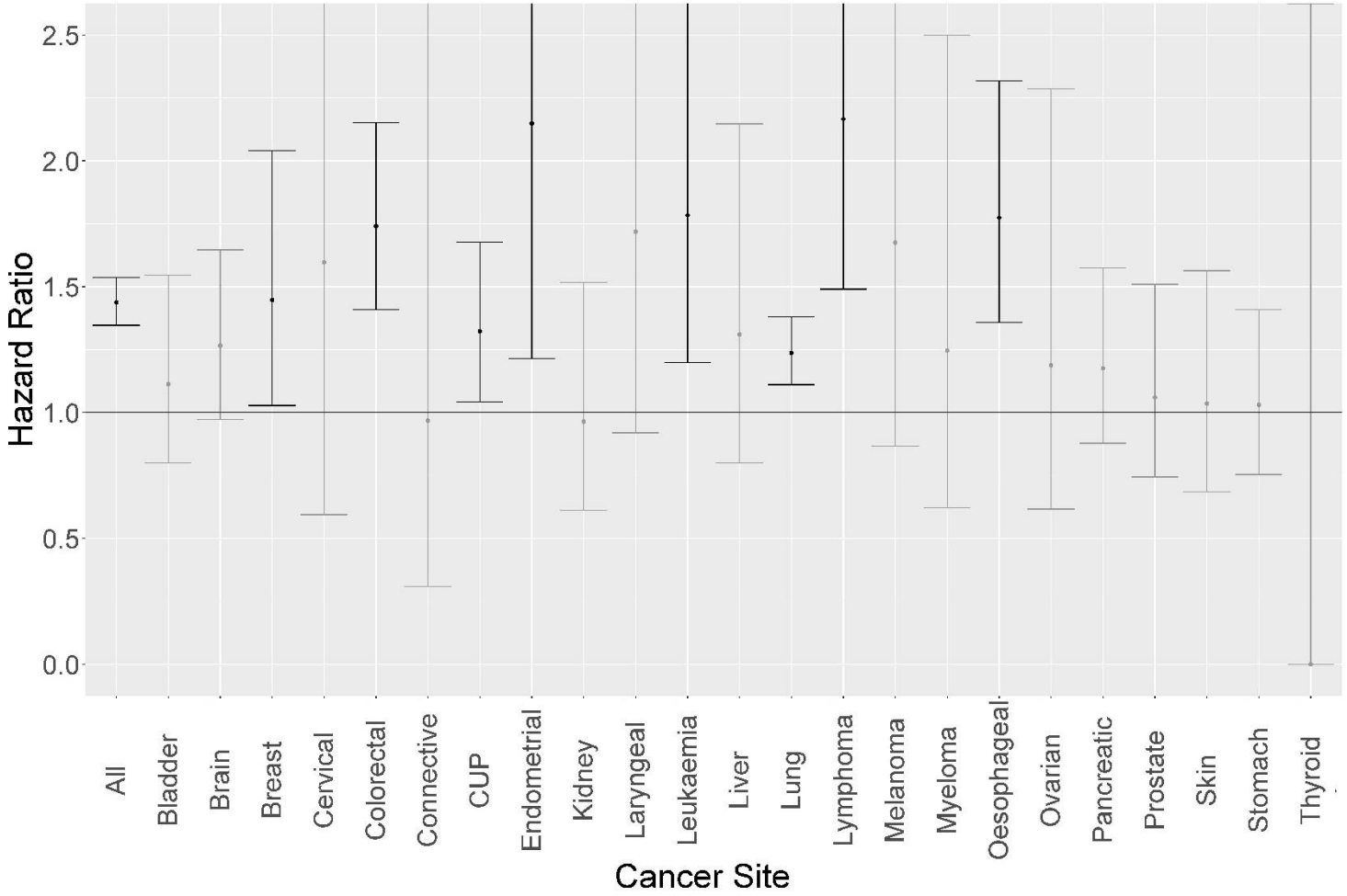


Figure 69: Cox Derived Cancer Cause-Specific Hazard Ratios for Stroke in All Cancer and Site Specific Cancer Cohorts *– Cox derived cancer cause-specific hazard ratios with 95% confidence interval associated with Stroke in the All Cancer and Cancer Site Specific Cohorts. Each estimate is derived from a standalone model adjusting for age, gender and deprivation. Those in black have a high probability of unidirectional hazard and those in grey have a low probability of unidirectional hazard. Where confidence intervals are large they may extend beyond the bounds of the plot.*

Precision

A total of 19 models met the precision cut offs. Only two cancer sites are represented with 7 comorbidities in lung cancer and one in colorectal cancer. The remaining models were based on the all cancer cohort. Of note, 9 of the 19 models are associations with a reduction in hazard. Within the site specific cohorts the largest precise estimate was for CCF which was associated with a hazard ratio of 1.3 (1.19-1.42).

Site	Comorbidity	Hazard Ratio	Lower CI	Upper CI	CI Span	CI Span as %
Lung	Diabetes	0.87	0.81	0.93	0.11	13.01%
	T2 Diabetes	0.86	0.80	0.93	0.13	14.58%
	Congestive Cardiac Failure	1.30	1.19	1.42	0.24	18.16%
	Hypertension	0.84	0.80	0.87	0.07	8.75%
	Asthma	0.82	0.75	0.90	0.15	17.98%
	COPD	0.86	0.82	0.91	0.09	10.27%
	Hyperlipidaemia	0.86	0.80	0.92	0.12	13.64%
Colorectal	Hypertension	0.89	0.83	0.96	0.13	15.06%
All	T2 Diabetes	1.17	1.12	1.21	0.09	8.01%
	Congestive Cardiac Failure	1.39	1.31	1.47	0.15	10.98%
	Arrhythmia	1.10	1.05	1.14	0.08	7.75%
	COPD	1.57	1.50	1.63	0.13	8.02%
	Peptic Ulcer Disease	1.14	1.07	1.21	0.13	11.84%
	Chronic Kidney Disease	1.16	1.08	1.24	0.17	14.37%
	Stroke	1.44	1.35	1.54	0.19	13.17%
	Rheumatoid Arthritis	1.16	1.05	1.28	0.22	19.39%
	Obesity	0.83	0.77	0.91	0.14	16.85%
	Hyperlipidaemia	0.91	0.87	0.94	0.08	8.32%
	Peripheral Arterial Disease	1.34	1.25	1.43	0.18	13.10%

Table 13: High Precision Comorbidity Cancer Cause-Specific Hazard Ratio Results *–Cancer cause-specific hazard ratios associated with comorbidity extracted from each Cox proportional hazard model. The confidence interval for the hazard ratio was extracted and is shown as the lower confidence interval, upper confidence interval and the total span of the confidence interval. The span is further expressed as a percentage of the hazard ratio point estimate. The results shown are limited to those that have a high probability of unidirectional hazard, a confidence interval span of less than or equal to 0.25 and where the span was less than or equal to 25% of the point estimate for the hazard ratio*

6.2.4 - Comparison with All-Cause Results

To facilitate model interpretation **Table 14** includes only results where both the all-cause hazard from chapter 5 and the cause-specific hazard have a high probability of unidirectional hazard and the effect estimate is in the same direction. **Table 15** includes those results where the precision cut off is achieved. A total of 58 site specific analyses showed unidirectional effects in the same direction across the all-cause and cause specific analysis. Of these 8 met the precision criteria and were all in the lung cancer cohort. 5 of these were for associations with reductions in hazard.

Site	Label	All-Cause Hazard	Cause-Specific Hazard	Span
	T1 Diabetes	2.54	1.77	2.02
	Congestive Cardiac Failure	2.97	2.12	1.23
	Arrhythmia	1.91	1.43	0.60
	Varicosities	0.69	0.51	0.54
Breast	Chronic Kidney Disease	2.53	1.97	1.47
	Paraplegias	2.23	2.04	1.96
	Stroke	1.85	1.45	1.01
	Dementia	3.08	3.02	2.13
	Peripheral Arterial Disease	2.23	1.76	1.69
Prostate	Dementia	2.28	2.94	2.61
	Peripheral Arterial Disease	2.05	1.48	0.95
	Diabetes	0.92	0.87	0.11
	T2 Diabetes	0.91	0.86	0.13
	Other Diabetes	1.26	1.26	0.57
	Congestive Cardiac Failure	1.36	1.30	0.24
	Hypertension	0.87	0.84	0.07
	Thromboembolic Disease	1.47	1.47	0.45
	Asthma	0.84	0.82	0.15
Lung	COPD	0.91	0.86	0.09
	Malabsorption	0.63	0.56	0.63
	Paraplegias	1.24	1.28	0.40
	Demyelination	1.85	2.21	2.52
	Stroke	1.25	1.24	0.27
	Dementia	1.43	1.34	0.39
	HIV	1.95	2.28	3.70
	Obesity	0.44	0.33	0.09
	Hyperlipidaemia	0.90	0.86	0.12
	Congestive Cardiac Failure	1.89	1.67	0.50
	Varicosities	0.63	0.47	0.46
	Thromboembolic Disease	1.60	1.58	0.81
	COPD	1.54	1.18	0.35
Colorectal	Chronic Kidney Disease	1.70	1.46	0.60
	Paraplegias	2.29	1.75	1.19
	Neuromuscular Disease	2.06	1.99	2.37
	Stroke	2.03	1.74	0.74
	Dementia	2.46	2.46	1.22
	Peripheral Arterial Disease	1.51	1.31	0.60
Skin	COPD	2.21	1.92	0.96
	Congestive Cardiac Failure	2.46	1.76	1.08
Lymphoma	Stroke	2.25	2.17	1.66
	Congestive Cardiac Failure	1.73	1.74	0.77
Pancreatic	COPD	1.41	1.40	0.54

Site	Label	All-Cause Hazard	Cause-Specific Hazard	Span
Leukaemia	Congestive Cardiac Failure	1.95	1.81	1.23
	COPD	1.78	1.66	1.08
	Stroke	1.71	1.78	1.45
Endometrial	Stroke	2.24	2.15	2.58
Oesophageal	Congestive Cardiac Failure	1.41	1.41	0.72
	COPD	1.32	1.31	0.51
	Stroke	1.78	1.77	0.96
Stomach	COPD	1.37	1.27	0.52
Liver	Diabetes	0.77	0.75	0.23
	Myocardial Infarction	1.48	2.24	1.90
	Congestive Cardiac Failure	1.66	1.53	0.95
Cervical	Diabetes	2.05	2.21	2.29
CUP	Myocardial Infarction	1.43	1.42	0.76
	Congestive Cardiac Failure	1.62	1.66	0.63
	COPD	1.32	1.34	0.44
	Stroke	1.42	1.32	0.63
All	Congestive Cardiac Failure	1.68	1.39	0.15
	COPD	1.70	1.57	0.13
	Stroke	1.58	1.44	0.19

Table 14: Comorbidity Cancer Cause-Specific Hazard Ratio Results Consistent with All Cause Hazard Results *–Hazard ratios associated with comorbidity were extracted from each all cause and cancer cause-specific Cox proportional hazard model. Where both showed the hazard to have unidirectional effect the hazard estimates and cause specific confidence interval span are shown.*

Site	Label	All-Cause Hazard	Cause-Specific Hazard	Span	Span as %
Lung	Diabetes	0.92	0.87	0.11	0.13
	T2 Diabetes	0.91	0.86	0.13	0.15
	Congestive Cardiac Failure	1.36	1.30	0.24	0.18
	Hypertension	0.87	0.84	0.07	0.09
	Asthma	0.84	0.82	0.15	0.18
	COPD	0.91	0.86	0.09	0.10
	Stroke	1.25	1.24	0.27	0.22
	Hyperlipidaemia	0.90	0.86	0.12	0.14
All	Congestive Cardiac Failure	1.68	1.39	0.15	0.11
	COPD	1.70	1.57	0.13	0.08
	Stroke	1.58	1.44	0.19	0.13

Table 15: High Precision Comorbidity Cancer Cause-Specific Hazard Ratio Results with Results Consistent with All-Cause Hazard Ratio *–All Cause and cancer cause-specific hazard ratios associated with comorbidity were extracted from each Cox proportional hazard model. . The results shown are limited to those that had a high probability of all-cause and cancer cause specific unidirectional hazard, a cause-specific confidence interval span of less than or equal to 0.25 and where the cause-specific confidence interval span was less than or equal to 25% of the point estimate for the hazard ratio.*

6.3 - Discussion

6.3.1 - Data Pre-processing

A large proportion of patients have no cause of death data, despite being recorded as deceased. This data is imported from external sources as the cause of death is recorded with the registry office and shared with the hospital trust at a later stage in some instances.[183] The missing data may be due to incomplete sharing of data with the hospital trust for patients. As a result, it might be possible to enhance this data by asking for a manual extraction of this dataset from the office for data release to enhance clinical records. Another possibility is that patients who die, but are not in the UK will not be subject to the same processes. As such, a patient may be reported as deceased but no data on cause of death that is made available, as the death is recorded in an alternative country. Although this may occur in a proportion of cases it seems implausible as the sole explanation given the high percentages shown in **Table 11**.

When looking at the trend in missing cause of death by year no clear trend is demonstrated up until 2017. This suggests that the proportion of cases where the data is missing is fairly static over time. From 2017 onwards there is a sharp rise in missing data reaching 100% of deaths from 2018 onwards showing that the data accuracy is worse in more recent years. This however is a product of how this data is derived with it needing to be collected by central government bodies and processed by them before it can be made available to the hospital. This is therefore simply a reflection of the long time lag that occurs between death and its registration, and the subsequent processing and release of cause of death data. The falloff in number of missing cause of death data items in 2020 is shown to be an artefact by virtue of the percentage of deaths remaining at 100%. It results from the fixed cut off for inclusion within the all cancer cohort at the end of 2018. The number of deaths will start to fall annually after this cut off as no new diagnoses are added and thus the number at risk will shrink.

Although the approach implemented was to exclude patients with missing data, this runs the risk of introducing bias to the analysis in the same way that complete case analysis does.[162] If the data are missing systematically, for certain types of patients, then this could create an inadvertent selection bias that may not be dealt with by adjustment with covariates alone. An alternative approach would be to include all patients however treat unknown deaths in the same way as non-cancer deaths and treat them as a censoring event rather than a death. This would have had the advantage of increasing patient numbers and enhancing precision. As the mechanism by which this data is missing is not fully understood handling the data in this way would have created potential challenges in knowing how to interpret the output of analysis, introduced misclassification error and may have introduced a form of censor bias.[175]

6.3.2 - Model Assumptions

Proportional Hazards

Similar results are identified within the cause-specific models as those found within the all-cause models in the previous chapter. The graphical approach identified no instances of a violation of the proportional hazards assumption occurring. As with the previous all-cause analysis, due to the large numbers of patients the graphical approach was the only method that was applicable to all the cohorts under investigation. The identification however of some instances where a given variable is derived from a smaller cohort, shows no graphical evidence of a violation, but does show numerical evidence of a violation, calls into question the robustness of the graphical approach. This suggests that the subjective elements of a graphical approach do need to be taken into consideration and that despite being assessed directly, that there may be some issues that remain in regards to the proportional hazard assumptions made.

Linear Assumptions

Several instances of violations of linear assumptions are identified. The non-linear relationships between age and outcomes appear to be more marked on average which is reflected in the non-linear relationship with age occurring in the all cancer cohort. The presence of non-linear relationships, suggests a possibility of a greater level of residual confounding when adjusting for these covariates within our Cox models. As such, those with identified non-linear relationships are likely to have estimates that deviate further from the true effect when compared to those without non-linear relationships. This is particularly true for leukaemia where the cohort shows evidence of non-linearity with regards to both age and IMD quintile.

Chapter 5 detailed approaches to overcome non-linearity through cohort stratification. Alternative approaches to regression modelling could also be implemented to combat non-linear relationships.[368] One approach is the use of polynomial regression models.[369] Here the non-linear variable is included in its raw form, quadratic form and increasing higher order polynomial forms. This is continued until as non-significant p value is obtained for the next order polynomial. This allows the regression model to capture non-linear relationships, however this will model non-linearity only to a certain extent. Thus more complex relationships still may not be accurately modelled with this approach.

To improve its performance, this approach can be combined with an approximation to the previously detailed stratification method by implementing splines. Here a number of cut off values are identified with a polynomial regression model fitted to the data that lies between each cut off for that variable. In such cases, finding the appropriate number of and values for cut off points, which are termed knots, is crucial in creating a model that appropriately reflects the non-linear relationship. In general a higher numbers of knots will result in a closer fit of the model to the data, however this also may result in overfitting.[370] This may reduce the external validity of the results obtained. Lower numbers of knots reduce the risk of overfitting, but may result in a less accurate representation of the non-linear patterns seen within the data. These models are also challenging to interpret clinically.

Influential Outliers

When assessing the impact of influential outliers in **Figure 59** and **Figure 60** the results suggest that in most cases the outlier values are pulling the effect estimate towards the null hypothesis. This may be due to the outliers having improved survival estimates, however this is less likely to be the case as the opposite effect was seen in the all-cause equivalent analysis. Instead, it is likely that the outliers are associated with an increase in the risk on non-cancer deaths. This general trend is in some ways helpful for the analysis, as it suggests that where influential outliers are not excluded, it is producing a null result where a true difference may exist. If this holds true across the other cancer cohorts, then our process of focussing on only those with a high probability of unidirectional hazard is unlikely to include results that are only high in probability due to outliers. The negative trade-off is however that in some cases we may be discounting some effects which have been deflated by the outlier cases.

6.3.3 - Cause of Death as a Source of Bias

Competing risks analysis is fundamentally reliant on cause of death data, whether implementing a sub-distribution or cause-specific approach. As such, errors in the cause of death data will result in misclassification error and biased estimates as a results.[218,371] Significant volumes of research have been undertaken in numerous countries to assess the quality of cause of death reporting.[371–373] In

almost all studies high volumes of death certificates have been shown to be incorrect. Despite these high volumes of errors identified, the error rate has been shown to be inconsistent when comparing different conditions. The accuracy of studies relating to cancer deaths show improved accuracy over diagnoses such as cardiovascular disease.[184,371,374] As a result, the focus of our analysis being on cancer related versus non-cancer related deaths, as opposed to individual cancers as a cause of death, or other individual diagnoses as a cause of death, should improve the accuracy of the cause of death data. Despite this, there will still be an inherent error rate with this data which will likely result in some degree of bias within the estimates generated.

6.3.4 - Interpretation of Cause-Specific Analysis

When interpreting the output of these cause-specific analyses it is not sufficient to analyse the estimated hazard on its own. As detailed in chapter two, competing risk methodologies can demonstrate evidence of reduced risk if there is a rise in a competing risk.[180,364] As such, the cause specific competing risk should be assessed in tandem with the overall all cause hazard. Where the overall mortality demonstrates increased hazard and the cancer-cause specific hazard is reduced then it is likely that the increased mortality is caused by non-cancer causes, however it is not necessarily true that cancer risk is truly reduced to the extent estimated. If patients who are frail are more likely to die from either cause when it occurs in isolation but the effect is greater in the non-cancer condition, then the process of censoring non-cancer death creates a selection bias. The remaining at risk population is in effect stripped of the most vulnerable patients and thus creates the appearance of a survival advantage. This is a form of censor bias.[175] The converse is also true such that if there is an improvement in the hazard of the competing risk then this may cause more frail patients to survive making the cause specific outcomes appear worse. Thus, when the all-cause hazard and cause-specific hazard show the same direction of effect it is likely that the cause-specific effect is a meaningful contributor to the overall effect seen. When however it shows the opposite direction of effect, then the analysis result can be considered unclear, it may be a true estimate or may be due to censor bias which is not possible to calculate or estimate.

A further issue arises due to the censoring approach and missing cause of death data. Within the cause-specific competing risks approach, deaths from the non-cancer cause are treated as censored rather than a true event. In addition the removal of patients without cause of death data, reduces the total at risk population. These two factors result in increased levels of uncertainty of the measures derived, with wider confidence intervals. This is reflected in the lower numbers of comorbidities and cancer sites showing a high probability of unidirectional hazard. It also accounts for the smaller numbers of models meeting the precision criteria presented in **Table 13**. It could therefore be argued that a larger study population may be needed when assessing these cause-specific effects.

Due to the limitations described above and in chapter 2 interpreting the findings of cause-specific analyses can be challenging where the hazard estimates for all-cause mortality and the cause-specific mortality show opposite directions of effect. As such, although all of the results are presented, the discussion will focus primarily on the results presented in **Table 14** where the cancer cause-specific mortality and all-cause mortality are in agreement on the direction of effect.

When reviewing **Table 14** it is tempting to compare the hazard ratios in more detail and assess what proportion of the total hazard is accounted for by the cancer related deaths. This however would be inappropriate, as the methods applied in each case serve a different purpose and thus must be interpreted differently. The all-cause estimate is the risk of one group versus another taking into account all potential causes of death. The cause-specific analysis, attempts to quantify the hazard

relating to dying from a particular cause, under the assumption that dying from all other causes were not possible. As such the cumulative incidence of events generated from the estimate will be higher than in the actual population.[179] Due to these differences, comparisons beyond the direction of effect should be done with extreme caution.

6.3.5 - Comparison to Previous Research

An important step in assessing the potential validity of the analysis output is to compare them to the results of previously published research. A significant number of previous studies have attempted to assess the impact of comorbidity on cancer survival in the cause-specific context.[365–367,375] These studies have predominantly focussed on high prevalence cancers such as breast, colorectal, lung and prostate cancer. These studies have also largely relied on the use of the Charlson index[38], rather than analysing the effect of individual comorbidity. The Charlson Index is calculated based on three broad components; diagnoses, importance of condition and severity of condition. Any patient diagnosed with one of the pre-specified comorbidities results in points for that condition. Some conditions are deemed more significant than others, such that they attract more points, for example dementia scores 1 point where non-metastatic cancer diagnoses attract 2. Some conditions where severity may be assessed, such as diabetes with and without end organ damage, may also attract further points. Diabetes for example without end organ damage scores 1 point and those with end organ damage score 2 points.

This approach has the advantage of improving the number of patients within each strata, thus improving the precision of results. It also incorporates disease severity information in a way that the binary indicator of comorbidity applied in the above analyses does not. Further it attempts to incorporate information about several comorbidities into a single analysis.

Despite these advantages, the approach also has a number of important downsides. The use of any composite score results in loss of information as it is possible to derive the same score via multiple routes. In addition, the use of the Charlson score introduces a number of assumptions that were not in any way tested for in these previous publications. It assumes that amongst conditions which attract the same score the effects are equivalent. It also assumes that the weighting of conditions is appropriate for this particular context, for example, it assumes having both dementia and chronic pulmonary disease is equivalent to having moderate renal impairment only. Further, it assumes that disease severity is equivalent to some diagnoses such that that a patient with mild liver dysfunction with CCF and a stroke has an equal score and therefore assumed equal risk as someone with moderate liver disease.

This fundamental difference in analysis approach prevents direct comparison of the results presented above to the results of previous literature for the most part. Further, the results of our analysis call into question the potential validity of this previous research. Our results clearly demonstrate different effect sizes, for different conditions within the same site specific cancer cohort. This would suggest a violation of one of the inherent assumptions of the majority of the previous research in this area.

One example of a disease specific approach previous employed was in the analysis of diabetes mellitus in colorectal cancer.[376] This demonstrated that diabetes was associated with a high probability of unidirectional hazard in terms of overall survival for both colon and rectal cancers with an increased hazard of 12% and 21% respectively. When assessing cancer cause-specific results, only rectal cancer had a high probability of unidirectional hazard, with an estimated 30% increase in hazard. The results of the analysis undertaken within chapter 5 showed similar results for overall survival with an estimated 15% (7-12%) increase in hazard. Our cause-specific analysis however

showed a low probability of unidirectional hazard with a confidence interval from a 6% reduction in hazard to a 5% increase in hazard. The differences seen in the PPM analysis results compared to the previously published results could be due to several factors. The published analysis separated out colon and rectal cancer results, where the PPM based analysis did not. Differences between these might exist such that had the PPM analysis made a similar distinction, then the results might have been different. An additional factor is the use of the hybrid definition of diabetes throughout the study. As described in chapter 3 relying on clinical coding only produces a more pessimistic estimate of the effect of diabetes. The enhancement of our data might therefore be appropriately reducing the hazard estimation for diabetes.

6.3.6 - Cause Specific Hazard Estimates

Lung

Of the key cancer sites, lung cancer had the greatest number of comorbidities where there was agreement between the all-cause and cause-specific effect estimates. Of these 8 are examples of a reduction in hazard in the comorbid group and 8 showed an associated increase in hazard. The number of conditions with a reduced hazard in the lung cancer population is far larger than those found in the other site specific cohorts. It is important to consider if there is something intrinsic to the lung cancer population, which makes this more likely. Firstly, the large population numbers with high multi-morbidity increases the number of patients with the comorbidity of interest. This in general has resulted in more precision, and therefore a greater likelihood of having a high probability of unidirectional hazard. The shorter survival seen in lung cancer also amplifies the effect of small survival improvements on hazard estimates, whether driven by a true effect or artefact, such as lead time bias. As such, a modest survival difference of just a few days or weeks will be reflected more in the derived hazard ratio than it would be for longer surviving cancers such as breast and prostate cancers. If patients die early from their cancer, then the negative health effects of their pre-existing health condition may not have sufficient time to effect either the all-cause or cancer cause-specific mortality.

When taking into account our precision thresholds, the lung cancer cohort was the only site specific cohort to generate associations with all-cause and cause-specific agreement and that met the precision cut offs. A total of 7 comorbidities in lung cancer met the precision threshold with 5 showing a reduction in hazard and 2 showing an increase in hazard. The point estimates in each of these cases was also of a scale that would be clinically significant with the smallest decrease in risk being 13% in Type 2 diabetes and the smallest increase being 24% in stroke.

Previous research into the impact of comorbidity on lung cancer outcomes has suggested that the effects seen may differ with different histological subtypes.[377] Within the analyses undertaken in this chapter and chapter 5, the impact of histological subtypes was not considered. As highlighted in chapter 5, histology could be considered as a potential mediator of effect, however given the previous differences identified in the literature further research is needed to assess if these previously identified patterns persist when assessing individual comorbidities rather than using Charlson score, and if so ,what mechanism is driving this.

Although several potential sources of bias may be affecting the hazard estimates obtained there is a need for more detailed and specific analysis to be undertaken in each of these comorbidities. A precisely specified model using DAGs with the use of a more detailed and complete dataset might offer the opportunity of attempting a causal analysis that accounts for more of the bias identified. This could provide a more definitive answer as to the true nature of the relationship between comorbidity and lung cancer survival. This would be particularly important in the cases of improved

survival outcomes, as it might offer a potential route to identify new treatment and management strategies to use in patients to improve care and outcomes.

Colorectal

The second highest number of comorbidities in agreement between all-cause and cause-specific hazards is seen in colorectal cancer. Here 8 of the 9 results were associations with increased hazard. The exception to this was varicosities which is a pattern identified in both chapter 4 and chapter 5 previously. The cause specific analysis results identify that varicosities are associated with a reduction in the cancer cause-specific hazard in both the colorectal and breast cancer cohort. This would suggest that the reductions in all-cause hazard are primarily attributable to changes in the cancer related deaths within these cohorts. A literature search revealed no previous studies in this area and identifies a potential future area of focus for further research to identify if this association is seen more widely in other centres and geographical regions. If this pattern of association is identified wore widely, then further investigation into this effect is warranted to understand in more detail the potential drivers of it.

Despite the large hazard differences identified by the point estimates generated in the colorectal cohort, none of the results met the precision criteria. Despite the relatively low precision, when looking at the minimum hazard difference indicated by the confidence interval, the scale of effect appears to be clinically relevant in many instances. Estimates suggest at least a 44% increase in hazard for CCF, 19% for CKD, 41% for stroke, 22% for thromboembolic disease, 26% for paraplegia and 13% for neuromuscular disease. These hazard differences translate to large differences in projected survival times despite their imprecision, which would likely be of relevance to clinical and patient decision making. Other conditions show what may only be more modest differences such as COPD at 1% and 4% for peripheral vascular disease. Further analysis is needed using additional cohorts that improve the representation of comorbid patients. By increasing the numbers it is likely that the precision of results would also increase accordingly. This will provide a clearer idea of the true scale of associations seen.

Breast

The breast cancer cohort identified 9 comorbidities with all-cause and cause specific agreement. Varicosities were the only comorbidity associated with a reduction in hazard. The effect size in all conditions was large with the smallest point estimate increase in arrhythmias suggesting a 42% increase in hazard. Despite the large effect sizes none met the precision cut offs. As with colorectal cancer the minimum effect size suggested by the confidence intervals was still clinically relevant in many cases. For example lower confidence interval estimate of effect size was a 37% hazard increase for CKD. This example along with those highlighted in colorectal cancer suggests issues with the application of multiple cut offs for precision, clinical relevance, confidence interval and all-cause versus cause specific agreement. By applying so many rules to filter results down, there is a risk that many valid and clinically meaningful results may be ignored.

Prostate

Of all the comorbidities analysed in prostate cancer only two showed a high probability of unidirectional effect in both the all-cause and cancer cause-specific models and with the same direction of effect. This low number may be representative of the fact that associated differences seen in prostate cancer and comorbidity are primarily driven by non-cancer related causes. This would fit with arguments made previously in the literature that the effect of comorbidity in cancers with a better prognosis is driven by the effect of comorbidity having enough time to manifest[366]. This

does not however remove the possibility that cancer interacts with comorbidity to increase the impact of comorbidity in this group of patients. Further study is therefore needed with control cohorts to assess the net survival difference when compared to the non-comorbid and comorbid population with and without cancer. Additionally the study of other causes of death or ill health after cancer might reveal other potential effects. Cancer and its treatment might speed up the pathogenic processes driving ill health. As such a study of the time to event of key health conditions such as cardiovascular events, might reveal cancer accelerates processes such as this. If this were undertaken, then a death would need to be treated as a competing risk and the results compared to a non-cancer control group. This might provide insight into the causes of the lower all-cause survival seen in prostate cancer patients with many common comorbidities. In the case of the two prostate cancer results found in **Table 14** neither of these met the precision thresholds. Despite this the effect size was marked, suggesting an impact that may be clinically relevant for patients. This is especially true given the relatively long survival times seen in prostate cancer when compared to other primary malignancies.

Other Cancer Sites

Ten further cancer sites are identified as meeting the requirements for **Table 14**, although none of these meet the precision threshold. Amongst these cancer sites certain comorbidities predominate including stroke, CCF and COPD. As detailed in chapter 5 there are a number of physiological mechanisms which may directly link these conditions with cancer and its pathogenesis. Furthermore many of the treatments for common cancers are more limited by the presence of poor lung function, cardiac dysfunction and previous ischaemic events. Given the consistency with which these comorbidities show an association with differences in cancer cause-specific outcomes they offer a compelling target for further investigation into the drivers of this association whether causal or artefactual.

The cause-specific results in liver cancer show that diabetes is associated with a reduction in hazard. This echoes the all-cause results shown in chapter 5. These results suggest that the all-cause survival advantage seen in this group may be due to a reduction in cancer related hazard with point estimates suggesting a 25% reduction in risk. As described in chapter 5 the effects seen may be driven by cirrhotic burden and thus further research is needed, with a more detailed dataset that can condition on the degree of cirrhosis and hepatitis virus.[316] If however, this pattern continued to be present after adjustment for these factors more research would be required to understand the mechanisms driving these differential survival outcomes.

6.4 - Summary

The results presented above demonstrate a significant limitation in the previous cause-specific cancer research relating to comorbidity. The previous reliance on aggregated comorbidity scores is shown to be potentially flawed due to the highly varied nature of different comorbidities within a given group of cancer patients and with the same comorbidity across different cancer groups. In many of the comorbidities of interest, the effect seen in all-cause mortality appears to be being driven in a large part, by an associated effect on cancer cause-specific mortality. The number of models with agreement between the all cause and cause specific results appears to be greater in those cases where a survival improvement is identified.

Despite the patterns seen, the majority of comorbidities assessed are not associated with a meaningful difference in cancer cause-specific hazard. Of those that do, a further proportion show an effect that is opposite to that identified in the all-cause models. This renders the results of these analyses uninterpretable. The reliability and interpretability are further hampered when combined with the previously detailed issues relating to data bias, potential violations of model assumptions and bias introduced through the methods applied, as described in the DAG section in chapter 5. In most instances it is not possible to quantify the degree of misestimation of the associations identified. An argument could therefore be made that an alternative approach is needed which is not reliant on the correct attribution of effects to any given variable. One such approach would be the application of a predictive framework rather than an inferential one. Here, the value of comorbidities as predictors of outcome can be used. In this way comorbidities can be used as potential markers for high risk individuals, whilst accepting that the comorbidity may be independent of the cause of the outcome effects which they predict. The subsequent chapter uses this alternative approach applying machine learning methods with no underlying model assumptions to assess comorbidity in this manner.

Chapter 7 – The Application of Random Survival Forests to Assess the Relationship between Baseline Characteristics and Cancer Survival Using a Predictive Framework

7.0 - Introduction:

Chapters 3-6 include a range of analyses that are descriptive, exploratory and inferential in nature. They apply traditional statistical methods, to assess relationships between data items and outcomes. As discussed in chapters 5 and 6, although we are not attempting to make causal inference, there are important aspects of causal inference methodology that can still be used to inform the interpretation of the analysis results, in a non-causal way. As discussed, the focus on the cancer population introduces a block in causal relationships temporally downstream of the cancer diagnosis and introduces collider bias that is impossible to estimate.[296] Thus, the associations seen could be true, inflated, deflated or even the reverse of the true effect, depending on the level of bias introduced.

An alternative approach is to use a predictive framework to derive insight. Here, instead of attempting to quantify the associations seen, prediction is used to identify at risk groups. The models can be used to derive descriptions of the patterns of association within predictions. Further insight can be generated through the identification of baseline characteristics that are most informative or influential in generating predictions. Although the traditional methods used thus far, such as Cox proportional hazards[108], could be used for prediction, these have underlying model assumptions that previous chapters have suggested may not hold in all circumstances. This is particularly true for linearity[178] and this may result in a reduced ability to produce accurate predictions. Within this chapter an alternative method is applied which employs ensemble learning using random survival forests[109]. This method has been utilised due to its lack of underlying model assumptions and relative ease of interpretability (see chapter 2 for an overview). Despite this method having been described and implemented more than a decade ago, to date, it has been used infrequently within medicine and even less frequently in oncology focussed studies.[378,379] Within this chapter a predictive perspective is used to interrogate the PPM cancer cohorts as an alternative approach to generating insight into patient survival outcomes.

7.0.1 – Aims and Objectives

Aims

1. Apply a predictive machine learning approach to identify baseline characteristics that can be used to identify at risk patients.

Objectives

a) Identify the optimal hyperparameters for the development of accurate random survival forests for cancer survival prediction.
b) Build random forest models for overall all-cause survival in breast, lung, colorectal and prostate cancer.
c) Compare the predictive accuracy of random survival forests to Cox proportional hazards and Kaplan-Meier methods.
d) Assess predictive information gain of individual baseline characteristics using average minimum tree depth.
e) Assess the permutation variable importance scores of baseline characteristic in survival predictions.

f) Assess pairwise permutation variable importance score of baseline characteristic in survival predictions.

7.1 - Methods

7.1.1 - Testing and Training Sets

For the development and testing of our predictive models the holdout method was used.[380,381] Here each cohort was divided into a training and a testing dataset. The training dataset comprised 80% of the cohort and the testing set 20% of the cohort. These were divided at random using standard packages in R (see appendix).

7.1.2 - Hyperparameter Tuning

Model hyperparameter optimisation for sample size of candidate variables at each node and minimum size of patient cohorts in terminal leaves was conducted.[189] This was achieved using the tune.rfsrc function in R. This implements the rfsrc.fast algorithm to assess out of bag (OOB) error[382] for hyperparameter combinations. The rfsrc.fast algorithm uses subsampling to create close approximations to the true forests in a fast and computationally efficient manner. The optimisation was run using a forest size of 500, with candidate variable sizes increasing from a minimum of 3 upwards in single variable increments. Terminal node size was assessed from 3 increasing up in one unit increments. These parameters were selected due to 3 being the minimum selected for each hyperparameter in standard practice. The optimisation was set to stop assessing combinations once a combination of hyperparameters had OOB error that was better than the 10 subsequent hyperparameter combinations tested. Optimal hyperparameters and subsequent model building was undertaken for the top cancer sites by incidence, namely; breast, colorectal, lung and prostate cancer.

7.1.3 - Model Building

Once the optimal hyperparameters had been identified, these were used to build the full random forest model using the rfsrc algorithm.[109] The forest size was set to 500 with permutation variable importance (VIMP) estimation and split depth by tree being calculated during model building. The number of time points used to calculate ensembles was set to 100 in order to enable calculation in a memory efficient way due to the limited hardware available for analysis. The models were developed using age, gender, deprivation, comorbidity status for each comorbidity of interest, histological type, stage and grade. Full model specifications can be found in the appendix (**Table 26**).

7.1.4 - Internal Error Assessment

To assess the suitability of the 500 tree forest size implemented for each model the out of bag (OOB) error[382] was calculated with each additional tree added to the forest. This was then used to generate plots of OOB error through the model development and assess if a steady state of error had been achieved.

7.1.5 - Model Accuracy

The predictive accuracy of the resulting models was assessed using integrated Brier score and time dependent Brier score.[383] These were calculated using the holdout testing dataset in order to ensure accurate results were not due to overfitting. To provide a performance benchmark Cox proportional hazards models were developed to include, age, gender, histology, stage, grade and comorbidities. A Kaplan Meier estimate was also generated for benchmarking. The performance metrics were then compared.

7.1.6 - Assessing Variable Importance

Interpretation of the model was delivered through a combination of metrics including permutation Variable Importance (VIMP) scores[188] and minimum tree depth.[109] Permutation VIMP was applied to identify the most important features (variables) in influencing the model's predictive accuracy. Where a VIMP score was less than or equal to zero it was determined to be non-informative. Where VIMP was positive but less than 0.002 it was deemed to be predictive but a low value predictor. Those with a VIMP of greater than or equal to 0.002 were deemed to be important predictors.[109]

A second method was applied to assess the utility of individual variables. Average minimum tree depth was used as an assessment of the relative information gain yielded by each variable.[109] This was assessed against the calculated model specific cut off[109] to identify which variables were important in informing the predictions of the models.

To assess the interactions between variables, pairwise permutation VIMP was also calculated. Here the pairwise effect was assessed by calculating the VIMP when both features in the pair were permuted simultaneously.[384] This was then compared to the additive VIMP of each variable permuted individually to identify relationships that were greater or less than would be expected given their individual contribution to predictive accuracy.

7.1.7 - Partial Dependence Plots

The relationships between important variables and outcomes was assessed using partial dependence plots also called partial plots.[188,385] Here each individual had their survival probability at 1,3 and 5 years calculated. Where a continuous variable was being assessed a smooth curve was fitted to demonstrate the relationship between increasing values and survival probability. Where categorical variables were assessed, box and whisker plots were used to demonstrate the trends in the data.

7.2 - Results

7.2.1 - Hyperparameter Tuning and OOB Error

Table 16 demonstrates the results from the optimisation algorithm. The optimal hyperparameters identified were not consistent across each of the site specific datasets used. Many of the top hyperparameter combinations were close in accuracy to one another varying by just a fraction of a percent in several cases.

Site	Number of Candidate Variable (mtry)	Minimum Number in Terminal Leaf (nodesize)
Breast	7	10
Colorectal	12	9
Lung	22	30
Prostate	9	15

Table 16: Optimised Hyperparameters for Random Survival Forests *– Optimised hyperparameter results obtained from the tune.rfsrc algorithm when applied to the training cohorts for the breast, colorectal, lung and prostate cancer populations.*

When these hyperparameters were implemented to build the full model, the OOB error demonstrates that the error rate stabilised at fewer than 100 trees across each of the four models.

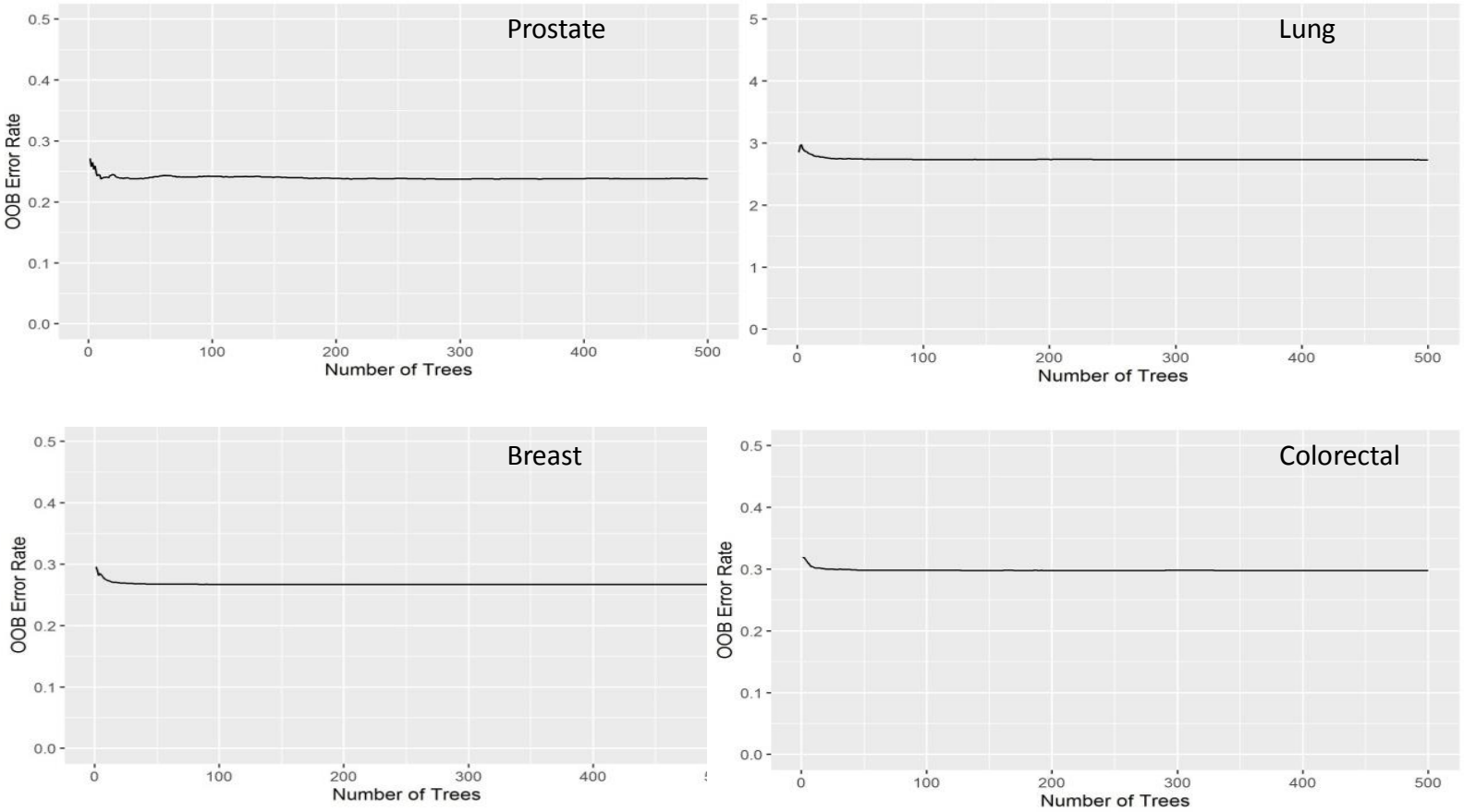


Figure 70: Effect of forest size on Out of Bag (OOB) Error *– Graphical representation of the OOB error rate, measured in C-Index, plotted against the number of trees within the forest. Results are shown for each of the four models developed using the PPM cancer site specific cohorts training data subsets with results of OOB error calculated as the forest were being trained.*

7.2.2 - Predictive Accuracy

The overall model predictive accuracy as assessed by integrated Brier score showed that overall the RSF was more accurate than the Cox proportional hazard or Kaplan Meier estimator in generating survival estimates for patients. The results of the integrated Brier scores for the models can be found in **Table 17**. All modelling strategies were better than both random guessing (0.5) and attributing all cases to a 50% risk (0.25).

	Kaplan Meier	Cox Proportional Hazard	Random Survival Forest
Prostate	0.114	0.111	0.079
Colorectal	0.193	0.148	0.129
Lung	0.118	0.118	0.097
Breast	0.198	0.153	0.134

Table 17: Comparison of KM, Cox and RSF Integrated Brier Scores *– Integrated Brier Scores for Kaplan Meier, Cox and random survival forest methods models trained on training datasets for prostate, colorectal, lung and breast cancer patients. Integrated Brier scores are based on the holdout testing dataset derived from each cohort.*

When assessing the model accuracy over time (**Figure 71**) in the case of all but prostate cancer, the RSF approach was more accurate at all time-points from 0 to 10 years. In the case of prostate cancer the model accuracy for RSF was greater from 0-8 years, however the RSF approach showed a reduction in accuracy after this time dropping to a lower level of accuracy than both Cox and KM approaches.

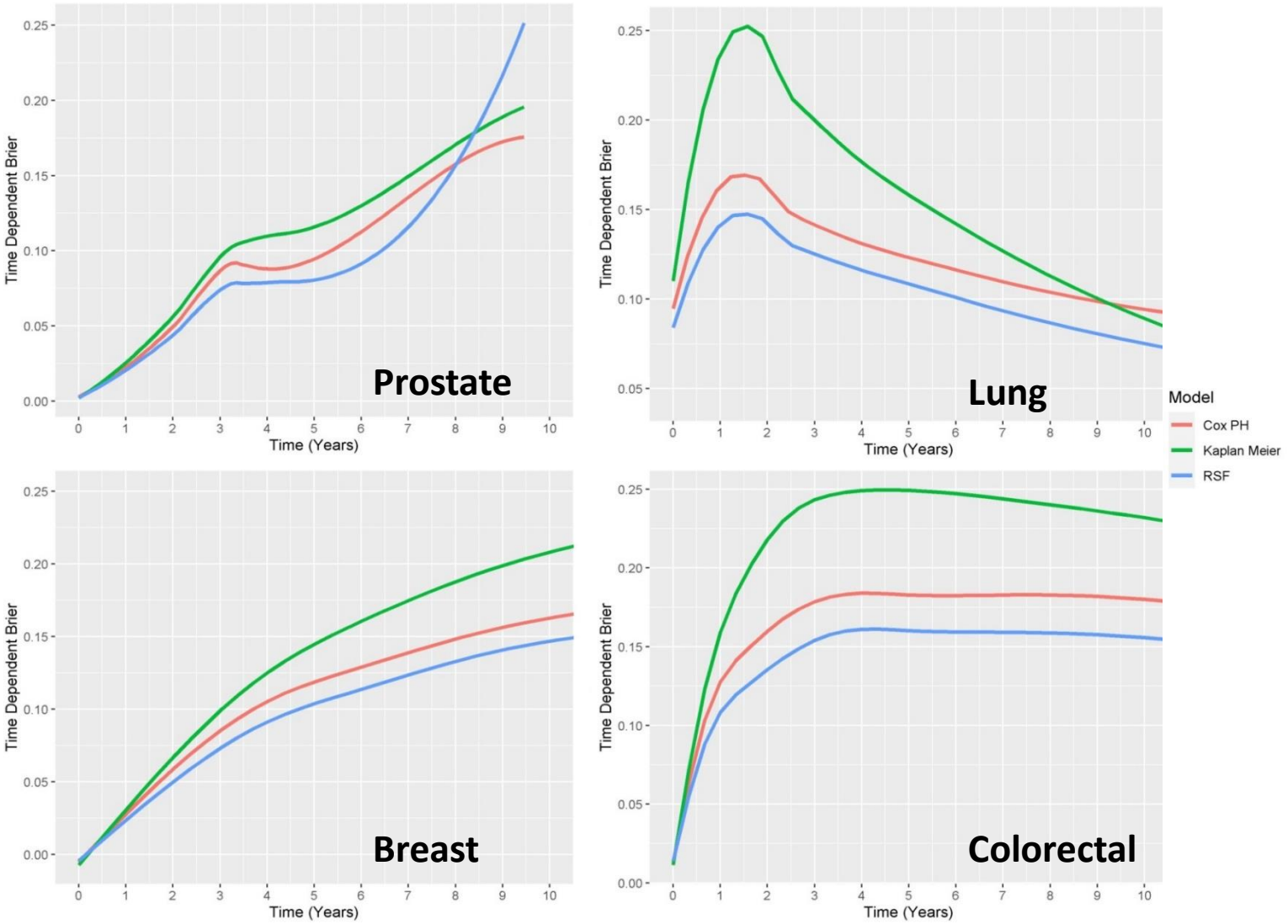


Figure 71: Time Dependent Accuracy with Brier Scores *- Comparison of predictive accuracy of Kaplan Meier, Cox and random survival forests modelling approaches using the testing holdout data subset in each of the four Cancer Site Specific Cohorts. Accuracy is assessed using time dependent Brier Scores.*

7.2.3 - VIMP

The variable importance scores derived from the breast cancer model identified 7 highly important predictors including age, stage, grade, histological subtype, deprivation quintile, coronary artery disease and hypertension (**Figure 72a**). Of these, age was the most important variable with a VIMP score 33% higher than stage, the next most important variable. 7 variables were identified as not being important predictors including gender, paraplegia, malabsorption, pancreatitis, spinal injury, MND and other respiratory diseases. The remaining variables show evidence of some but more limited predictive value.

The average minimum tree depth demonstrates a similar pattern overall with the top three variables being the same using this approach (**Figure 72b**). Histology is ranked as slightly lower than by the VIMP and COPD deemed more informative. Of note two of the variables classified as noise variables by VIMP are shown to have tree depths suggestive of useful information gain which includes paraplegia and gender. Three variables with low predictive VIMP have scores below the average tree depth cut off suggesting they do not provide significant information gain and these included Parkinson's, venous disease and cardiomyopathy.

The colorectal cancer VIMP scores showed stage to be the most important variable, being more than twice as important as the next variable of age (**Figure 73a**). A further 5 variables have VIMP scores suggesting a high predictive value which included grade, deprivation quintile, gender, arrhythmia, and CCF. Seven variables showed VIMP scores of less than 0 including venous disease, other respiratory disorders, spinal injury, MND, Parkinson's, demyelination and TIA. The other variables have VIMP scores suggestive of limited predictive value.

The average minimum tree depth measures show the same top three variables although the information gain advantage of stage over age was more limited than the difference seen in VIMP score (**Figure 73b**). Gender and IMD were ranked lower in their tree depth measures than by VIMP. TIA which was suggested to have no predictive utility by VIMP score was shown to demonstrate useful information gain by the tree depth measure. The non-informative variables as defined by tree depth, otherwise matched those suggested by the results of the permutation VIMP scores.

The lung cancer VIMP scores identified just 4 highly predictive variables which were stage, histology, age and grade (**Figure 74a**). The stage of cancer was scored very highly compared to the other variables with an importance score more than 8.5 times higher than age, the next most important variable. A larger number of variables were shown to have VIMP scores below zero, suggesting that they were not useful predictors. These included venous disease, MND, spinal injury, other respiratory diseases, rheumatoid arthritis, cardiomyopathy, demyelinating disease, liver dysfunction, neuromuscular disorders, other rheumatological diseases, peptic ulcer disease, TIA and Parkinson's.

Tree depth measures showed lower levels of agreement with VIMP than in the other cancer sites with only 3 of the top five variables being consistent across the two measures (**Figure 74b**) Tree depth suggested IMD and gender were less informative than the VIMP measures. Four of the variables with VIMP scores below zero were also ranked as showing information gain including liver dysfunction, rheumatological diseases, peptic ulcer disease and TIA. Peripheral arterial disease and malabsorption however fell below the tree depth cut off suggesting low information gain where VIMP suggested some predictive value.

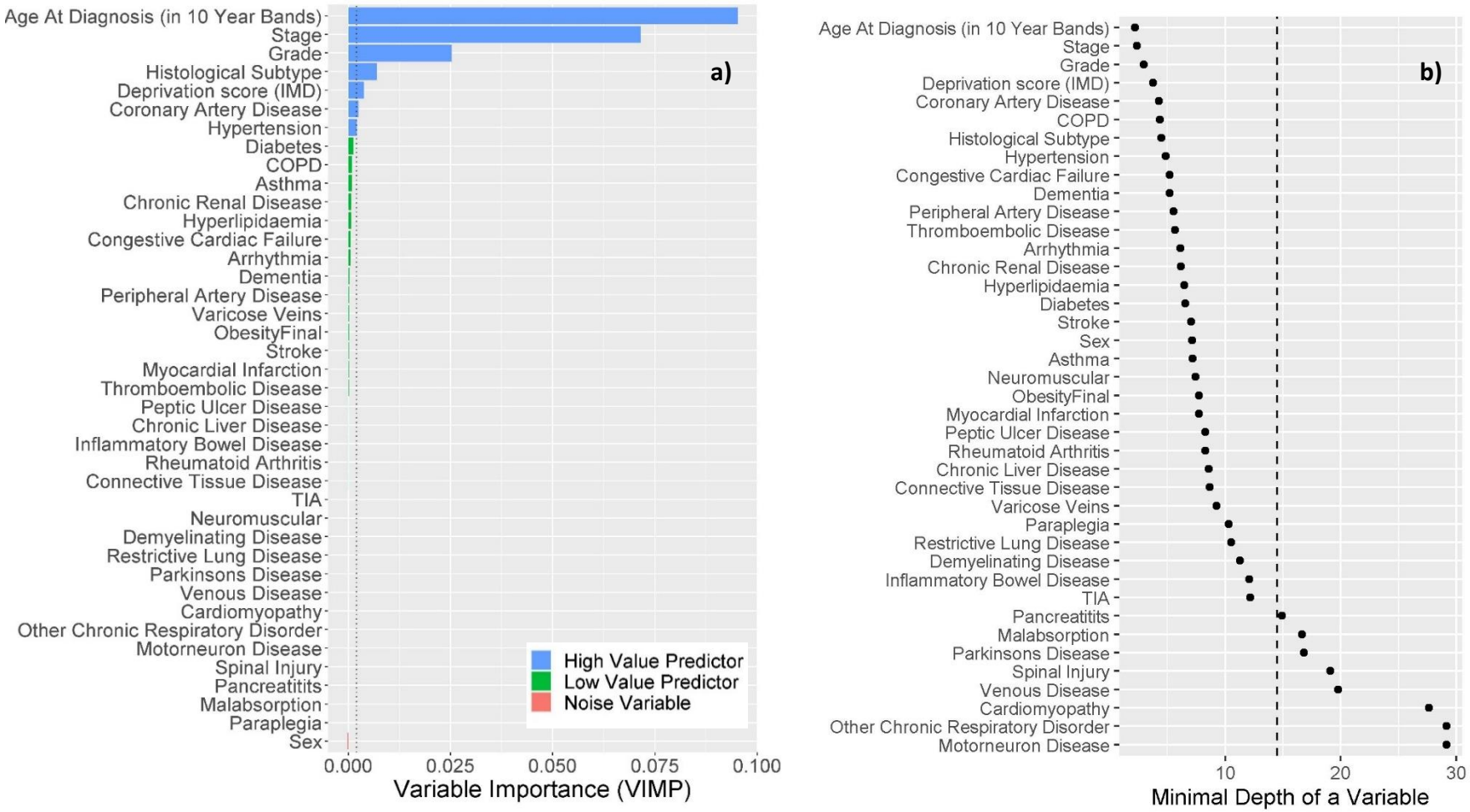


Figure 72: Breast Cancer Importance Scores – *a) Single variable permutation importance scores with colour denoting the predictive utility of the variable. b) Average minimum tree depth measures where points to the left of the vertical dotted line denotes a variable with useful information gain. Both measures are derived from the breast cancer random survival forest developed using the training subset of the Breast Cancer Site Specific Cohort.*

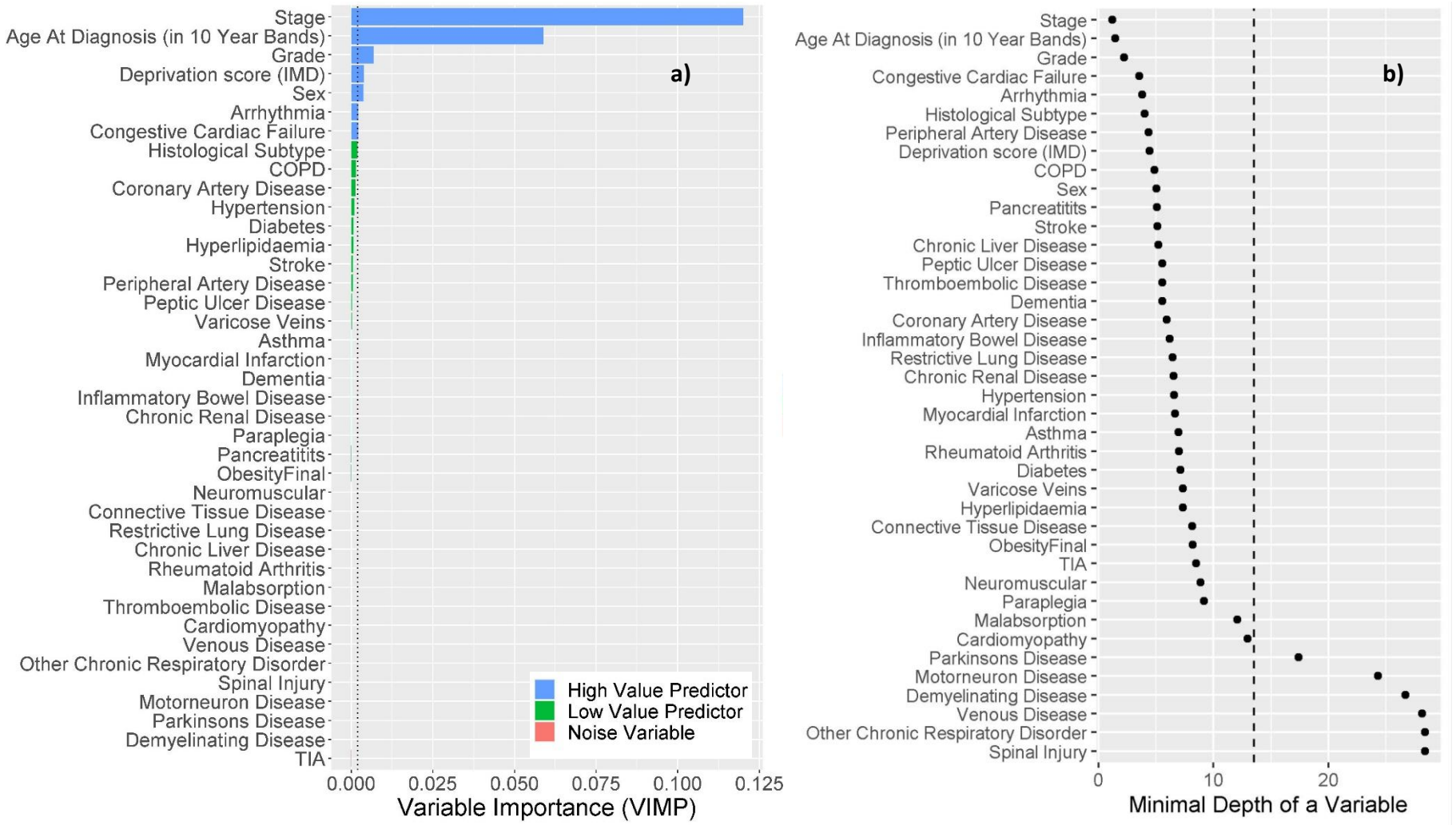


Figure 73: Colorectal Cancer Importance Scores *– a) Single variable permutation importance scores with colour denoting the predictive utility of the variable. b) Average minimum tree depth measures where points to the left of the vertical dotted line denotes a variable with useful information gain. Both measures are derived from the colorectal cancer random survival forest developed using the training subset of the Colorectal Cancer Site Specific Cohort.*

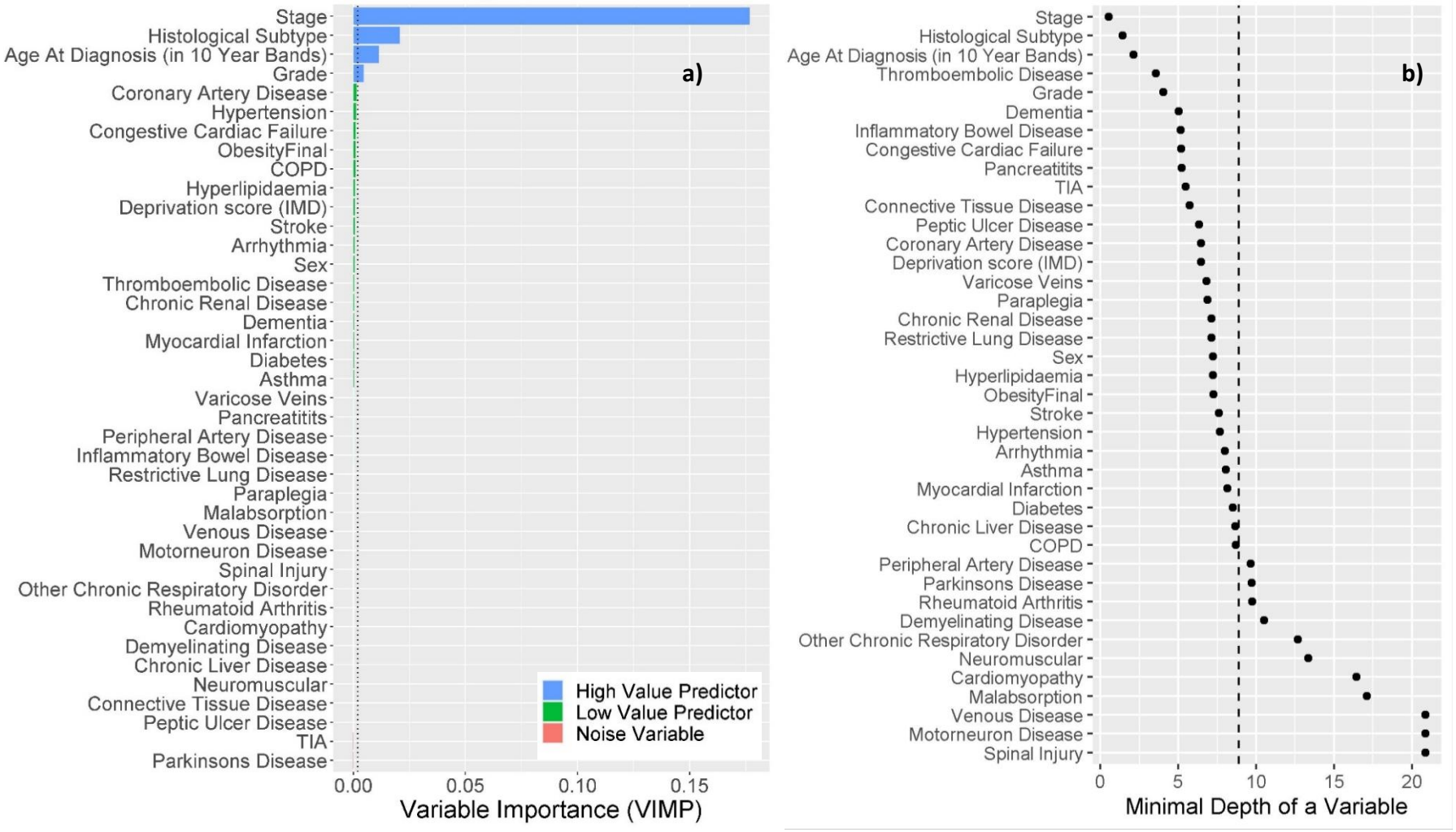


Figure 74: Lung Cancer Importance Scores *– a) Single variable permutation importance scores with colour denoting the predictive utility of the variable. b) Average minimum tree depth measures where points to the left of the vertical dotted line denotes a variable with useful information gain. Both measures are derived from the lung cancer random survival forest developed using the training subset of the Lung Cancer Site Specific Cohort.*

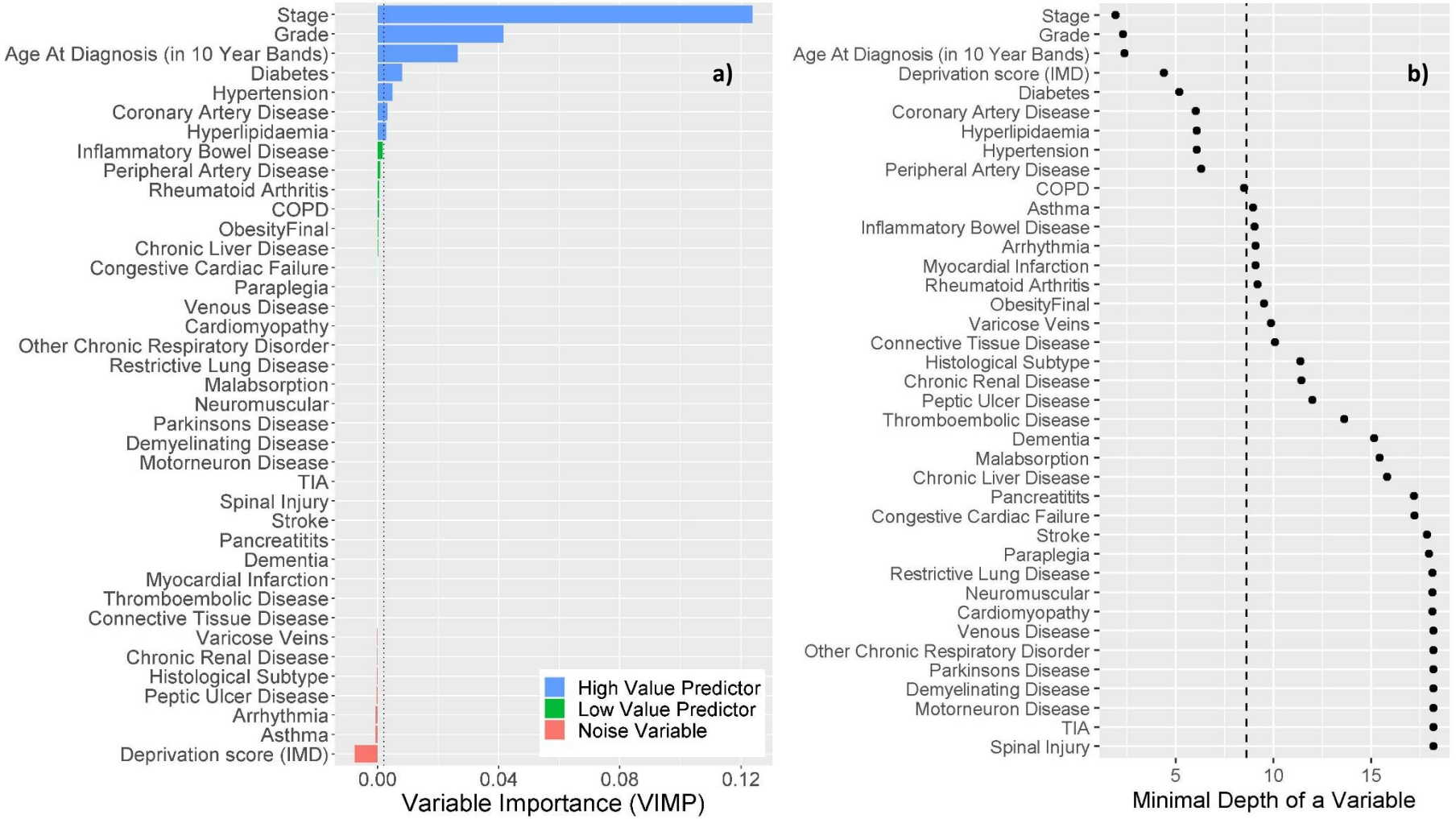


Figure 75: Prostate Cancer Importance Scores *– a) Single variable permutation importance scores with colour denoting the predictive utility of the variable. b) Average minimum tree depth measures where points to the left of the vertical dotted line denotes a variable with useful information gain. Both measures are derived from the prostate cancer random survival forest developed using the training subset of the Prostate Cancer Site Specific Cohort.*

The VIMP scores for the prostate cancer model demonstrate far fewer variables with predictive value (**Figure 75a**). Seven variables were shown as being highly predictive, 8 as lower value predictors but the majority were deemed of no predictive value. The most valuable predictor was stage which had a VIMP score over 3 times that of grade, the next most important variable. Other highly predictive variables included, age, diabetes, hypertension, coronary artery disease and hyperlipidaemia. Low value predictors included inflammatory bowel disease, peripheral arterial disease, rheumatoid arthritis, COPD, obesity, liver dysfunction, CCF and paraplegia.

The average minimum tree depth measures also suggest that fewer variables are important in the context of prostate cancer (**Figure 75b**). Deprivation levels are however deemed to provide information gain despite being the lowest in terms of VIMP score. All the other variables above the information gain threshold were also found to have VIMP scores above zero although some with VIMP above zero fell below the information gain threshold by minimum tree depth. These included inflammatory bowel disease, rheumatoid arthritis, obesity, congestive cardiac failure and paraplegia.

7.2.4 - Pairwise Permutation VIMP

When analysing the largest pairwise VIMP scores as shown in **Table 18**, there is overlap between the top single VIMP results shown in the above section and those represented in the pairs. Of the top pairwise results the majority are lower than the additive value in breast, colorectal and prostate cancer. In Lung cancer however all of the top results exceed the additive VIMP

The results in **Table 19** highlight that in breast cancer only one of the breast cancer results has a pairwise VIMP excess of over 0.002[109], the threshold for highly predictive results. None meet this threshold in colorectal cancer, all meet it in lung cancer and all but one meet it in prostate cancer.

Table 20 identifies the five most suppressed pairwise VIMP scores. Here results show that the pairwise VIMP is less than the sum of the single VIMP scores. Overall the level of predictive suppression seen is greater than the level of predictive value seen in the largest differences in **Table 19**. The exception to this is in lung cancer

	Variable 1	Variable 2	Variable 1 VIMP	Variable 2 VIMP	Paired	Additive	Difference
Breast	Age	Stage	0.09529	0.07154	0.17036	0.16683	0.00353
	Age	Grade	0.09529	0.02530	0.10666	0.12059	-0.01393
	Stage	Grade	0.07154	0.02530	0.09144	0.09684	-0.00540
	Age	Deprivation score (IMD)	0.09529	0.00382	0.09119	0.09911	-0.00792
	Age	Coronary Artery Disease	0.09529	0.00248	0.08816	0.09777	-0.00961
Colorectal	Stage	Age	0.11997	0.05890	0.17381	0.17887	-0.00506
	Stage	Grade	0.11997	0.00693	0.12610	0.12690	-0.00080
	Stage	Deprivation score (IMD)	0.11997	0.00389	0.11911	0.12386	-0.00475
	Stage	Congestive Cardiac Failure	0.11997	0.00215	0.11816	0.12212	-0.00396
	Stage	Arrhythmia	0.11997	0.00217	0.11792	0.12214	-0.00422
Lung	Stage	Histological Subtype	0.17669	0.02072	0.19913	0.19741	0.00172
	Stage	Age	0.17669	0.01134	0.18853	0.18803	0.00050
	Stage	Obesity	0.17669	0.00095	0.18095	0.17764	0.00331
	Stage	Coronary Artery Disease	0.17669	0.00147	0.17997	0.17816	0.00181
	Stage	Congestive Cardiac Failure	0.17669	0.00103	0.17974	0.17772	0.00202
Prostate	Stage	Grade	0.12384	0.04163	0.18613	0.16547	0.02066
	Stage	Age	0.12384	0.02652	0.15078	0.15036	0.00042
	Stage	Diabetes	0.12384	0.00812	0.11644	0.13196	-0.01552
	Stage	Other Chronic Respiratory Disorder	0.12384	0.00000	0.11415	0.12384	-0.00969
	Stage	Inflammatory Bowel Disease	0.12384	0.00169	0.11396	0.12553	-0.01157

Table 18: Top 5 Pairwise VIMP Results in Each Cancer Site Specific Cohort *- The table includes the site specific cohort the analysis was conducted in, the pair of variables for which the results are presented, the VIMP of the first variable on its own, the VIMP for the second variable on its own, the paired VIMP is where both variables are permuted simultaneously, the sum of the VIMP results for the two variables when estimated alone (additive) and the difference between the pairwise and additive VIMP*

	Variable 1	Variable 2	Variable 1 VIMP	Variable 2 VIMP	Paired	Additive	Difference
Breast	Age	Stage	0.09529	0.07154	0.17036	0.16683	0.00353
	Chronic Renal Disease	Chronic Liver Disease	0.00069	0.00009	0.00101	0.00078	0.00023
	Chronic Renal Disease	Neuromuscular	0.00069	0.00003	0.00087	0.00072	0.00015
	Chronic Renal Disease	Restrictive Lung Disease	0.00069	0.00001	0.00084	0.00070	0.00014
	Chronic Renal Disease	Cardiomyopathy	0.00069	0.00000	0.00083	0.00069	0.00014
Colorectal	Paraplegia	Parkinson's Disease	0.00009	0.00000	0.00012	0.00009	0.00003
	Paraplegia	Malabsorption	0.00009	0.00002	0.00014	0.00011	0.00003
	Paraplegia	Venous Disease	0.00009	0.00000	0.00011	0.00009	0.00002
	Paraplegia	Other Chronic Respiratory Disorder	0.00009	0.00000	0.00011	0.00009	0.00002
	Paraplegia	Motor Neuron Disease	0.00009	0.00000	0.00011	0.00009	0.00002
Lung	Stage	Obesity	0.17669	0.00095	0.18095	0.17764	0.00331
	Stage	Connective Tissue Disease	0.17669	-0.00005	0.17898	0.17664	0.00234
	Stage	Paraplegia	0.17669	0.00002	0.17904	0.17671	0.00233
	Stage	Restrictive Lung Disease	0.17669	0.00003	0.17904	0.17672	0.00232
	Stage	Peptic Ulcer Disease	0.17669	-0.00005	0.17895	0.17664	0.00231
Prostate	Stage	Grade	0.12384	0.04163	0.18613	0.16547	0.02066
	Age	COPD	0.02652	0.00050	0.03004	0.02702	0.00302
	Age	Deprivation score (IMD)	0.02652	-0.00754	0.02168	0.01898	0.00270
	Age	Chronic Renal Disease	0.02652	-0.00015	0.02842	0.02637	0.00205
	Inflammatory Bowel Disease	Arrhythmia	0.00169	-0.00067	0.00300	0.00102	0.00198

Table 19: Top 5 Feature Combinations in Each Cancer Site Specific Cohort Where Pairwise VIMP Exceeds Additive VIMP- *The table includes the site specific cohort the analysis was conducted in, the pair of variables for which the results are presented, the VIMP of the first variable on its own, the VIMP for the second variable on its own, the paired VIMP is where both variables are permuted simultaneously, the sum of the VIMP results for the two variables when estimated alone (additive) and the difference between the pairwise and additive VIMP*

	Variable 1	Variable 2	Variable 1 VIMP	Variable 2 VIMP	Paired	Additive	Difference
Breast	Grade	Histological Subtype	0.0253	0.00697	0.01523	0.03227	-0.01704
	Age	Histological Subtype	0.09529	0.00697	0.08667	0.10226	-0.01559
	Age	Grade	0.09529	0.0253	0.10666	0.12059	-0.01393
	Grade	Deprivation score (IMD)	0.0253	0.00382	0.0159	0.02912	-0.01322
	Stage	Histological Subtype	0.07154	0.00697	0.06606	0.07851	-0.01245
Colorectal	Age	Deprivation score (IMD)	0.0589	0.00389	0.04963	0.06279	-0.01316
	Age	Sex	0.0589	0.0038	0.04978	0.0627	-0.01292
	Age	Hypertension	0.0589	0.00099	0.04784	0.05989	-0.01205
	Age	Histological Subtype	0.0589	0.0019	0.04902	0.0608	-0.01178
	Age	Coronary Artery Disease	0.0589	0.00127	0.04848	0.06017	-0.01169
Lung	Histological Subtype	Grade	0.02072	0.00468	0.01635	0.0254	-0.00905
	Age	Grade	0.01134	0.00468	0.00851	0.01602	-0.00751
	Histological Subtype	COPD	0.02072	0.00093	0.01532	0.02165	-0.00633
	Histological Subtype	Arrhythmia	0.02072	0.00051	0.01508	0.02123	-0.00615
	Histological Subtype	Hyperlipidaemia	0.02072	0.00079	0.01546	0.02151	-0.00605
Prostate	Grade	Hypertension	0.04163	0.00491	0.02543	0.04654	-0.02111
	Grade	Diabetes	0.04163	0.00812	0.03054	0.04975	-0.01921
	Grade	Coronary Artery Disease	0.04163	0.00325	0.02631	0.04488	-0.01857
	Stage	Hyperlipidaemia	0.12384	0.00288	0.10903	0.12672	-0.01769
	Grade	Hyperlipidaemia	0.04163	0.00288	0.02683	0.04451	-0.01768

Table 20: Top 5 Feature Combinations in Each Cancer Site Specific Cohort where additive VIMP exceeds pairwise VIMP- *The table includes the site specific cohort the analysis was conducted in, the pair of variables for which the results are presented, the VIMP of the first variable on its own, the VIMP for the second variable on its own, the paired VIMP is where both variables are permuted simultaneously, the sum of the VIMP results for the two variables when estimated alone (additive) and the difference between the pairwise and additive VIMP*

There were no instances where both variables are singly deemed to be non-predictive but are predictive when combined

7.2.5 - Partial Plots

Age:

In the breast (**Figure 76**), colorectal (**Figure 77**) and prostate cancer (**Figure 79**) cohorts a non-linear relationship between age and survival outcomes it identified. In the breast cancer cohort this relationship suggests that younger patients below the age of 50 have worse outcomes than those between 50 and 69. Those aged 70 or over however, have worse outcomes than both younger groups. As time since diagnosis increases, the patterns become more apparent with greater survival differences identified within these age bands. The median probability of survival at 10 years for patients 30-39 was estimated to be 75%, 82% for 50-59 year olds and 56% for 70-79 year olds.

Within the colorectal cancer population, discerning a clear pattern is made more challenging by the higher levels of uncertainty introduced through the low numbers of patients within the younger age bands. The point estimate identifies a similar pattern as seen in breast cancer with worse outcomes in patients under 40, the best outcomes in patients aged from 40-59 before an increasing impact of age with each subsequent decade. The confidence interval of the curve applied in **Figure 77** shows that the upper bounds of the confidence interval for the younger age groups overlap with the upper confidence bound of the middle age patients. Thus it is unclear if the worse outcomes suggested are a meaningful pattern or not.

The patterns seen within the lung cancer population appears to be linear, with no large scale shifts from this identified (**Figure 78**). The survival differences between ages are more apparent early on in disease follow up, with outcomes differences becoming less pronounced across age groups with increasing time. This is demonstrated via the falling incline of the overall trend line. Despite this, younger patients still have more favourable predicted survival with median predicted survival of 7% at 10 years for patients in their 30s versus 4% for patients in their 80s.

Within the prostate cancer cohort there is a gradual decrease in the median survival probability with increasing age in a linear fashion until the age of 80 or above (**Figure 79**). At this age there is a large and rapid fall off in survival probability which becomes more marked with increasing time after cancer diagnosis. At 5 years the median probability of survival is 89% for patients 70-79 falling to 61% in those aged 80-89. By 10 years post follow up the linear trend with increasing age transitions to a sharp fall off at an earlier age of 70 with median predicted survival of 91% for patients aged 60-69, falling to 62% for patients aged 70-79.

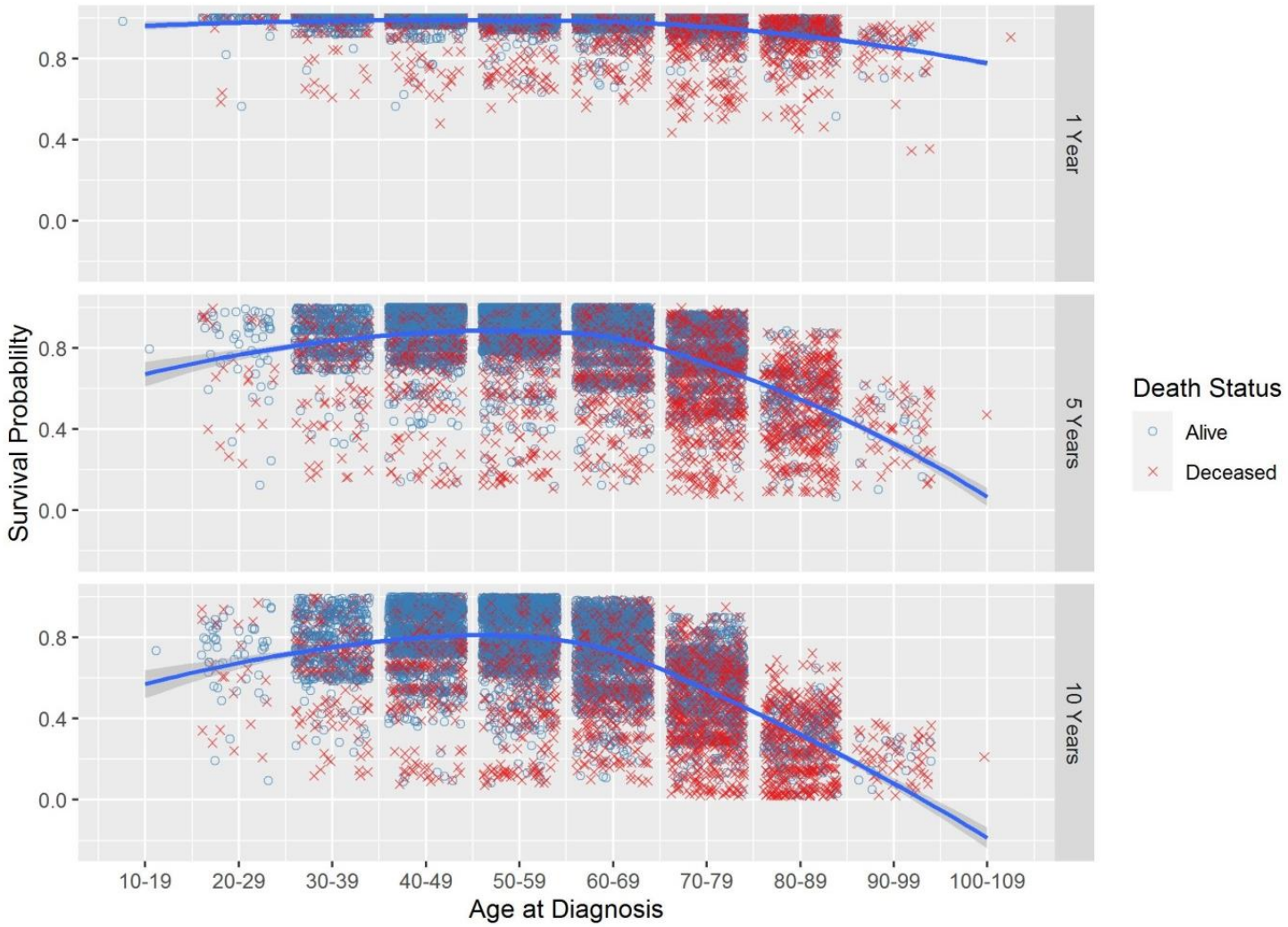


Figure 76: Relationship Between Age and Predicted Survival in Breast Cancer at 1, 5 and 10 Years - *Individual patient's predicted probability of survival derived from breast cancer random survival forest model is plotted against their age band. Patients who were alive at that point are plotted as blue circles. Those that were deceased are plotted as red crosses. The line is a smooth loess curve fitted to the data.*

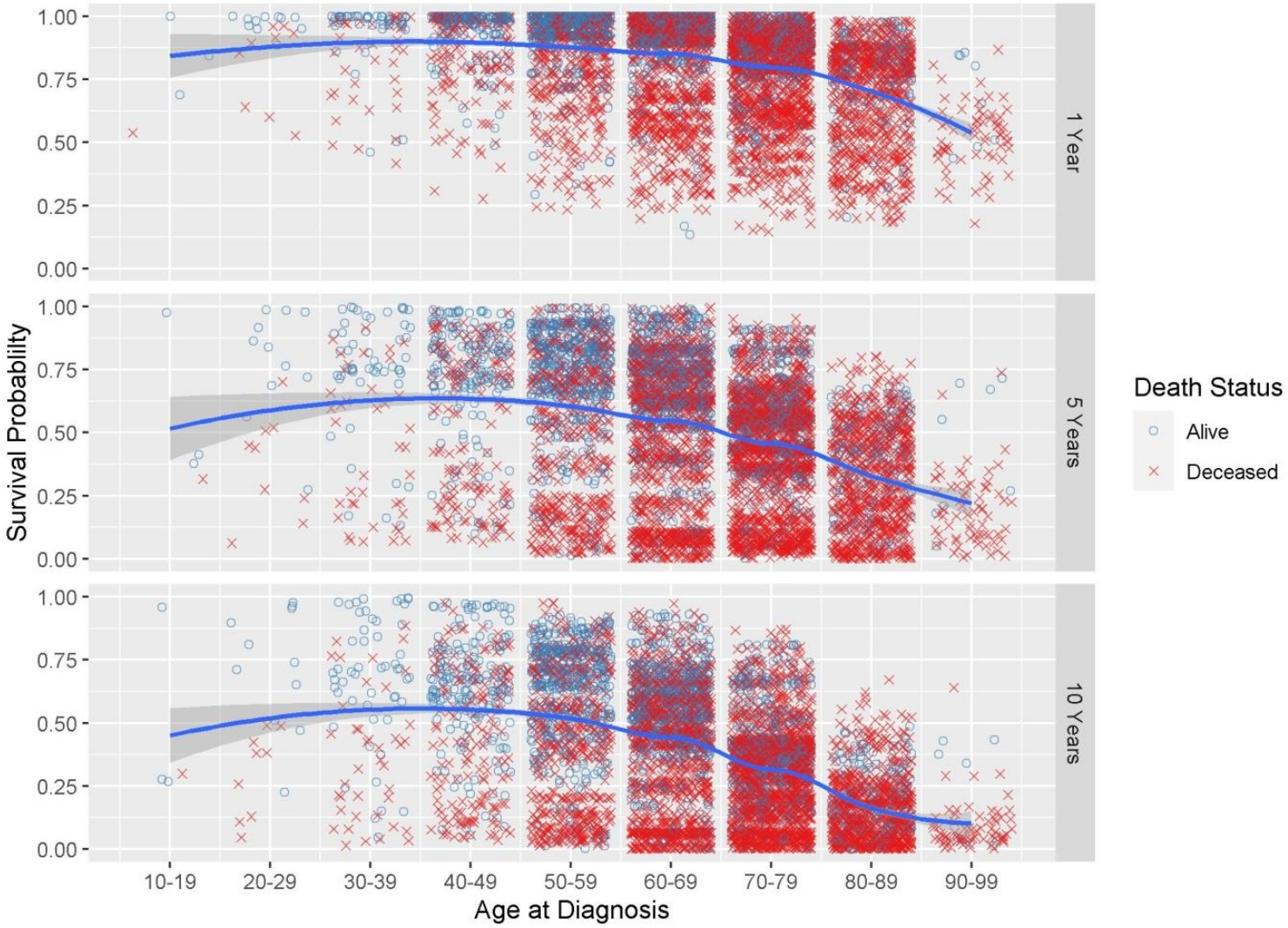


Figure 77: Relationship Between Age and Predicted Survival in Colorectal Cancer at 1, 5 and 10 Years *- Individual patient's predicted probability of survival derived from colorectal cancer random survival forest model is plotted against their age band. Patients who were alive at that point are plotted as blue circles. Those that were deceased are plotted as red crosses. The line is a smooth loess curve fitted to the data.*

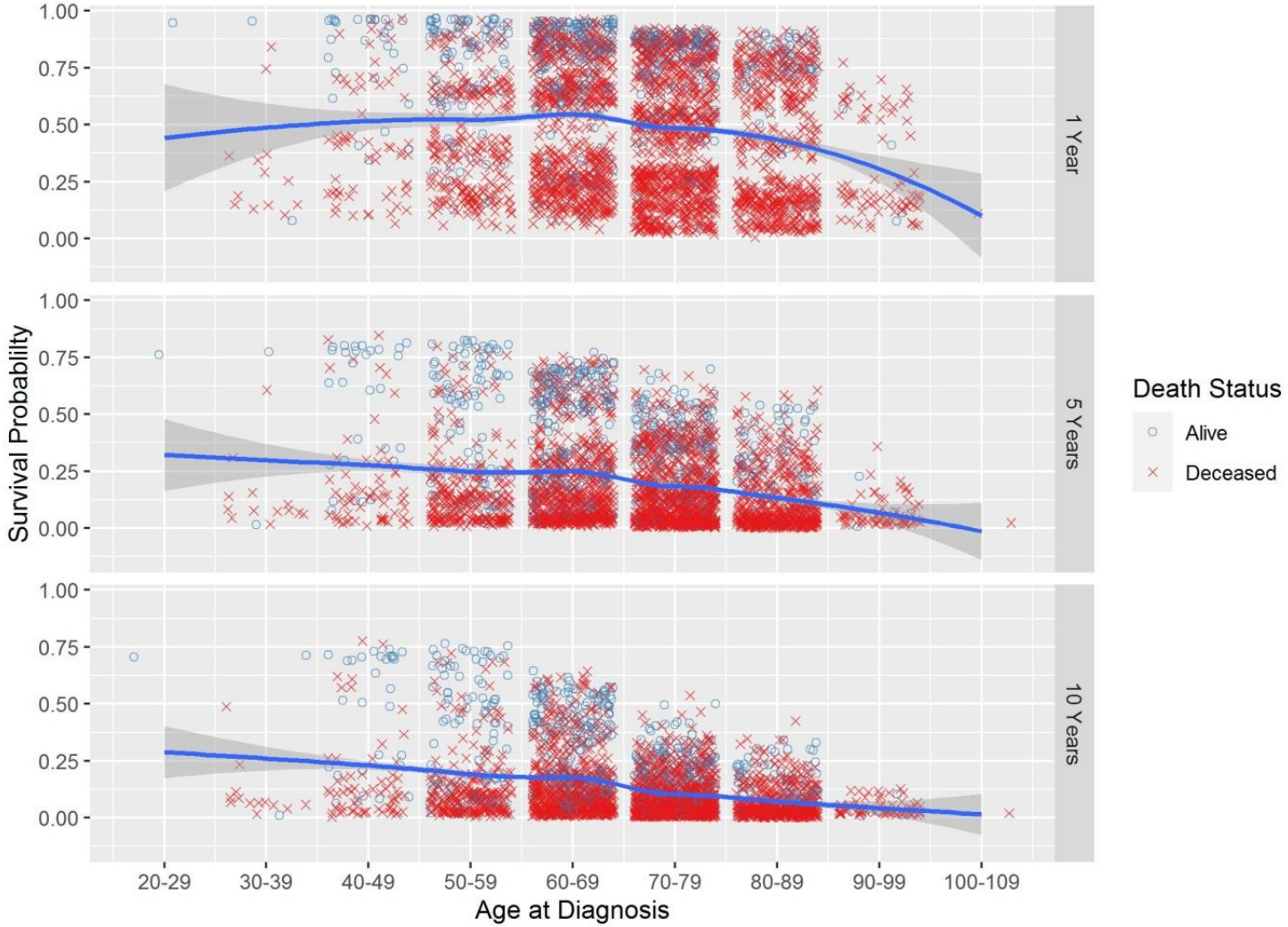


Figure 78: Relationship Between Age and Predicted Survival in Lung Cancer at 1, 5 and 10 Years *- Individual patient's predicted probability of survival derived from lung cancer random survival forest model is plotted against their age band. Patients who were alive at that point are plotted as blue circles. Those that were deceased are plotted as red crosses. The line is a smooth loess curve fitted to the data.*

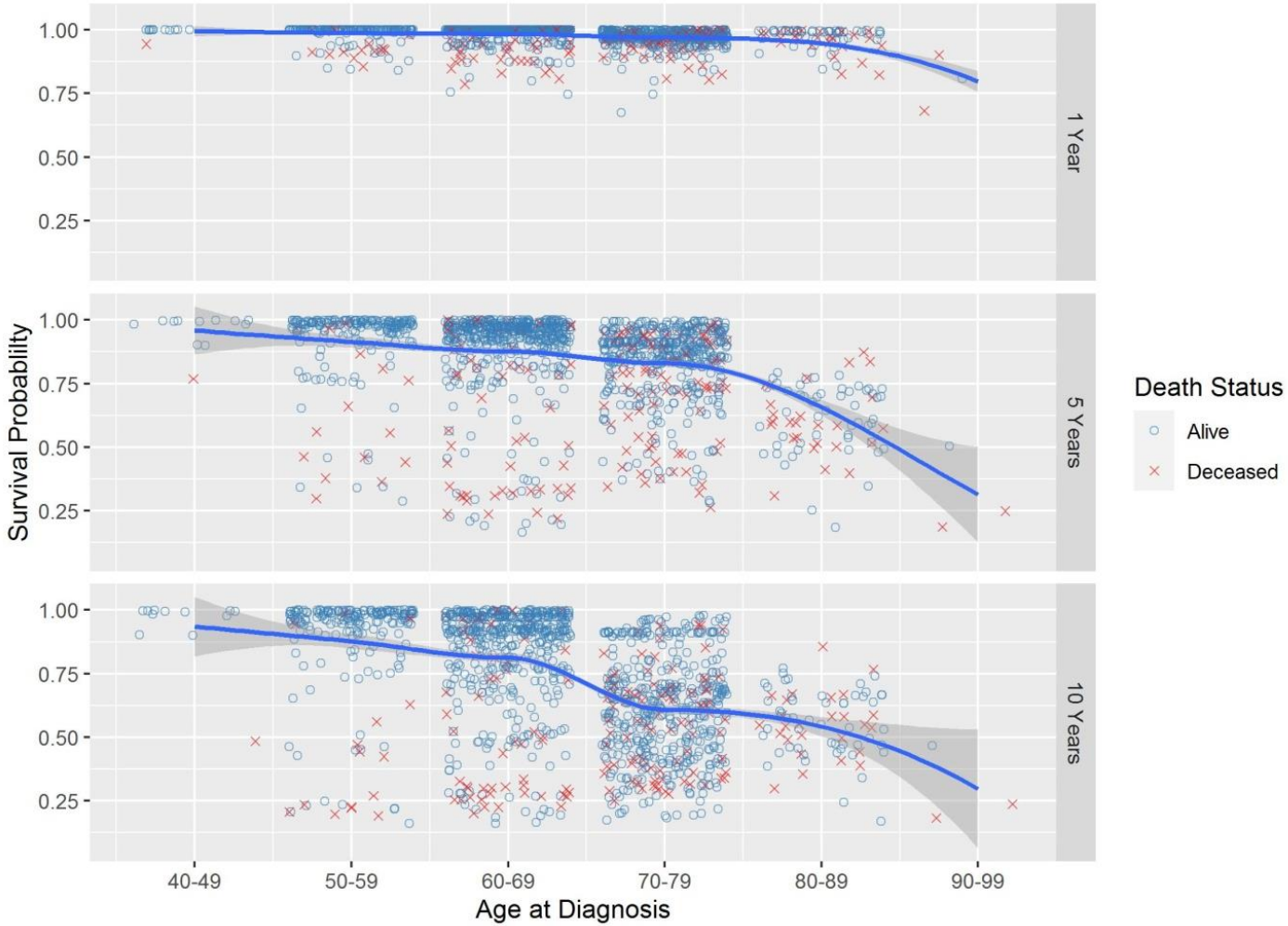


Figure 79: Relationship between Age and Predicted Survival in Prostate Cancer at 1, 5 and 10 Years *- Individual patient's predicted probability of survival derived from prostate cancer random survival forest model is plotted against their age band. Patients who were alive at that point are plotted as blue circles. Those that were deceased are plotted as red crosses. The line is a smooth loess curve fitted to the data.*

Stage

A similar pattern of results is seen across all the cancers. With increasing stage there is a corresponding fall in predicted survival. Of note however, the survival difference is not consistent in magnitude with increasing stage and is not consistent over time. In the case of breast cancer at one year although there is a decrease in predicted survival of stage 2 patients, relative to stage one, and stage 3 ,relative to stage 2, these differences are relatively small with median predicted survival for stage 3 showing a 2% absolute reduction compared to grade 1. Stage 4 however, even at 1 year, shows a large fall off in median predicted survival with a 31% absolute reduction in probability. It is however important to note that there is higher variance in the stage 4 results than stage 1-3 at one year. As time from diagnosis increases so too do the differences between stages. At 5 years the median predicted survival for Stage 2 is 13% lower than grade 1 and stage 3 is 22% lower than stage 2. Stage 4 shows an even larger fall with median predicted survival being just 21%. At ten years post diagnosis, median predicted survival for stage 1 is estimated to be 88%, 71% for stage 2, 46% for stage 3 and just 13% for stage 4. With increasing time from diagnosis the variance for each of the stages also increases.

In the prostate cancer cohort the survival differences seen by stage is modest even in the highest stages of patients, with stage 4 disease showing an absolute reduction in median predicted survival of just 7%. The differences seen increase over time, but in stages 1-3 the differences seen are still modest such that at 5 years post diagnosis median predicted survival for stage 2 and 3 is 1% lower than stage 1. Stage 4 however shows a large absolute decrease in median predicted survival, falling by 40% to 54%. Even at 10 years post diagnosis stage 1 and stage 2 show similar outcomes with a median predicted survival of 92% and 91% respectively. Stage 3 however shows a larger fall off in median predicted survival dropping to 78%. Stage 4 is however markedly worse with a median predicted survival of 42%. As with the breast cancer results with increasing time from diagnosis the variance of predicted survival increases across all stages. Stage 4 however shows higher variance at year 1 and year 5 however has lower variance than the other stages at 10 years.

Within the colorectal cancer cohort the effect of stage is more pronounced at year 1 compared to breast and prostate cancer. Median predicted survival for stage 2 and 3 is 5% lower than that of stage 1. Stage 4 is markedly lower at 56% representing a further absolute reduction of 34%. At 5 years post diagnosis clear differences are seen between all stages with median predicted survival being 81% for stage 1, 62% for stage 2, 52% for stage 3 and 10% for stage 4. By ten years the outcome differences continue with median predicted survival results of 66% for stage 1, 48% for stage 2, 40% for stage 3 and 6% for stage 4. The spreads of predicted survival in stages 1-3 increases with time from diagnosis. Of note, stage 4 results show the converse with decreasing variance in predicted survival with increasing time from diagnosis.

Lung cancer patients show the largest predicted survival differences by stage at 1 year. Stage 1 patients have a median predicted survival of 85%, 75% for stage 2, 53% for stage 3 and 18% for stage 4. The differences between stages narrows over time due to reduced survival in the earlier stages of disease such that at 10 year post diagnosis median predicted survival for stage 1 is 19%, 20% for stage 2, 18% for stage 3 and 3% for stage 4.

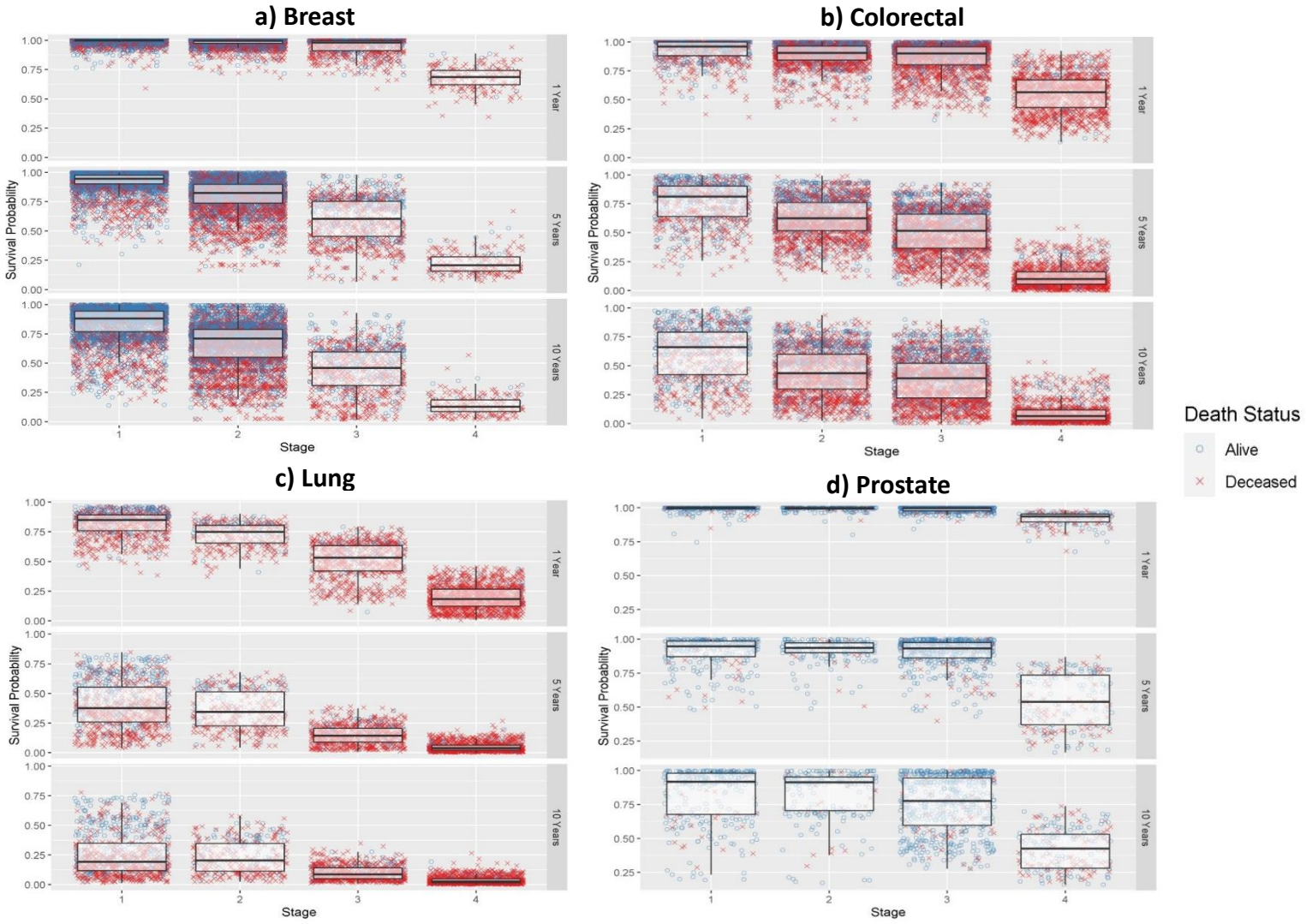


Figure 80: Partial Plots for Cancer Stage at Presentation *– Each individual plot represents the predicted survival derived from a random survival forest model for each patient at 1, 5 and 10 years with increasing stage at presentation along the horizontal axis. A box and whisker plot is the overlaid to demonstrate the distribution of predictions. Results are shown for the four cancer site specific models.*

Grade:

The partial dependence plots demonstrate that there is a corresponding decrease in predicted survival with increasing grade. This increase is shown to be inconsistent between grades, times and cancers. Within the breast cancer cohort, the association between grade and survival is minimal at 1 year but increases over time. At 1 year the median predicted survival for low grade tumours is 100% compared to 99% for high grade tumours. At five years, the disparity grows such that low grade tumours have a median predicted survival of 96% versus 89% for intermediate grade and 78% for high grade tumours. By 10 years, low grade has a median predicted survival of 91%, versus 76% for intermediate grade and 67% for high grade. Across all grades the variance of the survival predictions increases with increasing time since diagnosis.

The colorectal cancer cohort shows falling survival probability with increasing grade although the results are more marked at 1 year than in breast cancer. At one year post diagnosis low grade patients' have a median predicted survival of 92% compared to 90% for intermediate grade tumours and 73% for high grade tumours. By 5 years this falls to 61% for low grade, 56% for intermediate grade and 35% for high grade. At 10 years this falls further to 44% for low grade, 39% for intermediate grade and 35% for high grade. These results demonstrate that at 1 and 5 years the effect of predicted survival of high grade is far larger than that of intermediate grade. This difference does however shrink over time demonstrating that the impact is inconsistent with regards to time and increasing grade.

Within the lung cancer cohort the pattern is similar to that seen with increasing stage, in that the differences are most apparent earlier in the follow up period and become less marked as time goes on in terms of absolute survival. At 1 year median predicted survival is 78% in low grade tumours, 64% in intermediate grade and 46% in high grade. By 10 years this falls to 19% for low grade, 11% for intermediate grade and 6% for high grade. Despite the fall off in absolute difference in survival, the relative difference has in fact increased from 41.1% to 68.4% when comparing low to high grade at 1 and 10 years.

The comparisons for grade in prostate cancer are slightly different due to the use of the Gleason grading system. This results in grades from 6-10, as opposed to the low, intermediate and high grades for the other cancers sites presented above. With increasing grade there is evidence of a reduction is predicted survival on average. This difference increases over time with meaningful differences being demonstrated in both relative and absolute terms. At 1 year Gleason 6 patients have a median predicted survival of 100% versus 96% for Gleason 9 and 83% for Gleason 10. By 10 years survival predications fall to 97% for grade 6, 47% for grade 9 and 29% for 10.

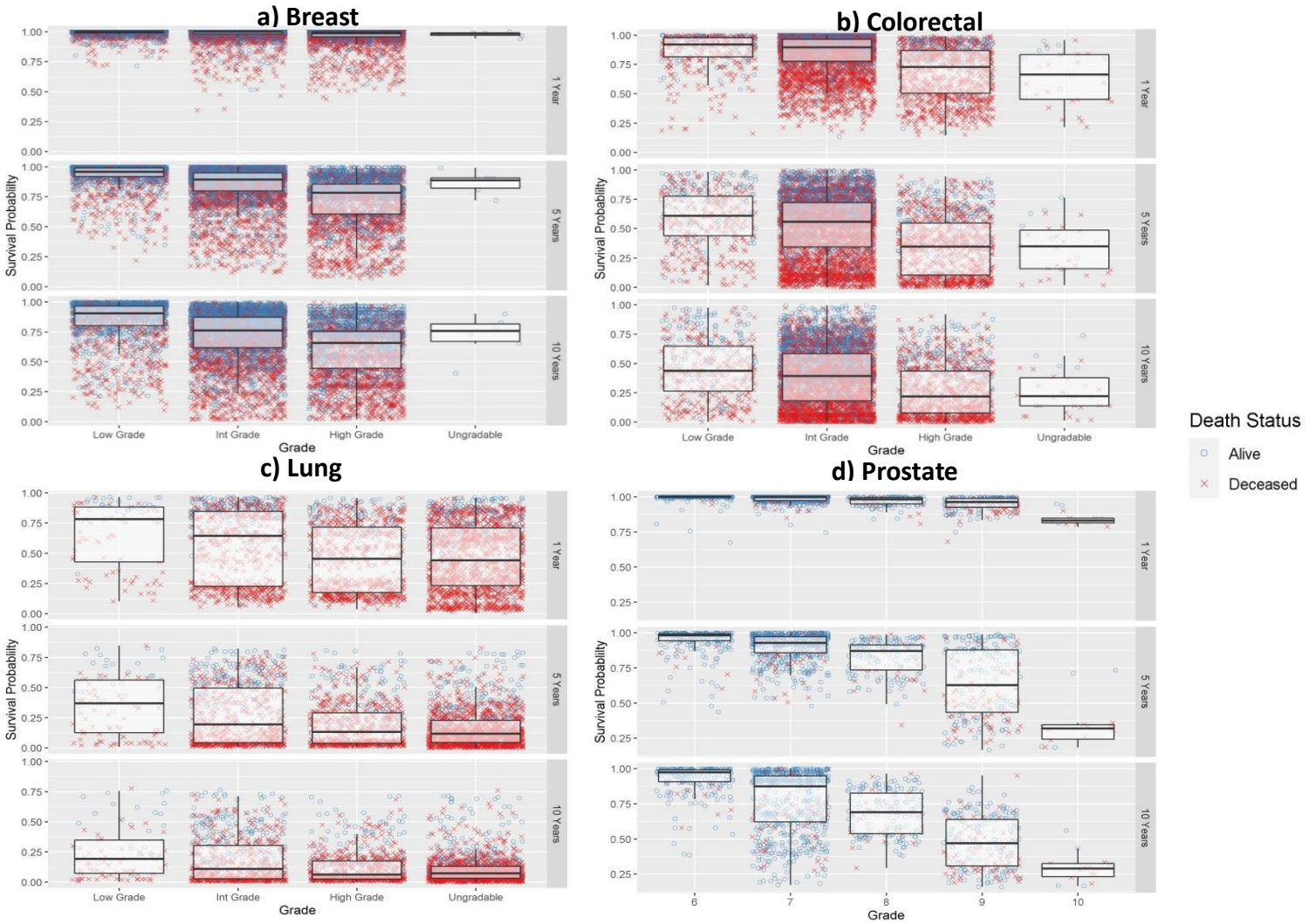


Figure 81: Partial Plots for Cancer Grade at Presentation *– Each individual plot represents the predicted survival derived from a random survival forest model for each patient at 1, 5 and 10 years with increasing grade at presentation along the horizontal axis. A box and whisker plot is the overlaid to demonstrate the distribution of predictions. Results are shown for the four cancer site specific models.*

7.2.6 - Survival Curves

The models generated for each site were used to derive survival curves for each of the key comorbidities of interest. CCF was associated with worse outcomes across all four cancer sites although the degree of effect is different for each (**Figure 82**). The largest survival difference is seen in breast cancer where the effect grows over time until the end of the analysis window. In lung cancer the effect of CCF increases with the greatest impact occurring in the first year before the impact gradually decreases until approximately 7 years after diagnosis where the effects plateau. In prostate cancer the pattern is less smooth and thus more difficult to interpret. Overall, CCF patients do worse, although the effects appear most marked from 26 months onwards. In colorectal cancer the association with adverse outcomes is greatest in the first 6 months before slowing gradually and reaching a plateau at 8 years post diagnosis.

The MI stratified curves show a similar patterns with breast cancer continuing to show increased mortality associated with MI patients throughout the analysis period (**Figure 83**). Lung cancer showing the greatest association with increased adverse outcomes in the first 6 months before slowing to a plateau between 8 and 9 years. Prostate cancer appears to show on ongoing association with adverse outcomes in the patients with previous MI throughout the analysis period however as with the CCF results, the lack of a smooth curve makes interpretation less precise. Colorectal cancer demonstrates that again MI is associated with a lower probability of survival with the net deaths associated with MI being at the highest rate for the first 6 months before levelling out and a falling rate of death plateauing at around 100 months post diagnosis.

The results for COPD are less consistent with breast, prostate and colorectal cancers all showing an association between COPD and adverse outcomes (**Figure 84**). In the case of lung cancer, the results echo those shown in chapters 4 and 5 with COPD being associated with improved survival. The improvements in predicted survival do however only appear to be apparent from 6 months to 9 years post cancer diagnosis. Before and after this time window the survival trajectory of patients both with and without COPD appear similar.

The predicted outcomes for stroke patients are worse in breast, lung and colorectal cancer (**Figure 85**). The effect size appears to be more modest in the case of lung cancer compared to breast and colorectal cancer. The pattern of results in prostate cancer is again difficult to interpret given the jagged nature of the curve, however would appear to suggest no clear difference in the prostate cancer patients with previous stroke.

Diabetes is shown to be associated with worse survival predictions in all four cancer sites (**Figure 86**). Breas cancer shows the largest survival disadvantage followed by colorectal cancer and prostate cancer. Lung cancer predicted outcomes are least associated with DM.

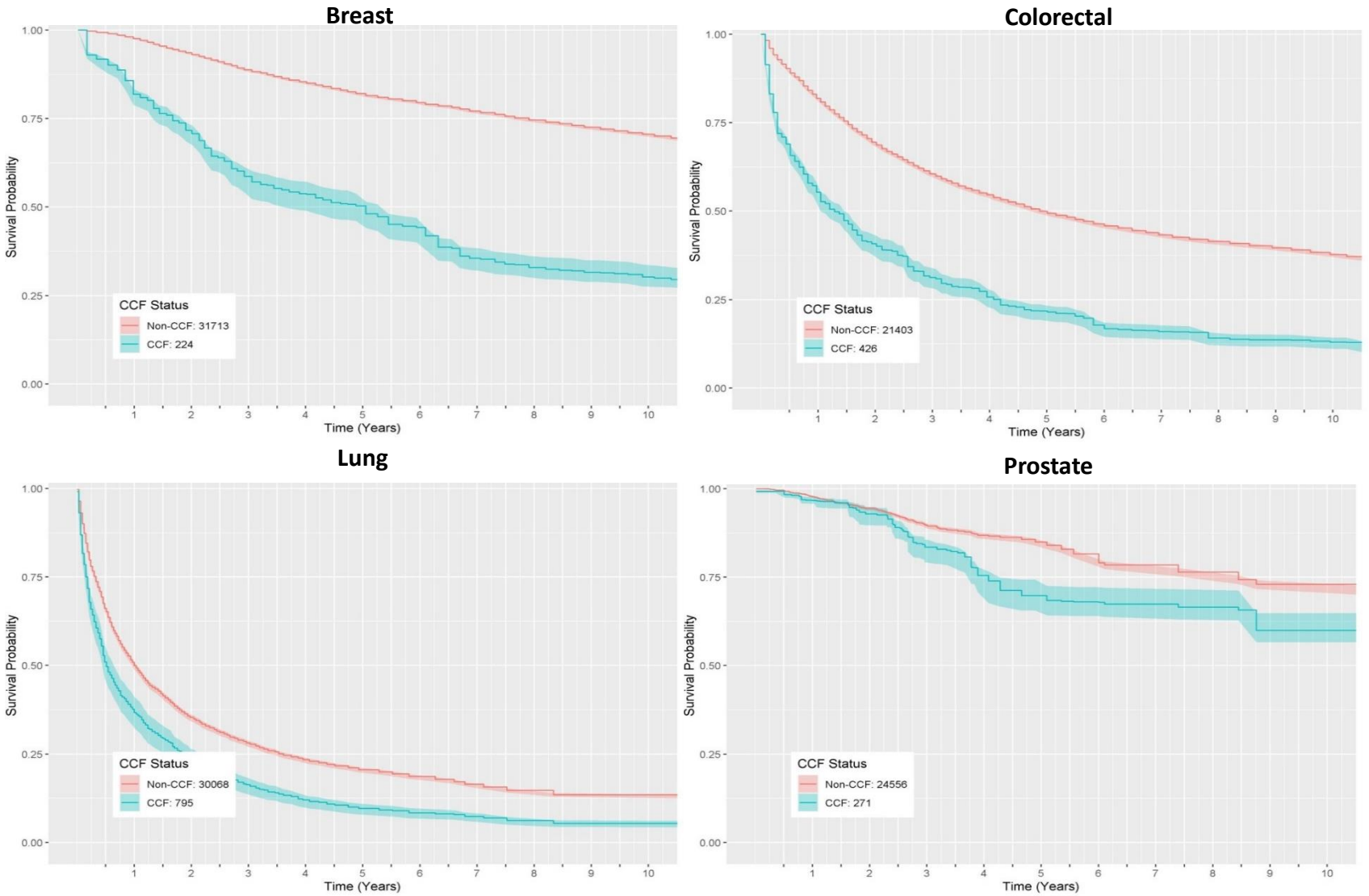


Figure 82: RSF Derived Stratified Survival Curves for Patients with Prior CCF *– Stratified survival curves generated from each of the site specific random survival forest models.*

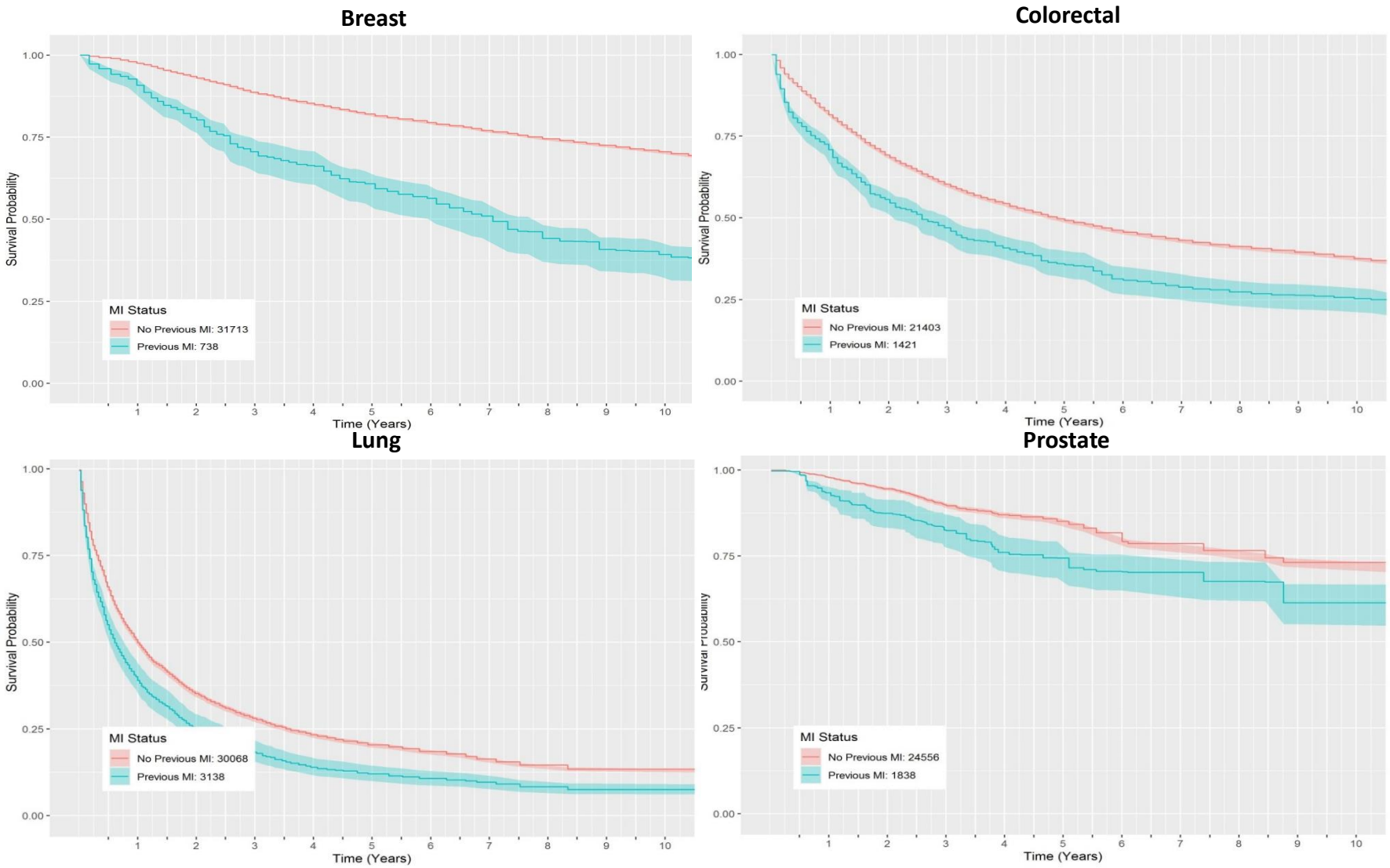


Figure 83: RSF Derived Stratified Survival Curves for Patients with Prior MI– *Stratified survival curves generated from each of the site specific random survival forest models*

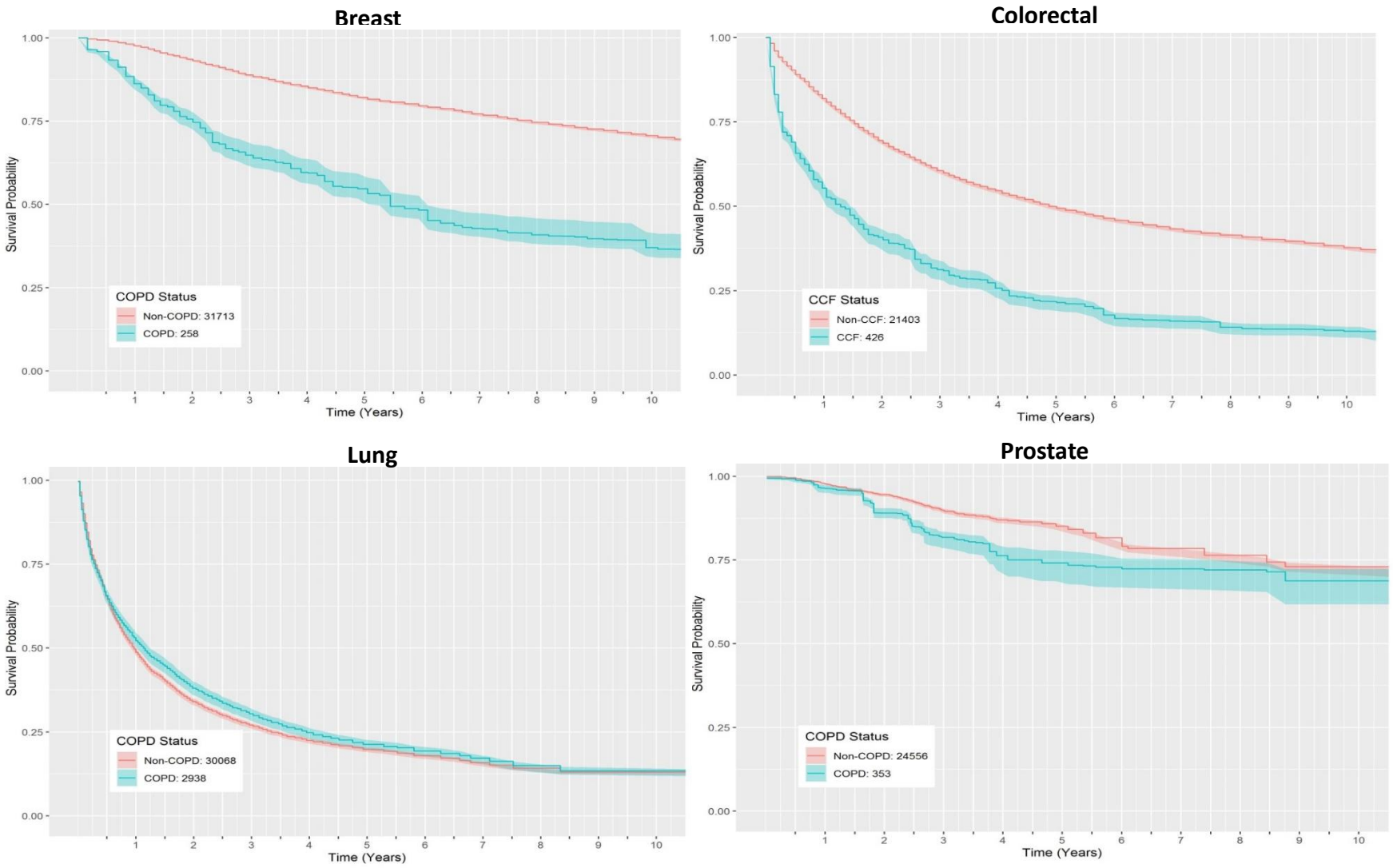


Figure 84: RSF Derived Stratified Survival Curves for Patients with Prior COPD *– Stratified survival curves generated from each of the site specific random survival forest models.*

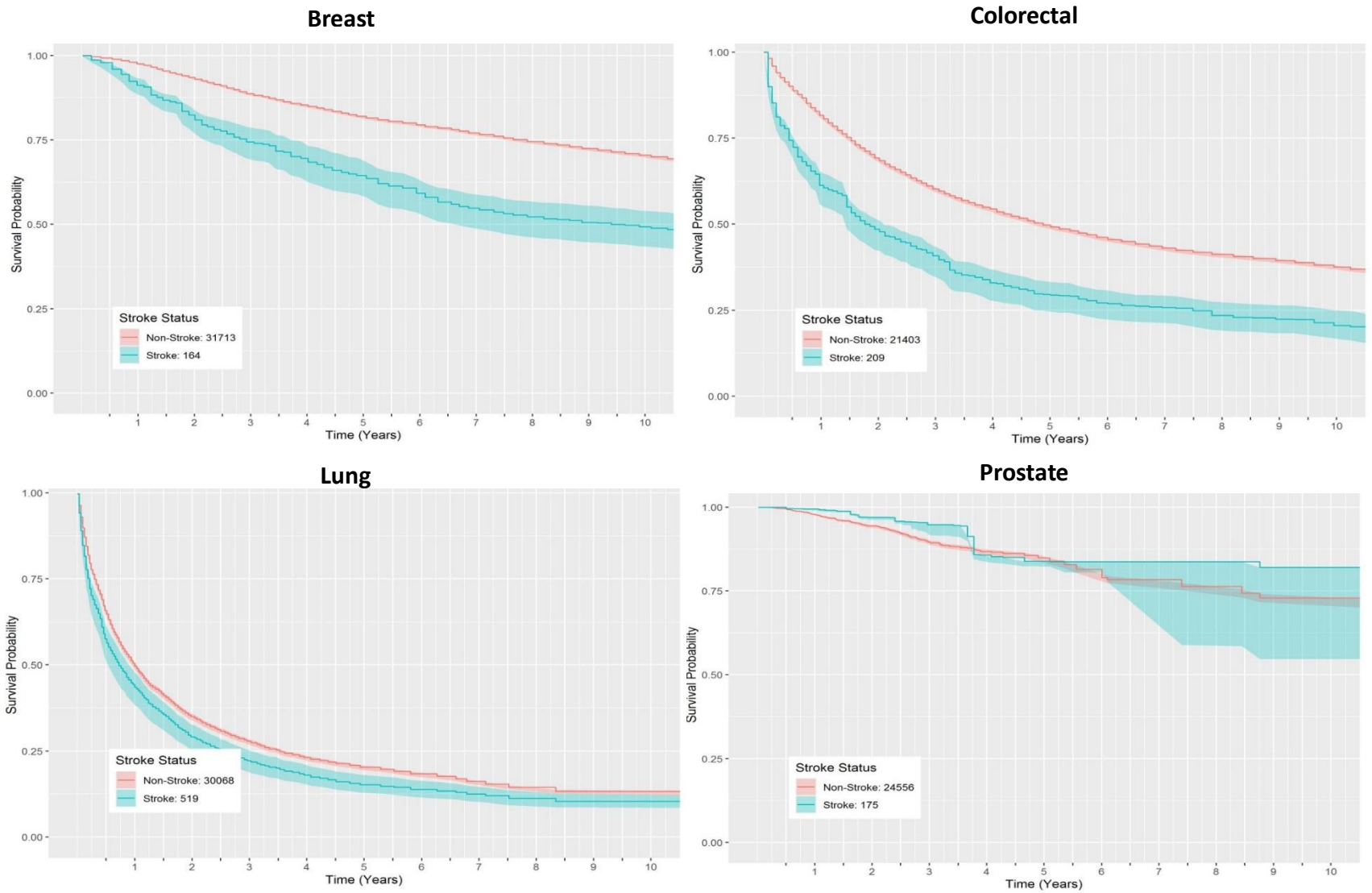


Figure 85: RSF Derived Stratified Survival Curves for Patients with Prior Stroke *– Stratified survival curves generated from each of the site specific random survival forest models.*

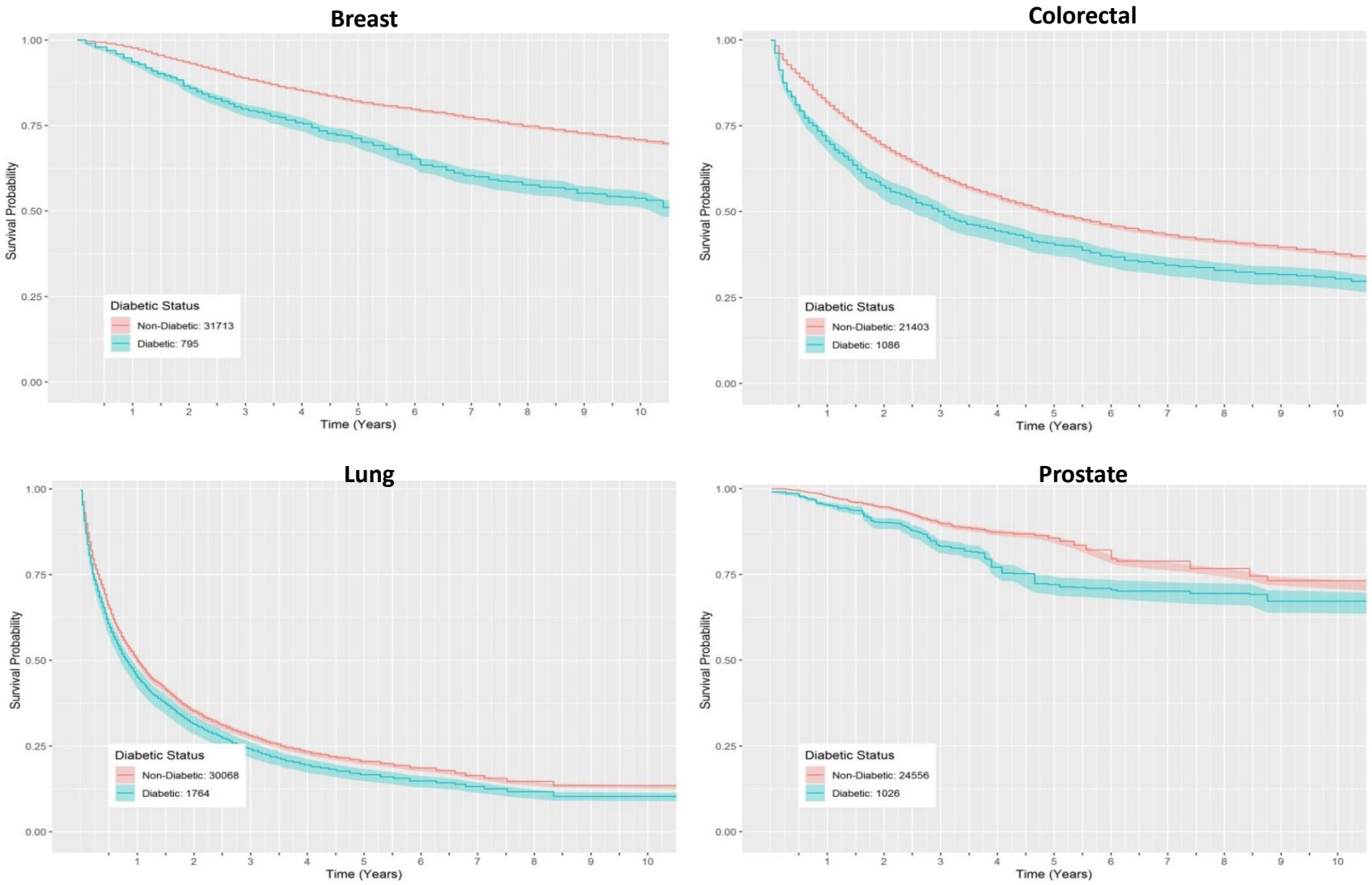


Figure 86: RSF Derived Stratified Survival Curves for Patients with Prior DM *– Stratified survival curves generated from each of the site specific random survival forest models.*

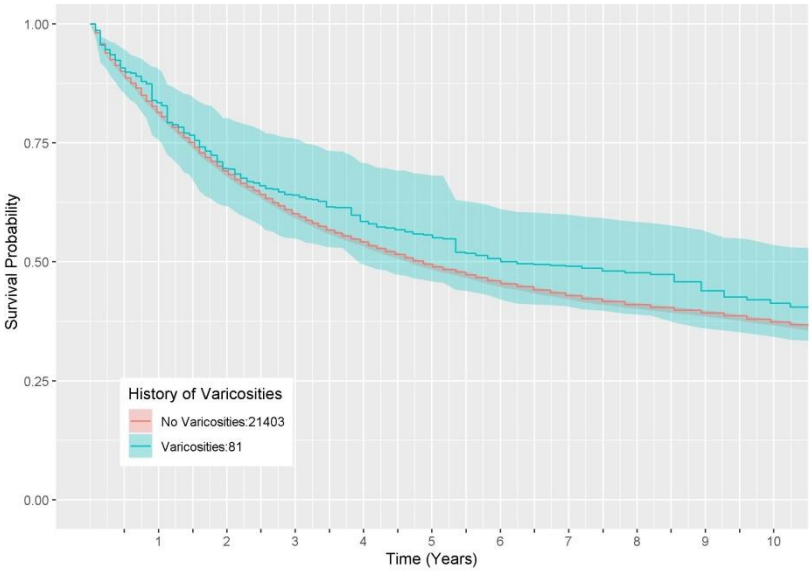


Figure 87: RSF Derived Stratified Survival Curve Varicosities in Colorectal Cancer *– Stratified survival curve for patients with and without varicosities derived from the Colorectal Cancer Site Specific Cohort RSF*

The stratified survival curves focussing on varicosities in colorectal cancer and obesity in lung cancer are shown in **Figure 87** and **Figure 88**. In the case of varicosities although the general trend suggests that patients with varicose veins have an association with worse outcomes the lack of precision of these estimates results in the curves overlapping therefore not providing evidence of different outcomes in this group. Obese lung cancer patients are however shown to have marked and consistently better predicted outcomes than their non-obese comparator group. These results echo those derived from the Cox modelling undertaken in chapter 5.

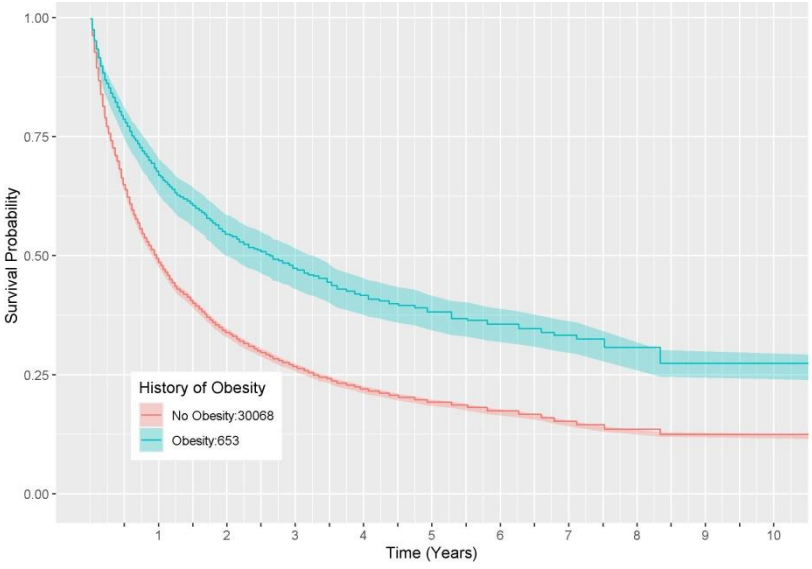


Figure 88: RSF Derived Stratified Survival Curve Obesity in Lung Cancer *– Stratified survival curve for patients with and without obesity derived from the Colorectal Cancer Site Specific Cohort RSF*

7.3 - Discussion

7.3.1 - Optimisation:

For results to be as informative as possible within a predictive framework of analysis the models developed need to have the greatest possible accuracy. As such hyperparameter tuning is a crucial step in developing all subsequent model output.[189] If a model is not sufficiently accurate in determining the true outcome for a patient or group of patients, then the relationships between features and predictions cannot be relied upon. The methods that have been implemented relied upon the fast forest approximation method. Although computationally efficient, the results provided are an estimate of the accuracy not the true accuracy. This is however accepted practice within the field.

If resources and time were unconstrained than a manual process of building full random forests with each hyperparameter combination would have been undertaken. These could then have been tested on the holdout dataset to identify the most accurate model. The implementation of the random survival forest algorithm is such that increasing numbers of variables have no effect on CPU runtime, increasing numbers of trees results in a linear increase in CPU runtime and increasing numbers of patients, results in an exponential increase in CPU run time. As the analyses are focussed on large cohorts the run time of the algorithm training is substantial taking around 20-30 minutes per model. To reach the same optimisation parameters identified using the fast forest approach would have taken approximately 20 days using full forest development. As a result, full optimisation using fully trained models was not possible due to computational and time limitations. This does however mean that in each case we have what we approximated to be the optimal hyperparameters, however there may be even better hyperparameter combinations that could be identified had full model training and testing been undertaken.

The analysis of the OOB error rate demonstrates that in each of the models fewer than 100 trees resulted in a steady state of error being achieved. This is relevant as it suggests that each model is comprised of a forest that is large enough in size to have yielded results that would not improve through increasing the size of the ensemble. The size of 500 does however exceed the steady error threshold suggesting that in future analysis it may be appropriate to reduce the forest size to for example 100-200 to improve the efficiency of the model development process.

7.3.2 - Predictive Accuracy

The comparison of the accuracy of the three methods identifies that the random survival forest approach outperforms both the Cox and KM methods. The reasonable scores shown when being assessed on the holdout dataset suggest that the RSF models are likely to generalise to a population beyond that on which it is trained.[54,380] It is however important to note that the results of chapter 3 highlight that the PPM population may be different to the wider cancer populations. As such the model may require further training using more representative data in order to perform well on other populations that include a different population make up. Despite the RSF approach being superior in its predictions, it could be argued that the comparison to Cox was not entirely valid as no optimisation was undertaken for the Cox model. A number of methods could have been employed including forwards and backwards elimination, using the VIMP to guide feature selection or a brute force approach (assessing all possible combinations of variables). It is therefore possible that the correct combination of features when included within a Cox model might outperform the random survival forest approach and this could be an area for further research.

The result showing enhanced performance of the RSF approach has been identified previously in other medical contexts.[378,386,387] Despite its advantages, the RSF approach has not been widely

adopted within the medical research domain. This may be due to it being less well established, a lack of implementation in commercial software offerings, less familiarity or a lack of applicability in causal estimation. The results of our analysis would seem to lend weight to the increased use of this approach in future medical research focussing on prediction. Future work could however focus on the use of condition inference forests[388] or deep neural network survival[389] approaches to assess how these compared in this analysis setting.

7.3.3 - Single Feature VIMP and Tree Depth

The use of permutation VIMP presents a number of distinct advantages.[188] As a measure of the model error introduced through the destruction of each feature's information in turn, it acts as a concise measure of insight into the whole model. In permuting each feature it removes the effect of that feature and all of its interactions allowing, that feature's global effects to be assessed (in random forests publications "interaction" is sometimes used to describe how the selection of one feature influences the subsequent use of other variables in split points, the use of interaction above is in the traditional sense[390]). Further the measure is derived directly from the model and does not require any further retraining to be calculated.

Despite this there are some disadvantages. Firstly the VIMP is integrally linked to the overall performance of the model, thus may be less informative in a poorly performing model. The reliance on randomness can result in VIMP scores that are different if repeated several times. Finally correlated features may result in the artificial deflation of VIMP such that although the model is not affected by collinearity, the VIMP score may be.[391] This is the case when two features encode the same information or some of the same information. Once one feature has been used for a given split the second variable adds no further information and this is deemed to not be predictive, although it is predictive in the absence of the other feature. Across the forest the two features may be used interchangeably and thus when permuted individually the forest may retain its performance.[190,392]

A further issue is whether the VIMP should be derived from the training data or the testing data. Using the training data increases the number of patients represented and removes the need for repeated subsampling of a smaller population. If however the model has overfit the data, there is a risk that using the training data might suggest a variable is important when it isn't.[393] Using the testing data overcomes the latter issue but introduces the first two. Within our analysis we have used the training dataset as we have demonstrated that the model is providing reasonable performance on the holdout dataset suggesting that there is generalisability[54] and thus the overfitting issues are a lesser concern.

When assessing the results of the single variable permutation VIMP and average minimum tree depth, it is important to understand that these measures are fundamentally different to one another, and reflect subtly different information. Minimum tree depth identifies important variables in a different way to VIMP, within the model development, at each node, the most discriminatory value of one of the selected candidate variables, is used to divide the cohort. The choice of variable and cut off value is optimised to maximise the survival difference between the groups created by the division. Thus the earlier a variable is used the greater the discriminatory power it has.[109]

To better conceptualise this it is possible to think of the game "20 questions", where an individual thinks of an object, item or person and the players have 20 questions with which to identify the thing that has been picked. Asking if the choice is a plant provides greater discrimination than asking if it is a rose, as in the former question a significant proportion of possible answers are removed no matter the response, whereas with the latter, if the item is not a rose all other answers are still possible. If as in the case with random survival forests the algorithm is optimised towards

discriminating two groups then the earlier the variable is used the greater information it is providing. Despite this, there are still a number of potential pitfalls. If a feature is continuous versus, ordinal versus binary the number of potential values for splitting by that feature changes. As such there is a greater likelihood of a variable being discriminatory where more potential values to delineate groups are available.[190] Age for example as a continuous variable as opposed to in 10 year age bands might be a better discriminator. Thus, when we interpret the results, it is unsurprising that in many cases comorbidities which are recorded as binary in our dataset, do not perform as well as other variables that are ordinal. Additionally where the number of cases of comorbidity are small the likelihood that the information about that variable is the most discriminatory is reduced. This is particularly true when there is variation in outcomes amongst the group with that condition. If the cohort had better representation of these conditions then the average minimum tree depth might also improve.

Across the models for breast, lung, prostate and colorectal cancer there was broadly agreement between the VIMP and average tree depth as to which variables were important. In each of the models one variable was shown to be contributing the most to predictive accuracy by a large margin, this being stage in prostate, colorectal and lung cancer, and age is breast cancer. Stage in breast cancer was shown to be the second highest VIMP score and age was in the top 5 for all models. The variables that would traditionally be used to determine outcomes based on clinical expertise which include, grade, stage, age and histological subtype were almost always in the highly predictive variable lists. Of interest, there are a number of occasions where variables were shown by VIMP to have low or no predictive importance that average minimum tree depth identified as providing significant information gain. These scenarios may represent occasions where VIMP is being impacted by collinearity.

Two significant drawbacks are present when applying either of these methods. Firstly it fails to capture interactions between variables as it incorporates the direct and interaction effect into one estimate. Secondly these scores do not provide any information as to the direction of effect that a given variable has. In order to address these two issues further analysis with pairwise VIMP and partial plots are demonstrated in the sections below.

7.3.4 - Pairwise VIMP

When analysing the pairwise VIMP there are a number of considerations. The most basic is the assessment of the ranking of the pairwise VIMP scores. This provides the most influential pairs of variables in determining the predictions. Further insight can however be gained by comparing the pairwise VIMP to the scores obtained from single variable VIMP permutation. These differences can be used to identify potential interactions or collinearity between these pairs of variables and can occur in a number of ways which are detailed below.

1) **Both variables are independently important, but their combined VIMP exceeds their additive VIMP.** This suggests that one or both of the variables are being deflated when assessed singly due to some degree of information overlap.
2) **Both variables are independently important, but their combined VIMP is less than their additive VIMP.** This suggests that there is an interaction between the two variables. As single VIMP removes all the direct effects and interaction effects the interaction effects are included in both single VIMP scores. As such the additive VIMP will exceed the pairwise VIMP.
3) **One variable is independently important but their combined VIMP exceeds the VIMP of independently important one.** Here this suggests that the single variable deemed to have

low importance may share overlap of predictive power with the other feature tested. This could be due to collinearity or both being a surrogate marker for other information.

4) **None are independently important but their combined VIMP is deemed predictive.** This could occur if both features encode the same information as one another about a patient either directly or indirectly. When testing only one at a time the predictive information is still available via the other feature in the analysis. If however they are both lost, then the shared encoding is also lost and thus they can be seen to be encoding important predictive information.

Due to the previously discussed issues around the potential for variation in VIMP scores the ideal results would be based on multiple calculations of VIMP and then taking an average across these. Unfortunately due to the computational time required to generate the pairwise VIMP this was impractical. Instead we will use the 0.002 threshold suggested by the previous research as being a cut off for VIMP suggesting a strong predictor.[109]

The results showing the top pairs of predictors includes results that are largely unsurprising (**Table 18**). Age, stage, grade and histology were shown to be strong predictors when assessed independently and are shown to be meaningful in previously published research.[394,395]

The results in **Table 19** identify cases where the pairwise VIMP exceeds the additive VIMP. As individual VIMP scores include the direct effect and interactions one would expect that if no interaction was present the pairwise and additive VIMP would be broadly the same. If an interaction effect was present then one would expect the pairwise VIMP to be lower as the interaction effect would be double counted in the additive VIMP. The results seen in this table therefore represent those where collinearity may have artificially supressed the individual VIMP score of one or more of the two variables. These pairs are therefore the ones whose collinearity is resulting in less reliable individualised VIMP scores.

Due to the low level of difference seen in the breast and colorectal cancer cohorts these results are unlikely to suggest the influence of collinearity. In the case of lung cancer however the top five all exceed the 0.002 threshold. All of these examples include stage and a comorbidity. This may suggest that in the context of lung cancer the stage data encodes information that an individual's comorbidities capture. Thus some of the results where individual comorbidities when assessed are suggested to be non-predictive or minimally predictive of outcomes may in fact have predictive value, but that the VIMP is being suppressed by correlation to stage data. The prostate cancer data is similar to that of lung in that 4 of the 5 results highlighted meet the 0.002 threshold. The fact that these two cancers are affected, one being a short prognosis cancer and the other a long prognosis cancer, suggests that this effect is not related to median survival time.

Table 20 shows results where the paired VIMP scores are lower than the additive VIMP. These suggest instances where there is an interaction between the two variables. In all cases these differences dramatically exceed 0.002 and the paired results exceed 0.002. This suggests that the variables are highly predictive together but have interactions that are in and of themselves predictive.

As the above section demonstrates, pairwise VIMP although potentially useful, is both difficult to interpret and inefficient. Although it would be possible to look at triplet VIMP and quadruplet VIMP this creates an exponential number of combinations that require assessment. This would be both computationally and analytically inefficient. Further work is therefore needed to apply additional

methods of model interpretability which might provide greater insight with a summary of the contribution of a given variable taking into account all other variables. An example of this would be the use of local interpretability with Shapley Values and global interpretability with global surrogate models.[396] Although still computationally complex these would provide an importance metric based on all possible combinations of other variables and thus be far simpler to interpret.

As with single VIMP scores the pairwise VIMP does not provide information on the pattern of association found and thus below we discuss the partial plots which demonstrate the patterns seen within our dataset.

7.3.5 - Partial Plots

When assessing the patterns seen within the partial plot analysis it is important to consider exactly what results are being reviewing and how this affects the interpretation of them . These plots are generated by creating a prediction of the survival probability for each of our patients at predefined times. In this analysis predictions were generated for 1 year, 5 years and 10 years post cancer diagnosis. The plots then represent the trend of predictions once grouped by a given characteristic. As the aim is to identify what characteristics can be used as an identifier of different or worse outcomes, this approach is a practical way of identifying this information. The estimates for each patient take into account all of their baseline characteristics, which are used to inform the predictions generated. It is however crucial to make the distinction between accounting for these factors and adjusting for them. Unlike in traditional methods such as Cox proportional hazards this approach is not designed to create an estimate of the effect of the feature being assessed, whilst removing the effect of the other included variables. As such when a pattern of outcomes is seen based on a given feature, e.g. age, it is important to understand that the relationship could be directly due to age, due entirely to other features correlated with age or a mixture of the two. It is not possible to identify the causal contributions of any given parameter using this approach. As such, the approach is in effect, using a predictive model to make inferences about how patients with a given characteristic may differ in terms of survival, but without making any attribution as to the cause for the difference that is identified.

Age

The results across the four cancer sites when assessing the relationship with age show signs of non-linear associations. This is consistent with the results presented in chapter 5 where Cox models were assessed for possible violations of the underlying linear assumptions. These partial plot analyses highlight the significant drawback of using a hazard ratio to describe the relationship between age and survival. These machine learning derived associations demonstrate that the relationship is far more complex and nuanced than traditional methods would suggest. This is particularly true in the case of breast cancer and may be true for colorectal cancer. These analyses suggest that younger patients may have worse outcomes than those who are middle aged. This is plausible if we look at the underlying physiology of cancer in young patients with these two oncological diagnoses. Patients with early cancer typically have inherited mutations which increase the likelihood of developing cancer. In breast and colorectal cancer this may include BRCA1 and BRCA2 mutations.[397] In colorectal cancer this may include additional genetic syndromes including familial adenomatous polyposis[398], attenuated familial adenomatous polyposis and Lynch syndrome.[399] The route to development of tumours as a result of these familial genetic risk factors results in cancers that behave in a fundamentally different way, responds differently to treatment, have different chances of recurrence and are often associated with risks of other cancers outside of that primary site. As a

result the colorectal cancer and breast cancer of the younger patients is fundamentally different from those who develop cancer at an older age. Thus outcome differences are unsurprising.

Previous research has highlighted the importance of age which has been shown to be associated with differential outcomes. In the context of breast cancer the non-linear effects have been demonstrated by placing patients into age bands of less than 40, 40-49 and 50+.[395] This showed that the best survival was seen in the 40-49 year olds followed by the under 40s and the worst outcomes in the patients aged 50 plus. Research using European data has suggested the survival disparity between middle aged and older aged patients is growing across seven cancers.[400] When looking just at patients offered curative surgery, which in effect focusses on the earlier diagnoses and fitter patients, age is associated with adverse outcomes.[401] Although these previous studies have highlighted age as an issue, the methods employed have been limited in how they distinguish between ages and use a traditional inferential approach. The results presented above provide greater granularity, do not rely on linear assumptions[178] between age cut offs or proportional hazards assumptions[402] and are thus more likely to demonstrate more complex patterns with respect to age where they exist.

Stage

The results for the partial plots showing the relationship between stage and survival show the pattern that would be expected based on the previous literature, which is that with increasing stage there is a corresponding reduction in survival probability.[403–405] This is unsurprising given how cancer stage is assigned and that the staging system is developed to provide outcome discrimination. With increasing stage there in increasing levels of local and distant spread, the more anatomically dispersed a cancer is the lower the likelihood of curative intervention being possible. As such, those patients with high stage disease are in most instances offered only palliative treatment. Although the use of radical intervention with surgery and radiotherapy with curative intent does not guarantee becoming disease free, the use of palliative treatment will almost never result in patients becoming disease free, thus physiologically accounting for disparities.

These results further highlight that the differences between stages are inconsistent. These inconsistences occur both when comparing different cancers and when comparing the same cancer at one or multiple time points. This suggests a violation of both linear assumptions and proportional hazard assumptions. This would result in incorrect estimates of the relationship between stage and survival using the traditional Cox approach. It highlights that when using staging information to inform clinical and patient decision making, the conversation may shift depending on when the discussion is taking place. How stage at diagnosis influences outcomes after a period of survival will be different from at the time of diagnosis. This highlights a clear area that is in need for further research which is survival predictions that are dynamic rather than static and produce estimates based on having already survived after a cancer diagnosis for a given period of time.

As demonstrated in chapter 4, the overall survival differs substantially between cancers. As such the wide variations seen when comparing across fixed time points is unsurprising. Further work could be undertaken to use time points for each cancer that represent a percentage of median survival. It could be possible that the effect of stage is more consistent if for example it was assessed not at 1, 5 and 10 years but instead at 50%, 100% and 150% of median overall survival for that cancer.

A further consideration when assessing this data is the high levels of missing staging data highlighted in chapter 3. This has resulted in many of the patients being excluded from these summary statistics as their stage was not known. If there are certain types of patients whose staging data is consistently

missed, this may result in biased results for those who have complete data. As such further work is needed to enhance the missing data and repeat the analysis to ensure that the results found are consistent when more complete data is available.

Grade

Tumour grade refers to the morphological characteristics of the cancer cell and how far they diverge from the original cell structure[406]. The increasing phenotypic changes seen with increasing grade reflect increasing numbers of genetic abnormalities within that cell. This may result in increasingly abnormal cell behaviour which is why high grade disease is often associated with an increased risk of metastatic spread, treatment resistance and more rapid progression[22,394]. Given these physiological differences it is unsurprising that the results of the above analysis show that on average with increasing grade there is an increase in risk of mortality. The effects are however different across the cancers studied and vary with respect to time from diagnosis. The use of flexible non-parametric modelling in the form of RSFs makes the identification of these complex relationship easier, especially given the absence of underlying model assumptions.

The outcome effects associated with grade appear less marked than those seen with stage data. This is likely to be due to the fact that grade encodes information about the risk of metastatic disease and disease severity whereas stage directly reflects the actual disease severity and presence or absence of metastatic disease.

Further Analysis and Work

Although not directly covered within the results above, the models generated could be used to assess the relationship between other variables and outcome predictions. Further work is therefore needed to assess the relationship with IMD which has been shown to be an influential predictor. The VIMP results suggest that in many instances there are interactions between variables and overlap of encoded information. Stratified partial dependence plots could therefore be used to identify differences in the patterns of predicted outcomes within substrata. An example of this would be the effect of grade in patients with and without hypertension in prostate cancer which had the largest suggested interaction of the results found in **Table 20**.

Previous research has called into question the validity of results obtained using both partial plots and VIMP in the context of correlated variables.[190,391,392] These previous studies have highlighted issues such as incorrect VIMP estimation and identification of stronger associations between features using partial dependence plots than are truly present within the data. Within the PPM dataset some correlation has been identified and thus may be a factor in the results seen. Future work could therefore use alternative approaches for example those based on permutation and relearning.[391] Here the value is permuted and the model retrained using this new permuted data. The resulting model accuracy can then be compared which combats the issues highlighted above. The significant trade-off is the vastly increased number of models that are needed to be trained and thus the amount of compute resource that these would require. Further these issues are far more important if trying to identify a biological mechanism driving a process.[392] As the approach employed above is solely trying to identify at risk individuals the bias is inconsequential so long as the outcome prediction is accurate. As a result of these issues, this approach was not employed despite its theoretical advantages.

7.3.6 - Associations between Comorbidity and Predicted Survival

When attempting to interpret the results for the stratified survival curves it is notable that the prostate cancer curves are jagged and therefore difficult to interpret. This is due to a limitation of the model building process. During the model building process it is possible to specify the parameter ntime as part of the hyperparameters. This is a count of the number of time points that survival differences are estimated during the model building process which is subsequently used to derive the stratified survival curves. Correspondence with the authors of the original random survival forest publication[109] (see appendix) has highlighted that reducing the ntime time reduces the memory requirements for model training and improves the model training time without significantly impacting on accuracy. By default the "ntime" is set to the number of events within the dataset. During our model building it was set to 100, which was done due to the limited computational resources available. It was found that increasing "ntime" above 100 results in insufficient memory being available to train the models, thus resulting in failure to generate the models. This in effect means that the survival curves shown are generated from 100 equally spaced estimates completed between time zero and the last recorded death. In the case of prostate cancer where the last recorded death is a longer time after diagnosis that the other patients this results in fewer estimates within the 10 year analysis period and thus a more jagged and less precise line being output. Future work building upon this should ideally take advantage of improved compute to allow ntime to equal the number of events yielding more precise stratified survival curves.

The results of the stratified survival curves largely mirror the results found in chapter 4 and chapter 5 with comorbidity being associated with worse outcomes across the four most common cancers. A couple of notable exceptions to this are once again highlighted including COPD in lung cancer and obesity in lung cancer. The results assessing varicosities in colorectal cancer do not however agree with those derived from the Cox models in chapter 5 with no difference clearly demonstrated with the RSF model approach. The large scale survival differences noted are important to contrast with the relatively low importance scores seen for most within the VIMP analysis. This suggests that comorbidities are likely to be closely related to several other key features such as age, grade and stage. Thus it may be true that these other features act as potential confounders, mediators or can be used to encode much of the same information as is provided by comorbidity status. This may in part be due to the binary nature of the comorbidity information that has been used as the basis of the analysis undertaken. As shown in the time of diagnosis analysis undertaken in chapter 3 using more granular information can provide additional insight. Future work should therefore look to move beyond a binary indicator of comorbidity and incorporate alternative information such as time from comorbidity diagnosis to cancer diagnosis. This would allow further analysis as to whether comorbidity diagnosed at different time points provides different information either due to its long standing nature or due to proximity to cancer suggesting a possible direct connection.[407–410] Additional measures of disease severity could be introduced such as those based on blood results. Previous research in the area of diabetes has demonstrated that HbA1c is important in assessing severity but can be assessed based on HbA1c variance in an individual which suggest labile blood sugar control and therefore poor control.[411–413] Other measures such as this could improve predictive performance, as well as greater insight into the behaviour of comorbidity subgroups. The more granular the data the greater need for increased number of comorbid patients and thus larger samples might be needed to generate enough precision of estimates.

7.4 - Summary

Our analysis demonstrates that when comparing traditional methods in the form of Cox and Kaplan Meier models to random survival forests, the machine learning approach has improved predictive accuracy. The models provide valuable information about which features are most informative or influential in generating those predictions. The features can therefore be used to identify individuals who may have either higher or lower risk of adverse outcome and could be the target of further study and research. The use of partial dependence plots provides insight into the complex relationships that individual features have in relation to outcomes. The complex nature of these relationships are such that a standard Cox model is unable to provide accurate estimations of the true pattern without multiple complex adjustments.

Despite these benefits the model development process required more steps, was more time consuming and required much more computational resources to be delivered. The large impact of increasing case numbers on model development time provides a potential barrier to the use of this method when using national datasets unless significant compute resources are available. More work is needed to explore how these models can be used for individualised predictions with accompanying explanations and to assess this approach prospectively. Overall however this modelling approach has yielded results that would be challenging to capture using traditional approaches and should be considered for use more widely within other time to event analyses in healthcare.

The results with respect to the use of comorbidity as predictors of outcome is somewhat mixed. Patients with comorbidity on average are shown to have similar outcome differences to those highlighted in chapters 4-6. Despite this, in many instances comorbidities provide little additional predictive benefit when used in combination with other features such as grade, age and stage. This suggests that although comorbid patients have worse outcomes there are more efficient predictors of outcome that can be used instead.

The final chapter that follows attempts to bring together the findings of all the analyses undertaken across chapters 3-7. It will address our original research questions and provide further details on what the results mean for both existing research and future research in this area.

Chapter 8 – Conclusions and Further Work

8.0 - Introduction

Through the investigation of the seven key aims of the research presented within this thesis a number of key questions are brought to the fore. Is hospital data accurate reliable and representative? What is the impact of comorbidity on cancer outcomes? What information can be used to predict cancer outcomes for patients? Does machine learning provide better survival predication for cancer patients? Although these may appear superficially to be basic and simple questions, the results presented prove that the attempts to answer them are anything but simple and are hugely dependent on the situation and interwoven with a need for nuance and caveats. Each of these key questions and the aims to which they refer will be discussed below within the context of the results presented. This concluding chapter will build on the individual chapter discussions and summarise the overall results within the context of potential future work whilst also giving focus to the limitations of the analyses undertaken.

8.1 - Accuracy, Reliability and Representativeness of Hospital Data

If we first take the issue of hospital data accuracy, then it is clear from the results that there are a number of flaws. In some cases the information is missing in large volumes, diabetes data has been shown to be inaccurate in a large percentage of the population and the cohort found within the hospital record was not representative of the wider region and UK. It could be argued that basing findings on missing, inaccurate and unrepresentative data is of little use.

The reality is that the relative utility is not black and white, and is very much dependant on the way in which one hopes to use the information. It would be inappropriate to extrapolate the findings of these analyses from one population, to another which is known to be fundamentally different. An example of this is knowing that hospital clinical coding appears to capture the diabetic patients with worse outcomes, results based on clinical coding alone should only be used when looking at other patients with hospital clinical coding for diabetes. This provides insight for this more limited group and although not as useful as it would be if the information was relevant to all diabetic patients, it still represents a step forwards, over and above having no information at all.

The results of the research presented cover large numbers of cancer sites and comorbidities, many of which have never been studied in combination before. Even where studied previously, these have been based on inconsistent approaches and definitions.[240,414,415] Despite the underlying data limitations, the results provide valuable new information that was not previously known. It is important to ensure that as with the diabetes example the interpretation of results is appropriately constrained. It should be based on the identification of patients using the same method and in a population that is similar. For the information to be used in a less constrained manner further research would be needed to improve the quality of data or use a more representative population and further information on this can be found in section 8.5 below.

The identification of limitations in hospital data also has potentially profound and far reaching effects on the interpretation of previous comorbidity research, not only in oncology, but other medical fields as well. The heavy reliance on clinical coding to identify comorbidity in hospital datasets is shown to be problematic. This approach results in substantial misclassification error and incorrectly assigns the timing of first diagnosis often by a period of several years in the context of diabetes mellitus. Although not directly studied, it is plausible to assume that similar issues may affect clinical coding in other conditions. Beyond the diagnostic information simply being incorrect,

the results of our analyses demonstrate that this has a clinically meaningful impact of the analysis results.

In the majority of previous comorbidity research based on hospital clinical coding the results are presented as being representative for all patients with that diagnosis without drawing a distinction between the coded patient population and the true population with that condition.[416–418] In the case of diabetes mellitus it is reasonable to say that all research using only clinical coding should be extrapolated beyond the study population in a highly constrained way to avoid incorrectly informing policy and clinical decision making that could disadvantage patients. It is likely that same is also true for other conditions and thus the output of previous research into comorbidity using clinical coding should be used with a much higher degree of caution and constraint. Further research is needed to assess this issue across other common conditions and quantify the extent of the problem both with the UK and internationally.

The implications of this misclassification error[218] also cross over into the use of comorbidity scoring systems. It is common practice to develop a risk score using coding data and then validate it using prospective coding data or a holdout dataset.[38,39,41] If the score is then implemented in clinical practice such that clinicians are getting comorbidity information directly patients, then the information is not being used with the appropriate constraints that should be applied.

Concerningly, the results of the data accuracy analysis identified that diabetes mellitus was associated with worse outcomes using any of the three data definitions, however the scale of effect differed substantially. This could result in a situation where if a score system based on diabetes mellitus clinical coding was implemented using broader data, such as self-reporting, and was subsequently assessed, it would still potentially appear to be useful, as it would correctly identify worse outcomes for diabetic patients. The issue however is that the scale of risk quoted to large numbers of patients would be incorrect and potentially negatively influence their management inappropriately. As a result, comorbidity scoring systems currently in regular use within the clinical setting need to be reviewed to identify any potential examples that may be subject to this pitfall across all areas of medical practice. Examples of scores being used in this way can be see frequently within the literature.[419–421]

8.2 - The Impact of Comorbidity on Cancer Patient Survival

When attempting to address the question of what is the impact of comorbidity on survival outcomes in cancer, our analyses have looked at this in three broad ways, descriptive, inferential and predictive. Across all the approaches implemented a common theme arises, the associations seen vary for the same comorbidity across cancers and for the same cancer with different comorbidities. The variation of associations seen is substantial, to the extent that in some instances comorbidities are associated with improved survival outcomes and in other cases worsened survival outcomes.

The variation of effects identifies a further potential issue with previous research. Many comorbidity studies in oncology may focus on multimorbidity rather than individual comorbidities to conduct their analysis. This is commonly achieved through the use of a comorbidity score such as the Charlson score.[38] The different outcomes identified with different comorbidities suggests that a composite comorbidity score is likely to be inappropriate, as the score will draw false equivalence of effect between different comorbidities. This may result in analysis outcomes which are an average effect estimate that will be influenced not only by the effect sizes for each condition, but also how well represented each condition is, relative to one another, within the population. The results of

these previous studies using this approach should therefore be viewed with caution and a degree of scepticism.

This concept can be taken a step further, as the approach implemented within this thesis although assessing comorbidity individually, treats all individuals with that condition as the same. The clinical reality is that within the population of patients with a given condition there is wide variation in the severity of disease and degree of impact on each individuals' activities of daily living. Within the population of patients with a given condition there is likely to be variation in its impact on outcomes and the use of more granular data might enable this to be better understood. Further discussion on this can be found in section 8.5.

The interpretation of many of the results presented is somewhat challenging due to a range of potential sources of error and bias.[163,218,275,296,371] In addition to the issues of comorbidity misclassification error discussed above, our results identified potential selection bias, violations of model assumptions, and through the application of causal methods, highlighted collider bias that could introduce further error into the estimates obtained. When combined with potential organisational forms of error such as lead time bias, different patterns of and timing of referrals, identifying what is a clinically relevant association is made more difficult. The use of observational data preventing the use of a statistical test precludes the application of a single simple test to identify which results are "significant".[59] Instead the analysis becomes reliant on confidence intervals to estimate likely direction of effect and precision of effect estimates. The lack of a commonly used universal threshold for each of these measures gives the impression of the results being far more arbitrary in the selection of a cut off. The reality is however that a 95% confidence threshold for a p value is just as arbitrary, however is more established.

The interpretation of results is further complicated when using competing risk analyses as is the case in Chapter 6. To interpret the cancer cause-specific results they must be assessed in comparison to the all-cause results.[179–181] This creates a scenario where the direction of effect and precision interpretation is applied to both, such that in many cases so many thresholds are applied that the number of "positive" results are small. This may be entirely appropriate, however may also be a reflection of dismissing meaningful results through the application of so many cut offs. When combined with known issues of cause of death reporting errors[185,371], providing robust estimates on which one can draw meaningful conclusions is difficult.

When taking into account the various sources of error and bias from the data and the methods applied it is reasonable to wonder whether there is actual value in the results obtained. It could be argued that as the sources of error cannot be removed or quantified, it is impossible to know if the results seen are inflated, deflated or even reversed. If one is trying to attribute the effects seen to a cause then this concern is entirely valid. If however one is simply interested in knowing if patients with a given condition do worse than those without, then there is still value to be obtained. It could therefore be argued that in this scenario the descriptive analysis undertaken in chapter 4 is more appropriate than the results seen in chapters 5 and 6, as the multivariable approaches, although not trying to assign causality, does attempt to attribute an effect to particular variables.

The use of simple description is however of limited clinical benefit. Knowing that patients with a given condition have different outcomes is useful in prognostication, however does not provide any insight into how the delivery of care could be altered to improve outcomes or even if that is possible. In many cases the associated differences in outcomes may be totally or partially explained by other related variables hence the need for adjustment. This may also result in misleading information being used in clinical decision making. An example of this would be that if a given condition is

associated with older aged patients and it is their age that results in worsened survival, a young patient with that condition may be falsely thought of as having a worse prognosis even though this is not the case. This can be overcome to an extent with a multivariable approach which brings the argument full circle and back to the issues around methods and error highlighted above.

An alternative is the predictive approach which allows the analysis to move beyond simple description and can be used to identify higher risk populations based on one or more factors. The results of the predictive analysis identified that although comorbidity did contain information in many cases that was useful for prediction, other patient level characteristics were in general much more useful in the identification of a patient's likely outcome. This approach enables the investigation of complex non-linear patterns of association. The predictive approach is however not particularly useful in explanatory terms. If we identify that a given characteristic has a pattern of association with outcomes, the natural question that will follow is why? Predictive methods simply cannot answer this. To answer this sort of question a causal analysis in needed. As highlighted in chapter 5 the issues around collider bias and temporal ordering of conditions makes undertaking robust causal analysis both challenging and time consuming, and would be impractical to conduct across the cancers and comorbidities assessed within the analyses presented within this thesis.

Thus the reality is that despite the limitations of descriptive, inferential and predictive analysis they are a necessary part of answering these why questions. They perform the important role of identifying clinically meaningful outcome differences that are worthy of further investigation. They can be used to identify the groups or targets of further study in which well thought out and meticulously designed causal analyses may be undertaken. If a well-designed causal study identifies a potential explanation, then this can be used to inform an interventional study which could be in the form of an RCT. As discussed in chapter 1 RCTs cannot assign comorbidity but RCD based observational studies, in the way described, can provide an evidence base to enable the study of improving outcomes related to comorbidity, through carefully identified interventions. In this situation it would therefore seem that the argument of what is "better" descriptive, inferential predictive or causal analysis, observational or randomised studies are not in fact helpful. Instead a greater focus is needed on how these different analyses can be used together in a complimentary manner to affect improvements in knowledge and patient care in the most robust, time efficient and cost efficient manner.

8.3 – Information for Predicting Cancer Outcomes

The analyses undertaken in chapter 7 provide information on the relative contribution of a number of baseline characteristics in predicting cancer outcomes. Although the relative value of each characteristic differs to some extent based on the interpretability measure used and the model assessed, it is clear the age, stage, grade, morphology and gender are in general the most significant contributors to predictions. This is in many ways reassuring as historically these have been the features consistently relied upon by clinical professionals to determine likely outcomes for cancer patients. Other characteristics such as the presence or absence of comorbidities was more variable in its contribution to predictions and although in many cases they had some predictive capability, it was modest.

It is important to consider that the predictive contributions quantified in each case is within the context of having the information of the other characteristics to rely on. As such comorbidity if used alone may provide valuable information, however if this is true, then its value is diminished when other more useful information is available. This is extremely important when interpreting the clinical relevance of the results presented within chapters 4-7 which show that when looking at comorbid

populations, those with comorbidity commonly have different outcomes in cancer patients. When however faced with an individual patient with comorbidity, if the clinician is attempting provide a survival prediction, then if key baseline characteristics such as the patient's age, gender, stage and grade have already been accounted for then the presence or absence of comorbidity is unlikely to further inform the survival prediction given. This is in some ways counterintuitive however highlights the need for further study into the causal relationships and effect overlaps between these characteristics.

8.4 - Traditional Statistical Methods versus Machine learning Methods

When looking at the results of the predictive analyses the more traditional methods of Cox and KM were show to have inferior predictive capabilities when compared to the random survival forest approach. The ability of random forests to identify non-linear patterns within the data also suggests potential interpretability advantages over the more traditional methods which are constrained by linear assumptions and proportional hazards assumptions. Despite this, it could be argued that the random forest approach is inferior in a number of other ways. Firstly the process of developing the optimised survival models was both time and computer resource intensive. This grows with increasing patient numbers such that for many institutions using this approach to analyse their data would be impractical. The interpretation of which variables are the most important predictors is also challenging due to known issues with the bias seen with VIMP scores and partial plots.[391] Although this issue can be overcome with alternative interpretation methods, these alternatives make the development process require even more resource.

To understand the true value of Cox in prediction further iterations of variable specification would be needed to include interaction terms and different combinations of these with other variables. It could in fact be the case the correct specification of a Cox model could outperform the RSF approach.

In either case however it would appear to be that with increasing model accuracy there is an increasing cost in terms of resource both time and computational. The degree of accuracy needed is going to very much depend on how one intends on using the model. If the aim is to identify broad population level patterns to inform future study it might be acceptable to have a lower accuracy requirement than if the intention was to provide prospective survival predictions to individuals.

8.5 – Study Limitations

Within the preceding chapters and the above sections a number of potential pitfalls and limitations have been highlighted. The following section outlines some of these areas with greater detail and the subsequent section 8.6 focuses on the potential next steps in overcoming these issues and building upon the work presented within this thesis

1) Unrepresentative Populations

The results presented in chapter 3 clearly demonstrate a number of ways in which both the PPM population and more specifically the PPM cancer population differ from the wider population. These differences in the basic demographics of the cohorts and levels of the population with comorbidity may impact on the external validity of the results. Thus the reliance on a single UK centre is a limitation to the potential applicability of the results both nationally and internationally. Even if the results do generalise their being unrepresentative may result in poor external validity of results.

2) Accuracy of Clinical Coding

Our result have demonstrated that within the PPM dataset clinical coding when used in isolation is not a reliable method for the identification of all patients with known diabetes mellitus. The issues with clinical coding may also extend beyond DM and affect other conditions. Data enhancement was only undertaken for DM data and obesity data but not the other comorbidities of interest.

3) Missing Data

The lack of complete data for grade and stage resulted in its exclusion from the analyses undertaken using traditional statistical techniques. They were however shown in chapter 7 to provide important predictive information. The lack of their inclusion due to missingness in chapter 4-6 may have resulted in inaccurate results if these factors are important within the context of inferential analyses.

4) Reliance on Structured Data

The analyses undertaken within this thesis rely solely on structured data. It is estimated that the majority (circa 80%) of clinical information is however recorded within the unstructured data of an EPR.[64] The lack of access to this information may introduce misclassification error.

5) Treating Comorbidity as Binary

The method of describing comorbidity was to define the presence or absence of each condition at the point of cancer diagnosis. This is however a simplification of the complex reality of comorbidity, where disease severity and length of time with the condition may also be of importance. These aspects where not considered.

6) Overall Survival as the Only End Point

The analyses rely on overall survival or cause-specific survival as the only end point for analysis. This fails to consider disease progression, treatment free time and quality of life. These alternative end points are in many cases as, if not more important to many patients and have not been assessed.

7) Limited Cox Optimisation

When comparing the performance of Cox to RSFs efforts were made to optimise the RSFs however the same level of optimisation was not applied to the Cox approach. It could be argued that this may have resulted in a potential performance advantage of RSFs over Cox when comparing their relative accuracy.

8.6 - Further Work

1) Inclusion of More Data

In order to overcome the highlighted issues of how representative the PPM population is, data could be gathered from other settings. This could include the addition of other cancer centres in the UK and internationally, but could also include the addition of linked primary care data. The use of national registry data such as national HES data could provide further data making the dataset more representaive.[126]Although this approach would be subject to additional legal and ethical considerations, adding data would serve to improve how representative the study population is and may also help to overcome some of the issues of missing and inaccurate data as well. For this to take place the study would need to be undertaken within an NHS data safe haven.[67] As a result, this research may be best suited to the analysis being undertaken within central organisations such as NHS digital.

2) Improving the Accuracy of Clinical Coding

As in the previous section the analysis could be completed on national HES data.[126] This would require some form of data linkage to occur as routine blood results do not currently form part of any of the nationally collected datasets. This approach would provide a much more accurate understanding of how widespread the limitations of clinical coding for DM are and whether significant local variation is identified. A further similar investigation could also be undertaken using international data to identify whether similar issues occur when using administrative data from other geographical locations.

DM is an easy target for analysis due to the easy to identify diagnostic test of HbA1c. Many other conditions are not as simple to identify from a single test, these include chronic kidney disease, chronic liver disease and autoimmune conditions. Involvement with domain experts could be utilised to create rule based approaches to blood based definitions being incorporated. This could allow assessment of whether other conditions require enhanced data definitions in the same way as DM within our analysis. The accuracy could be further improved by linking together primary and secondary care data. This has been shown, particularly during the COVID-19 crisis to yield additional insight that cannot be provided by secondary care data alone.[422]

3) Accessing Unstructured Information

Natural language processing methods could be applied to clinic letters to extract further information.[423] This could be used not only for the purposes of comorbidity information but could also be used to identify missing information such as the stage and grade data that was missing in many of the diagnoses recorded within the PPM dataset.

4) More Granular Comorbidity Data

Further analysis could be undertaken to include comorbidity data in a more granular way. One example could be to represent comorbidity as a time before cancer diagnosis. This would allow the relationship between when the patients develops comorbidity to be assessed for its association with survival outcomes. Comorbidity severity could also be included in some way. In the case of diabetes, HbA1c blood results could be used in the form of summary statistics which might provide further information to help inform analyses

5) Alternative End Points.

Assessing the relationship between comorbidity and patient reported outcomes could further our understanding of the patient experience with and without comorbidity. This sort of quality of life study would be unlikely to be possible using retrospective data as patient reported outcomes data is not currently routinely collected for patients. This information could however be collected via survey data and subsequently analysed.

The lack of the availability of this data routinely for patients does highlight a large gap in the available information for patients in the UK. With quality of life issues being of such huge importance, the absence of this information creates an information bias when clinicians and patients are making decisions. Either a stronger focus is placed on survival, where there is more extensive evidence and information, or there is a greater reliance of professional opinion and weak trial data when discussing quality of life. This is therefore an area that needs further research to address this current limitation on informed decision making.

Additionally, the relationships between recurrence, progression and toxicity could be studied as alternative end points, however as with quality of life, this information the data is not readily available and may benefit from approaches such as natural language processing as discussed above.

6) Late Effects

The analysis undertaken focussing on the relationship between cancer and the later development of serious health conditions was only analysed in a descriptive manner within this thesis. As detailed in chapter 2 little research has been undertaken to fully understand how cancer and its treatment impact on the long term health of patients living beyond their oncological diagnosis. This further analysis could be done in a number of ways. The impact of the later development of chronic health problems could be assessed to identify whether this is associated with altered survival outcomes in the same way that up front health conditions have been shown to. This would need to overcome the issue of immortal time bias through the use of time dependent variables.

The use of additional control cohorts could also provide information as to whether cancer has a greater association with later health problems than the background population. Using a matched control cohort the prevalence of comorbidity in a fixed period after index date could be compared across the cancer and control groups. An alternative approach would be to analyse the development of comorbidity as a time-to-event analysis. This could identify not only differences in the cumulative incidence but also time variations such that if cancer and its treatment is associated with speeding up a pathological process, this could be identified even in the absence of changes in the prevalence.

7) Further Comparison of Cox to RSFs

Our results have demonstrated the superiority of random survival forests over Cox models constructed in several ways. Further investigation is needed to determine whether there are circumstances in which Cox models could outperform RSF. This could be done by comparing forwards elimination, backwards elimination, VIMP informed variable selection and all possible combinations of variables achieved by brute force. Additionally, the inclusion of interaction terms might alter the performance of a Cox model. It could be the case that Cox models could outperform RSFs however it would require the correct model specification to achieve.

8) Enhancing RSF Analysis

Although our analysis suggested that RSFs outperform Cox models a number of other pitfalls relating to variable selections and interpretability have been highlighted in the literature. Further analysis could therefore be undertaken using alternative approaches to model interpretability using for example permute and retrain or totally different metrics altogether. A comparison to conditional inference forests would also allow for assessment as to whether the ordinal nature of some of the data is influencing the accuracy of the prediction or the interpretation of the models.

9) Using RSFs for Unsupervised Learning

The random survival forest methods employed can be used to generate proximity and distance measures. The proximity measures are based on the proportion of the time that two patients are in the same terminal leaf within the forest. The distance measure is a measure of the percentage of the total possible maximum distance between two patients along the edges of the graph representing the tree. This data can be used as a distance matrix and subjected to unsupervised learning for clustering using either K-means or Hierarchical clustering.[424] Clustering could be used to identify whether groups with different patterns of survival outcomes can be identified. The defining characteristics of these groups could then be used to identify combination clinical phenotypes that

may indicated different survival outcomes. This would in effect use the random survival forests as a dimensionality reduction method for multidimensional clustering to create an efficient approach to an otherwise computationally complex process. This could be particularly useful in identifying particularly high or low risk groups for further study but that incorporate information about as wide range of patients level characteristics simultaneously.

10) Other forms of Machine learning

Within the results presented only a single machine learning approach has been implemented. As discussed above there are a number of forest based methods that offer slight variations on the approach used which could be employed. There are however also a number of other approaches that could be undertaken using completely different approaches. Some examples of survival analysis optimised machine learning methods have been developed and described within the literature including support vector machines[425] and deep neural networks[389]. Further analysis could be undertaken by implementing these alternative approaches to identify whether they can be used in this clinical context to improve the predictive performance of the models.

11) Multimorbidity

Although some of the methods suggested above would provide insight into multimorbidity a more focussed approach could also be undertaken, particularly in the area of prediction. If a score system was to be used then this could be done using the comorbidity data and optimising the score towards the specific survival task. Given the results we have presented this would likely need to be done on a cancer site by cancer site basis.

12) Causal Inference Analysis

One area of analysis that has been intentionally avoided is that of causal inference. From our conditions that have been shown to have an association with differential outcomes, formal causal analysis could be undertaken. Here precise expert informed DAGs could be developed to inform the analysis and design of a study. It is likely that the precise estimand would need to be well defined and might need to differ from the idea estimand, in order to produce a reliable estimate.

13) Exploring Missing Data

Further work could be undertaken to assess whether the data that is missing within the dataset was associated with other parameters such as comorbidity, age etc. This along with data visualisation could enable a better understanding of the missingness of the data and offer the potential of the use of imputation methods where further data cannot be added or obtained.

14) Covid-19

The data from this study is from 2018 and before, and as such Covid-19 is irrelevant to the analyses undertaken. Covid-19 has however resulted in substantial changes to how medicine is being practiced and is also impacting on outcomes from a range of diseases with cancer being just one example. The methods employed to assess comorbidity within this thesis could equally be applied to assess comorbidity in Covid-19 patients or treat Covid-19 as a comorbidity in cancer patients.

8.7 - Summary

Despite the limitations stated and potential areas for further study, overall the analyses presented enable us to draw a number of conclusions.

1. Clinical Coding alone does not robustly record diabetes mellitus within the PPM cancer population.
2. The diabetes mellitus misclassification error occurring as a result of clinical coding affects the results of survival analysis in cancer patients.
3. The PPM dataset is not representative of the wider cancer population.
4. Patients with comorbidities commonly have different outcomes than those without comorbidity although this is not universally true.
5. The effect associated with comorbidity is dependent on the cancer that the patient has.
6. It is unclear as to whether the outcome differences seen are attributable to that comorbidity.
7. The associated effects of comorbidity are not always negative.
8. Binary indicators of individual comorbidity perform poorly as predictors of outcomes when compared to other patient level characteristics.
9. Using composite scores that combine comorbidities risk providing inaccurate estimates of effects associated with comorbidity.
10. Random survival forests may outperform Cox proportional hazards for cancer survival prediction.

These conclusions add to the growing body of literature in the area of comorbidity research and should serve to inform and enhance studies of this nature in future, both in oncology and beyond.

Appendix

Site	Histology Group
Bladder	Adenocarcinoma, Leiomyosarcoma, Other, Papillary, Squamous Cell, Transitional Cell, Unspecified
Brain	Astrocytoma, Glioblastoma, Glioma, Medullobastoma, Oligodendroglioma, Other, Unspecified
Breast	Ductal, Lobular, Medullary, Mixed, Mucinous, Other, Papillary, Tubular, Unspecified
Cervical	Adenocarcinoma, Mixed, Other, Small Cell, Squamous, Unspecified
Colorectal	Adenocarcinoma, Mucinous Adenocarcinoma, Neuroendocrine, Other, Squamous Cell, Unspecified
Connective	Adenocarcinoma, Other, Sarcoma, Unspecified
CUP	Adenocarcinoma, Other, Squamous, Unspecified
Endometrial	Adenocarcinoma, Clear Cell, Mixed, Other, Papillary Serous, Unspecified
Kidney	Adenocarcinoma, Other, Transitional Cell, Unspecified
Laryngeal	Other. Squamous, Unspecified
Leukaemia	Acute Lymphoblastic Leukaemia, Acute Myeloid Leukaemia, Acute Promyelocytic Leukaemia, Chronic Lymphoblastic Leukaemia, Chronic Myeloid Leukaemia, Myelodysplastic Syndrome, Other, Unspecified
Liver	Adenocarcinoma, Cholangiocarcinoma, Hepatocellular Carcinoma, Mixed, Other, Unspecified
Lung	Adenocarcinoma, Large Cell, Mixed, Neuroendocrine, Other, Small Cell, Squamous, Unspecified, Unspecified Non-Small Cell
Lymphoma	Hodgkin Lymphoma, Mixed, Non-Hodgkin Lymphoma
Melanoma	Acral lentiginous, Lentigo Maligna, Nodular, Other, Superficial Spreading, Unspecified
Myeloma	Myeloma, Plasmacytoma
Oesophageal	Adenocarcinoma, Other, Squamous, Unspecified

Site	Histology Group
Ovarian	Adenocarcinoma, Clear Cell, Endometroid, Mixed, MMMT, Mucinous, Other, Serous, Unspecified
Pancreatic	Adenocarcinoma, Cholangiocarcinoma Duct Cell, Neuroendocrine, Other, Unspecified
Prostate	Adenocarcinoma, Other, Unspecified
Skin	Basal Cell Carcinoma, Other, Squamous Cell Carcinoma, Unspecified
Stomach	Adenocarcinoma, Other, Unspecified
Testicular	Non-germinal, Non-Seminoma, Other, Seminoma, Unspecified
Thyroid	Follicular, Medullary, Other, Oxyphilic, Papillary, Unspecified

Table 21: Histological Groupings Applied in Each Cancer Site

Comorbidity	Code Definitions
Acute MI	I21, I22, I23, I24.1
Angina/CAD	I20, I24.8, I24.9, I25, I70
Ankylosing Spondylitis	M45, M08.1
Arrhythmia	I44, I45, I47, I48, I49
Asthma	J45, J46
Cardiomyopathy	I42, I43
CCF	I50
COPD	J41, J42 , J43, J44, J47
Dementia	G30, G31, F03, F00, F01, F02, F03
Demyelination	G35, G37
Diabetes Mellitus (Other)	E12, E13, E14, O24.2, O24.3, G59.0, G63.2, H28.0, H36.8, I79.2, M14.2, N08.3
Diabetes Mellitus (Type 1)	E10, O24.0
Diabetes Mellitus (Type 2)	E11, O24.1
Gout	M10
HIV	B20, B21, B22, B23, B24, F02.4, R75, z21
Hyperlipidaemia	E78
Hypertension	I10, I11, I12, I13, I15
IBD	K50, K51
Liver Dysfunction	K70, K71, K72, K73, K74, K75.2, K75.4, K75.8, K75.9, K76.1, k76.3, K76.5, K76.6, K76.7, K77.8
Malabsorption	K91.2, K90
MND	G12.2
Neuromuscular	G10, G11, G12, G13, G23, G32, G70, G71, G72, G73, G93.1, G95.0
Obesity	E66
Other Rheumatological Disease	M09, M08, M30, M31, M32, M33, M34, M35, J99.0, J99.1, N16.4, N08.5, G05.8, G73.7, M07
PAD	I70.2, I73.9
Pancreatitis	K85, K86.0, K86.1
Paraplegia	G14, G80, G81, G82, G83
Parkinsons	G20, G21, G22
Peptic Ulcer Disease	K22.1, K25, K26, K27, K28
Psoriatic Arthritis	M07.0, L40.5, M09.0, M07.1, M07.2, M07.3
Renal Dysfunction	N01, N03, N05, N07, N08, N11, N14, N16, N18
Respiratory (Other)	J95.3, J96.1, J99.0, J99.1
Restrictive Lung Disease	J60, J61, J62, j63, J64, j65, J66, J67, J68.4, J70.1, j70.3, J84
Rheumatoid Arthritis	M05, M06
Spinal Cord Injury	T09.3, S34.1, S14.1, S24.1, S34.3
Stroke	G46, I60, I61, I63, I64, I69
TIA	G45
Varicose Veins	I83
Venous Insufficiency	I87.2
Venous Thromboembolic Disease	I26, I81, I82

Table 22: Comorbidity Code Definitions *- ICD-10 Code definitions used for string matching in clinical coding data*

Package	Version	Package	Version	Package	Version	Package	Version	
abind	1.4-5	formatR	1.7	readr	1.4.0	TH.data	1.0-10	Data Wrangling
base	4.0.2	formattable	0.2.0.1	readxl	1.3.1	tibble	3.0.4	
bitops	1.0-6	glue	1.4.2	reshape	0.8.8	tidyr	1.1.2	
blob	1.2.1	haven	2.3.1	reshape2	1.4.4	tidyselect	1.1.0	
boot	1.3-25	hms	0.5.3	rio	0.5.16	tidyverse	1.3.0	
caTools	1.18.0	janitor	2.0.1	rlang	0.4.8	timereg	1.9.8	
cellranger	1.1.0	lubridate	1.7.9	rpart	4.1-15	units	0.6-7	
class	7.3-17	magrittr	1.5	rstatix	0.6.0	usethis	2.0.0	
data.table	1.13.2	Matrix	1.2-18	snakecase	0.11.0	xlsx	0.6.4.2	
dbplyr	2.0.0	mgsub	1.7.2	SparseM	1.78	xlsxjars	0.6.1	
desc	1.2.0	openxlsx	4.2.2	stats	4.0.2	XML	3.99-0.5	
dplyr	1.0.2	pillar	1.4.6	stats4	4.0.2	xml2	1.3.2	
DT	0.16	plyr	1.8.6	stringi	1.5.3	xopen	1.0.0	
forcats	0.5.0	pryr	0.1.4	stringr	1.4.0	yaml	2.2.1	
foreach	1.5.1	purrr	0.3.4	summarytools	0.9.8	zoo	1.8-8	
fastmap	1.0.1	maps	3.3.0	raster	3.4-5	sp	1.4-4	GIS
isoband	0.2.2	maptools	1.0-2	rgdal	1.5-18	spatial	7.3-12	
leaflet	2.0.3	OpenStreetMap	0.3.4	RgoogleMaps	1.4.5.3			
leaflet.providers	1.9.0	osmdata	0.1.4	sf	0.9-6			
broom	0.7.2	MatrixModels	0.4-1	randomForest	4.6-14			Modelling
car	3.0-10	minqa	1.2.4	randomForestSRC	2.9.3			
cluster	2.1.0	modelr	0.1.8	ranger	0.12.1			
clValid	0.6-9	My.stepwise	0.1.0	riskRegression	0.02.05			
cmprsk	2.2-10	nlme	3.1-148	risksetROC	1.0.4			
conquer	1.0.2	nnet	7.3-14	rms	6.0-1			
e1071	1.7-4	pbkrtest	0.4-8.6	sm	2.2-5.6			
exactRankTests	0.8-31	pec	2019.11.03	splines	4.0.2			
generics	0.0.2	polspline	1.1.19	SQUAREM	0.5			
KernSmooth	2.23-17	polynom	1.4-0	statmod	1.4.35			

km.ci	0.5-2	prodlim	9.11.13	survival	3.1-12			
KMsurv	0.1-5	pscl	1.5.5	survivalROC	1.0.3			
lava	1.6.8	psych	2.0.9	survMisc	0.5.5			
lme4	1.1-25	qap	0.1-1	tmvnsim	1.0-2			
lmtest	0.9-38	quantreg	5.74					
crosstalk	1.1.0.1	shiny	1.5.0	shinyWidgets	0.5.4			**Web Tools**
httpuv	1.5.4	shinyBS	0.61	webshot	0.5.2			
httr	1.4.2	shinycssloaders	1.0.0					
rvest	0.3.6	shinydashboard	0.7.1					
brew	1.0-6	ggsci	2.9	latticeExtra	0.6-29			**Plotting and Table Generation**
cli	2.1.0	ggsignif	0.6.0	pander	0.6.3			
colorspace	1.4-1	gplots	3.1.1	plotly	4.9.2.1			
corrgram	1.13	graphics	4.0.2	plotrix	3.7-8			
corrplot	0.84	grid	4.0.2	png	0.1-7			
cowplot	1.1.0	gridBase	0.4-7	prettydoc	0.4.0			
dendextend	1.14.0	gridExtra	2.3	prettyunits	1.1.1			
dichromat	2.0-0	gtable	0.3.0	RColorBrewer	1.1-2			
ellipsis	0.3.1	gtools	3.8.2	rmarkdown	2.5			
gclus	1.3.2	hexbin	1.28.1	rpart.plot	3.0.9			
GGally	2.0.0	htmlTable	2.1.0	scales	1.1.1			
ggcorrplot	0.1.3	htmltools	0.5.0	sunburstR	2.1.5			
ggfortify	0.4.11	htmlwidgets	1.5.2	survminer	0.4.8			
ggmap	3.0.0	igraph	1.2.6	treemap	2.4-2			
ggplot2	3.3.2	jpeg	0.1-8.1	viridis	0.5.1			
ggpubr	0.4.0	kableExtra	1.3.1	viridisLite	0.3.0			
ggRandomForests	2.0.1	labeling	0.4.2	xtable	1.8-4			
ggrepel	0.8.2	lattice	0.20-41					
askpass	1.1	crayon	1.3.4	highr	0.8	multcomp	1.4-14	**Other**
assertthat	0.2.1	credentials	1.3.0	Hmisc	4.4-1	munsell	0.5.0	
audio	0.1-7	curl	4.3	ini	0.3.1	mvtnorm	1.1-1	
backports	1.1.10	d3r	0.9.1	iterators	1.0.13	nloptr	1.2.2.2	

base64enc	0.1-3	datasets	4.0.2	jsonlite	1.7.1	nortest	1.0-4
beepr	1.3	DBI	1.1.0	knitr	1.3	numDeriv	2016.8-1.1
BH	1.72.0-3	devtools	2.3.2	later	1.1.0.1	openssl	1.4.3
bit	4.0.4	digest	0.6.27	lazyeval	0.2.2	pacman	0.5.1
bit64	4.0.5	doParallel	1.0.16	lifecycle	0.2.0	parallel	4.0.2
callr	3.5.1	evaluate	0.14	magick	2.5.2	pkgbuild	1.1.0
carData	3.0-4	fansi	0.4.1	markdown	1.1	pkgconfig	2.0.3
checkmate	2.0.0	farver	2.0.3	MASS	7.3-51.6	pkgload	1.1.0
classInt	0.4-3	foreign	0.8-80	matrixStats	0.57.0	praise	1.0.0
clipr	0.7.1	Formula	1.2-4	maxstat	0.7-25	processx	3.4.4
codetools	0.2-16	fs	1.5.0	memoise	1.1.0	progress	1.2.2
commonmark	1.7	gert	1.0.2	methods	4.0.2	promises	1.1.1
compiler	4.0.2	gh	1.2.0	mgcv	1.8-31	proto	1.0.0
covr	3.5.1	gitcreds	0.1.1	mime	0.9	ps	1.4.0
cpp11	0.2.3	grDevices	4.0.2	mnormt	2.0.2	Publish	0.10.27
R6	2.4.1	roxygen2	7.1.1	whisker	0.4		
rappdirs	0.3.1	rprojroot	1.3-2	withr	2.3.0		
rapportools	1	rstudioapi	0.11	xfun	0.18		
rattle	5.4.0	rversions	2.0.2	zip	2.1.1		
rcmdcheck	1.3.3	sandwich	3.0-0				
Rcpp	1.0.5	selectr	0.4-2				
RcppArmadillo	0.10.1.0.0	seriation	1.2-9				
RcppEigen	0.3.3.7.0	sessioninfo	1.1.1				
RcppGSL	0.3.8	sourcetools	0.1.7				
RcppZiggurat	0.1.6	sys	3.4				
registry	0.5-1	tcltk	4.0.2				
rematch	1.0.1	testthat	2.3.2				
remotes	2.2.0	tinytex	0.27				
reprex	0.3.0	tools	4.0.2				

rex	1.2.0	translations	4.0.2	
Rfast	2.0.1	TSP	1.1-10	
rJava	0.9-13	utf8	1.1.4	
rjson	0.2.20	utils	4.0.2	
RODBC	1.3-17	vctrs	0.3.4	

Table 23: Complete List of R Packages and Versions Used Grouped by Package Utility

Cancer Site	Code Definition
Brain	C71
Breast	C50
Cervical	C53
Colorectal	C18, C19, C20
Connective Tissue	C49
CUP	C80
Endometrial	C54
Laryngeal	C32
Leukaemia	C91, C92, C93, C94, C95
Liver	C22
Lung	C34
Lymphoma	C81, C82, C83, C84, C85, C86
Melanoma	C43
Myeloma	C90
Oesophageal	C15
Ovarian	C56
Pancreatic	C25
Prostate	C61
Renal	C64
Skin	C44
Stomach	C16
Testicular	C62
Thyroid	C73

Table 24: Cancer Site Specific Cohort ICD-10 Code Definitions

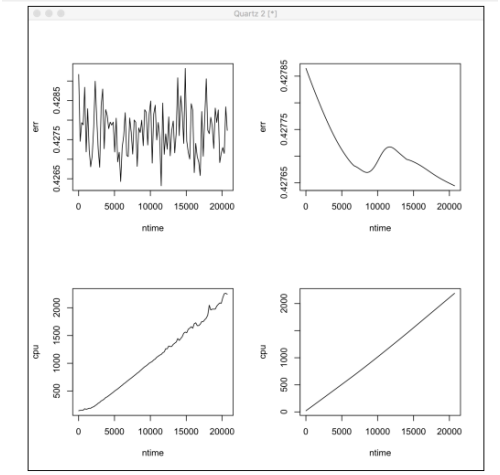
Cancer Cohort	Model Specification
Bladder Brain Breast Colorectal Connective Tissue CUP Laryngeal Leukaemia Liver Lung Lymphoma Melanoma Myeloma Oesophageal Pancreatic Renal Skin Thyroid	Survival ~ Comorbidity of Interest + Age + Gender + Deprivation Quintile
Cervix Endometrium Ovarian Prostate Testicular	Survival ~ Comorbidity of Interest + Age + Deprivation Quintile

Table 25: Model Specification for Cox Models *- Covariates specified in the model development process for the all-cause and cancer cause-specific models in chapters 5 and 6*

Cancer Cohort	Model Specification
Breast Colorectal Lung	Survival ~ Age + Gender + Deprivation Quintile + Stage + Grade + Morphology + Diabetes + CCF + COPD + MI + Stroke + Type 1 Diabetes + Type 2 Diabetes + Other Diabetes + Obesity + Hypertension + Hyperlipidaemia + HIV + Rheumatoid Arthritis + Ankylosing Spondylitis + Psoriatic Arthritis + Gout + TIA + Dementia + MND + Neuromuscular Disease + Other Rheumatological Disease + PAD + Spinal Injury +Paraplegia + IBD + Demyelination + Parkinson's + Liver Dysfunction + Renal Dysfunction + Pancreatitis + Malabsorption + PUD + Venous Disease + Varicosities + Arrhythmias + Angina + Thromboembolic Disease
Prostate	Survival ~ Age + Deprivation Quintile + Stage + Grade + Morphology + Diabetes + CCF + COPD + MI + Stroke + Type 1 Diabetes + Type 2 Diabetes + Other Diabetes + Obesity + Hypertension + Hyperlipidaemia + HIV + Rheumatoid Arthritis + Ankylosing Spondylitis + Psoriatic Arthritis + Gout + TIA + Dementia + MND + Neuromuscular Disease + Other Rheumatological Disease + PAD + Spinal Injury +Paraplegia + IBD + Demyelination + Parkinson's + Liver Dysfunction + Renal Dysfunction + Pancreatitis + Malabsorption + PUD + Venous Disease + Varicosities + Arrhythmias + Angina + Thromboembolic Disease

Table 26: Random Survival Model Specification *- Detailed model specification for each RSF model developed in each site specific cohort*

On 3/6/18 1:06 PM, Hemant Ishwaran wrote:



We've run some simulations to assess CPU usage and to assess performance (C-index) as ntime increases. For our simulation we used a data set with n=25000 and p=15. All forests used ntree=50, nodesize=100, nsplit=10.

In the attached figure, the top row displays performance and the bottom row displays CPU use. Left panels are raw data, right panels are smoothed values using lowess().

Our conclusion is that CPU time increases with increasing ntime and that performance is relatively robust to ntime.

Regarding the latter point, it should emphasized that ntime is merely subsetting the requested survival curves to a small number of time points but it does not degrade the estimate at those time points. In other words, if you used a larger value of ntime, the survival estimates would coincide with the smaller ntime estimates on the intersected time points.

Attaching the code that was used in the analysis:

Hope this helps.
Hemant

Figure 89: Correspondence with Dr Hemant Ishwaran *- Evidence of stable error with reducing ntime hyperparameters when training RSF models*

		Without Stage and Grade	With Stage	With Grade	With Grade and Stage
Bladder	CCF	1.42 (1.20-1.69)		1.36 (1.12-1.66)	
	COPD	1.37 (1.20-1.58)		1.29 (1.10-1.51)	
	DM	1.03 (0.91-1.15)		0.99 (0.87-1.12)	
	MI	1.02 (0.84-1.24)		1.06 (0.86-1.31)	
	Stroke	1.42 (1.16-1.75)		1.29 (1.01-1.65)	
Breast	CCF	2.97 (2.56-3.45)		3.15 (2.66-3.74)	
	COPD	2.11 (1.81-2.47)		2.35 (1.98-2.79)	
	DM	1.39 (1.25-1.55)		1.45 (1.29-1.62)	
	MI	1.66 (1.35-2.06)		2.01 (1.59-2.53)	
	Stroke	1.85 (1.51-2.26)		1.93 (1.54-2.41)	
Cervical	CCF	2.20 (1.14-4.26)	1.21 (0.5-2.93)	1.97 (0.81-4.77)	1.05 (0.34-3.3)
	COPD	1.40 (0.77-2.54)	1.38 (0.69-2.78)	1.35 (0.64-2.84)	1.31 (0.62-2.77)
	DM	2.05 (1.39-3.04)	1.64 (1.09-2.46)	1.72 (1.10-2.70)	1.59 (1.27-1.92)
	MI	1.06 (0.44-2.57)	0.48 (0.178-1.27)	0.52 (0.13-2.10)	0.43 (0.11-1.73)
	Stroke	2.06 (0.97-4.34)	1.24 (0.51-3.02)	1.95 (0.87-4.40)	1.15 (0.42-3.13)
Endometrial	CCF	1.60 (1.08-2.35)	1.70 (1.03-2.79)	1.36 (0.87-2.11)	1.72 (0.97 - 2.63)
	COPD	1.37 (0.93-2.02)	1.38 (0.86-2.19)	1.29 (0.83-2.01)	1.36 (0.84-2.20)
	DM	1.35 (1.13-1.61)	1.45 (1.20-1.76)	1.50 (1.24-1.81)	1.57 (1.27-1.92)
	MI	1.03 (0.58-1.81)	1.00 (0.48-2.11)	1.13 (0.60-2.11)	1.03 (0.46-2.31)
	Stroke	2.24 (1.43-3.53)	2.05 (1.21-3.49)	2.30 (1.46-3.62)	1.94(1.14-3.30)

	CCF	1.97 (1.35-2.84)	2.44 (1.58-3.76)		
	COPD	1.61 (1.03-2.54)	1.67 (1.02-2.74)		
Melanoma	DM	1.29 (1.03-1.63)	1.53 (1.18-1.99)		
	MI	1.25 (0.82-1.90)	1.38 (0.87-2.21)		
	Stroke	1.59 (0.98-2.56)	2.25 (1.32-3.83)		
	CCF	1.58 (1.03-2.44)	1.43 (0.85-2.38)	1.36 (0.75-2.47)	1.45 (0.80-2.64)
	COPD	0.91 (0.60-1.39)	1.66 (1.02 - 2.68)	1.07 (0.64-1.78)	1.57 (0.90-2.72)
Ovarian	DM	0.84 (0.64-1.09)	0.98 (073-1.31)	0.90 (0.66-1.23)	1 (0.72-1.39)
	MI	1.07 (0.55-3.06)	1.57 (0.65-3.78)	1.29 (0.54-3.10)	1.57 (0.59-4.20)
	Stroke	1.47 (0.87-2.48)	1.23 (0.66-2.29)	0.97 (0.48-1.95)	1.11 (0.55-2.22)
	CCF	8.22 (1.14-59.13)	18.17 (2.47-133.47)		
	COPD	2.24 (0.3-16.32)	4.36 (0.58-32.70)		
Testicular	DM	2.15(0.68-6.81)	7.01 (2.16-22.72)		
	MI	0.00 (0-inf)	0.00 (0-inf)		
	Stroke	NA	NA		

Table 27: Sensitivity Analysis of All-Cause Mortality Cox Models *- Comorbidity hazard ratio point estimates and confidence intervals from Cox Models with and with the inclusion of stage and grade data on a complete case analysis for cancer sites with 30% or less missing data for each of these data items. Combinations of cancer and comorbidity where the models without stage and grade showed a low probability of unidirectional hazard are coloured grey.*

Table 27 presents the results of the sensitivity analysis completed using the all-cause mortality cox models. Each cancer cohort where less than 30% of stage and grade data was missing was analysed with and without the inclusion of stage and grade data. Where one of these data items had more than 30% missing this was not analysed and is represented as a greyed out section within the table. This sensitivity analysis highlights that in all examples where the results without stage and grade show a high probability of unidirectional hazard the results with adjustment for stage and grade data either singly or together maintains the same direction of effect for the point estimate. In most instances the degree of change within the point estimate is modest however several examples a more clinically meaningful change is seen. One such example is CCF in cervical cancer which changes from 2.20 (1.14-4.26) to 1.05 (0.34-3.3) after adjustment for stage and grade. It is however worth noting that in this case the original point estimate is within the new confidence interval for

the models with full adjustment. As would be expected due to an increased number of covariates and smaller numbers of cases due to the exclusion of incomplete cases the confidence intervals for the models adjusting for stage and grade have wider confidence intervals. This effect is amplified by the cohort size with more uncertainty occurring in those with smaller populations such as testicular cancer. This also explains to some extent the larger shifts in point estimates seen in these populations. The results of this sensitivity analysis suggest that the results obtained within chapter 5 are likely to be robust in terms of direction of effect, however may vary somewhat in terms of scale of effect when grade and stage data is included. There does however remain the debate as to whether this form of adjustment should be conducted as previous studies have demonstrated how inclusion of stage and grade adjustment biases results.[363]

		Without Stage and Grade	With Stage	With Grade	With Grade and Stage
Bladder	CCF	0.89 (0.66-1.22)		0.66 (0.45-0.95)	
	COPD	0.91 (0.72-1.14)		0.76 (0.58-1)	
	DM	0.73 (0.61-0.88)		0.63 (0.51-0.78)	
	MI	0.80 (0.59-1.07)		0.87 (0.64-1.18)	
	Stroke	1.11 (0.80-1.55)		0.79 (0.52-1.20)	
Breast	CCF	2.12 (1.59-2.82)		1.79 (1.27-2.53)	
	COPD	1.16 (0.85-1.60)		1.15 (0.81-1.64)	
	DM	0.86 (0.71-1.04		0.86 (0.70-1.06)	
	MI	1.29 (0.88-1.89)		1.21 (0.78-1.89)	
	Stroke	1.45 (1.03-2.04)		1.47 (1.02-2.14)	
Cervical	CCF	1.65 (0.53-5.13)	0.89 (0.36-2.07)	0.51 (0.43-1.94)	0.44(0.06-3.14)
	COPD	0.96 (0.40-2.31)	1.52 (0.48-2.78)	0.98 (0.36-2.61)	0.98 (0.35-2.63)
	DM	2.21 (1.35-3.63)	1.41 (0.85-2.33)	1.53 (0.86-2.73)	1.43 (0.79-2.58)
	MI	1.13 (0.42-3.03)	0.45 (0.14-1.41)	0.34 (0.05-2.43)	0.28 (0.04-2.01)
	Stroke	1.60 (0.59-4.29)	1.19 (0.44-3.21)	1.27 (0.40-3.97)	1.04 (0.33-3.29)

	CCF	1.21 (0.67-2.20)	0.86 (0.72-3.65)	0.92 (0.44-1.94)	0.74 (0.06-3.15)
	COPD	0.93 (0.50-1.73)	0.67 (0.28-1.62)	0.73 (0.33-1.63)	0.58 (0.22-1.55)
Endometrial	DM	0.98 (0.75-1.29)	1.00 (0.74-1.35)	1.07 (0.79-1.44)	1.06 (0.76-1.47)
	MI	1.17 (0.52-2.62)	0.70 (0.18-2.82)	1.36 (0.57-3.29)	0.93 (0.23-3.73)
	Stroke	2.15 (1.22-3.80)	1.80 (0.85-3.79)	2.40 (1.36-4.25)	1.61 (0.76-3.41)
	CCF	0.88 (0.42-1.86)	1.63 (0.72-3.65)		
	COPD	1.09 (0.52-2.30)	1.46 (0.65-3.28)		
Melanoma	DM	0.79 (0.52-1.19)	0.86 (0.51-1.44)		
	MI	0.53 (0.22-1.27)	0.69 (0.30-1.86)		
	Stroke	1.68 (0.87-3.23)	2.58 (1.22-5.46)		
	CCF	1.45 (0.89-2.39)	1.41 (0.78-2.56)	1.18 (0.56-2.49)	1.25 (0.59-2.65)
	COPD	1.01 (0.63-1.63)	2.02 (1.19-3.44)	1.26 (0.71-2.22)	2.05 (1.13-3.74)
Ovarian	DM	0.69 (0.49-0.96)	0.76 (0.52-1.12)	0.69 (0.46-1.04)	0.77 (0.50-1.19)
	MI	1.03 (0.49-2.17)	1.28 (0.41-3.98)	0.97 (0.31-3.02)	1.06 (0.27-4.27)
	Stroke	1.19 (0.62-2.29)	0.93 (0.42-2.07)	0.77 (0.32-1.87)	0.91 (0.38-2.18)
	CCF	0 (0-inf)	0 (0-inf)		
	COPD	0 (0-inf)	0 (0-inf)		
Testicular	DM	1.41 (0.19-10.29)	15.01 (1.88-119.83)		
	MI	0 (0-inf)	0 (0-inf)		
	Stroke	NA	NA		

Table 28: Sensitivity Analysis of Cancer Cause-Specific Mortality Cox Models *- Comorbidity hazard ratio point estimates and confidence intervals from cancer cause –specific Cox Models with and with the inclusion of stage and grade data on a complete case analysis for cancer sites with 30% or less missing data for each of these data items. Combinations of cancer and comorbidity where the models without stage and grade showed a low probability of unidirectional hazard are coloured grey.*

The results for the sensitivity analysis in the cancer cause specific models are presented in **Table 28**. These show similar results to those of the all-cause mortality models, with those modes showing a high probability of unidirectional hazard maintaining their direction of effect when adjusting for stage and grade. As the precision of estimates for the cause specific models is already reduced due to the removal of cases with no cause of death data and higher levels of censoring the further removal of cases with missing data for stage and grade makes these estimates even less precise. In many instances this results in the adjusted models showing a low probability of unidirectional hazard due to wide confidence intervals. This makes the interpretation of these more challenging. In order to gain a clearer understanding of the effects of these adjustments a larger dataset would be needed in order to improve the precision of results and could be a focus of future work.

Bibliography

1. Cancer Research UK. Cancer Survival Statistics. Published 2017. Accessed December 20, 2020. http://www.cancerresearchuk.org/health-professional/cancer-statistics/survival/all-cancers-combined

2. Lifetime risk of cancer. Cancer Research UK. Published May 13, 2015. Accessed December 10, 2020. https://www.cancerresearchuk.org/health-professional/cancer-statistics/risk/lifetime-risk

3. The National Audit Office, The Comptroller and Auditor General. *Progress in Improving Cancer Services and Outcomes in England*.; 2015.

4. The Office For National Statistics. Expenditure on Healthcare in the UK: 2013. Published 2015. Accessed July 29, 2020. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthcaresystem/articles/expenditureonhealthcareintheuk/2015-03-26

5. Laudicella M, Walsh B, Burns E, Smith PC. Cost of care for cancer patients in England: evidence from population-based patient-level data. *Br J Cancer*. 2016;114(11):1286-1292. doi:10.1038/bjc.2016.77

6. The National Audit Office, The Comptroller and Auditor General. *Progress in Improving Cancer Services and Outcomes in England*.; 2015.

7. Arnold M, Rutherford MJ, Bardot A, et al. Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (ICBP SURVMARK-2): a population-based study. *Lancet Oncol*. 2019;20(11):1493-1505. doi:10.1016/S1470-2045(19)30456-5

8. Coleman MP, Gatta G, Verdecchia A, et al. EUROCARE-3 summary: cancer survival in Europe at the end of the 20th century. *Ann Oncol*. 2003;14:v128-v149. doi:10.1093/annonc/mdg756

9. Munro AJ. Interpretation of EUROCARE-5. *Lancet Oncol*. 2014;15(1):2-3. doi:10.1016/S1470-2045(13)70566-7

10. De Angelis R, Sant M, Coleman MP, et al. Cancer survival in Europe 1999-2007 by country and age: results of EUROCARE--5-a population-based study. *Lancet Oncol*. 2014;15(1):23-34. doi:10.1016/S1470-2045(13)70546-1

11. Souhami R. Are UK cancer cure rates worse than in most other European countries? *Br J Gen Pract*. 2010;60(571):81-82. doi:10.3399/bjgp10X483102

12. Woods LM, Coleman MP, Lawrence G, Rashbass J, Berrino F, Rachet B. Evidence against the proposition that “UK cancer survival statistics are misleading”: simulation study with National Cancer Registry data. *BMJ*. Published online 2011. doi:10.1136/bmj.d3399

13. Wilkinson E. Questions remain over validity of EUROCARE data. *Lancet Lond Engl*. 2009;374(9694):964-965. doi:10.1016/s0140-6736(09)61648-2

14. Shrestha A, Martin C, Burton M, Walters S, Collins K, Wyld L. Quality of life versus length of life considerations in cancer patients: A systematic literature review. *Psychooncology*. 2019;28(7):1367-1380. doi:10.1002/pon.5054

15. Lavelle K, Downing A, Thomas J, Lawrence G, Forman D, Oliver SE. Are lower rates of surgery amongst older women with breast cancer in the UK explained by co-morbidity. *Br J Cancer*. 2012;107(7):1175-1180. doi:10.1038/bjc.2012.192

16. Lavelle K, Sowerbutts AM, Bundred N, et al. Is lack of surgery for older breast cancer patients in the UK explained by patient choice or poor health? A prospective cohort study. *Br J Cancer*. 2014;110(3):573-583. doi:10.1038/bjc.2013.734

17. Paul C, Boyes A, Hall A, Bisquera A, Miller A, O'Brien L. The impact of cancer diagnosis and treatment on employment, income, treatment decisions and financial assistance and their relationship to socioeconomic and disease factors. *Support Care Cancer*. 2016;24(11):4739-4746. doi:10.1007/s00520-016-3323-y

18. Bates T, Evans T, Lagord C, Monypenny I, Kearins O, Lawrence G. A population based study of variations in operation rates for breast cancer, of comorbidity and prognosis at diagnosis: Failure to operate for early breast cancer in older women. *Eur J Surg Oncol*. 2014;40(10):1230-1236. doi:10.1016/j.ejso.2014.06.001

19. Blanco JAG, Toste IS, Alvarez RF, Cuadrado GR, Gonzalvez AM, Martín IJG. Age, comorbidity, treatment decision and prognosis in lung cancer. *Age Ageing*. 2008;37(6):715-718. doi:10.1093/ageing/afn226

20. Dehal A, Abbas A, Johna S. Comorbidity and outcomes after surgery among women with breast cancer: analysis of nationwide in-patient sample database. *Breast Cancer Res Treat*. 2013;139(2):469-476. doi:10.1007/s10549-013-2543-9

21. Ahn DH, Mehta N, Yorio JT, Xie Y, Yan J, Gerber DE. Influence of medical comorbidities on the presentation and outcomes of stage I–III non-small cell lung cancer. *Clin Lung Cancer*. 2013;14(6):644-650. doi:10.1016/j.cllc.2013.06.009

22. Gayar OH, Patel S, Schultz D, Mahan M, Rasool N, Elshaikh MA. The impact of tumor grade on survival end points and patterns of recurrence of 949 patients with early-stage endometrioid carcinoma: a single institution study. *Int J Gynecol Cancer Off J Int Gynecol Cancer Soc*. 2014;24(1):97-101. doi:10.1097/IGC.0000000000000018

23. Grann AF, Frøslev T, Olesen AB, Schmidt H, Lash TL. The impact of comorbidity and stage on prognosis of Danish melanoma patients, 1987-2009: A registry-based cohort study. *Br J Cancer*. 2013;109(1):265-271. doi:10.1038/bjc.2013.246

24. Sarfati D, Koczwara B, Jackson C. The impact of comorbidity on cancer and its treatments. *CA Cancer J Clin*. 2016;66(4):337-350. doi:10.3322/caac.21342.

25. Wolff JL, Starfield B, Anderson G. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. *Arch Intern Med*. 2002;162(20):2269-2276. doi:10.1001/archinte.162.20.2269

26. Starfield B. Threads and Yarns: Weaving the Tapestry of Comorbidity. *Ann Fam Med*. 2006;4(2):101-103. doi:10.1370/afm.524

27. Fortin M, Soubhi H, Hudon C, Bayliss EA, van den Akker M. Multimorbidity's many challenges. *BMJ*. 2007;334(7602):1016-1017. doi:10.1136/bmj.39201.463819.2C

28. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: A cross-sectional study. *The Lancet*. 2012;380(9836):37-43. doi:10.1016/S0140-6736(12)60240-2

29. Baijal P, Periyakoil V. Understanding frailty in cancer patients. *Cancer J Sudbury Mass*. 2014;20(5):358-366. doi:10.1097/PPO.0000000000000068

30. Hartley P, Adamson J, Cunningham C, Embleton G, Romero-Ortuno R. Clinical frailty and functional trajectories in hospitalized older adults: A retrospective observational study. *Geriatr Gerontol Int*. 2017;17(7):1063-1068. doi:10.1111/ggi.12827

31. Valderas JM, Starfield B, Sibbald B, Salisbury C, Rloand M. Defining comorbidity: implications for understanding health and health services. *Ann Fam Med*. 2009;7:357-363. doi:10.1370/afm.983.Martin

32. Leyfer O, Brown TA. The Anxiety-Depression Spectrum. The Oxford Handbook of Clinical Psychology. doi:10.1093/oxfordhb/9780199328710.013.033

33. Agrawal S. Late effects of cancer treatment in breast cancer survivors. *South Asian J Cancer*. 2014;3(2):112-115. doi:10.4103/2278-330X.130445

34. Lenihan DJ, Cardinale DM. Late cardiac effects of cancer treatment. *J Clin Oncol*. 2012;30(30):3657-3664. doi:10.1200/JCO.2012.45.2938

35. Maher EJ, Denton A. Survivorship, Late Effects and Cancer of the Cervix. *Clin Oncol*. 2008;20(6):479-487. doi:10.1016/j.clon.2008.04.009

36. Nekhlyudov L, Aziz NM, Lerro C, Virgo KS. Oncologists' and primary care physicians' awareness of late and long-term effects of chemotherapy: Implications for care of the growing population of survivors. *J Oncol Pract*. 2014;10(2):29-36. doi:10.1200/JOP.2013.001121

37. Ganz PA. Late effects of cancer and its treatment. *Semin Oncol Nurs*. 2001;17(4):241-248. doi:10.1053/sonu.2001.27914

38. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J Chronic Dis*. Published online 1987. doi:10.1016/0021-9681(87)90171-8

39. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity Measures for Use with Administrative Data. *Med Care*. Published online 1998. doi:10.1097/00005650-199801000-00004

40. Clegg A, Bates C, Young J, et al. Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age Ageing*. 2016;45(3):353-360. doi:10.1093/ageing/afw039

41. Gilbert T, Neuburger J, Kraindler J, et al. Development and validation of a Hospital Frailty Risk Score focusing on older people in acute care settings using electronic hospital records : an observational study. *The Lancet*. 391(10132):1775-1782. doi:10.1016/S0140-6736(18)30668-8

42. Divo MJ, Martinez CH, Mannino DM. Ageing and the epidemiology of multimorbidity. *Eur Respir J*. 2014;44(4):1055-1068. doi:10.1183/09031936.00059814

43. Gurney J, Sarfati D, Stanley J. The impact of patient comorbidity on cancer stage at diagnosis. *Br J Cancer*. 2015;113(9):1375-1380. doi:10.1038/bjc.2015.355

44. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. *BMJ Evid-Based Med*. 2016;21(4):125-127. doi:10.1136/ebmed-2016-110401

45. Petrisor B, Bhandari M. The hierarchy of evidence: Levels and grades of recommendation. *Indian J Orthop*. 2007;41(1):11-15. doi:10.4103/0019-5413.30519

46. Speich B, von Niederhäusern B, Schur N, et al. Systematic review on costs and resource use of randomized clinical trials shows a lack of transparent and comprehensive data. *J Clin Epidemiol*. 2018;96:1-11. doi:10.1016/j.jclinepi.2017.12.018

47. Frieden TR. Evidence for Health Decision Making — Beyond Randomized, Controlled Trials. *N Engl J Med*. 2017;377(5):465-475. doi:10.1056/NEJMra1614394

48. Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. *Soc Sci Med*. 2018;210:2-21. doi:10.1016/j.socscimed.2017.12.005

49. Silverman SL. From Randomized Controlled Trials to Observational Studies. *Am J Med*. 2009;122(2):114-120. doi:10.1016/j.amjmed.2008.09.030

50. Vivek H. Murthy, Harlan M. Krumholz CPG. Participation in Cancer Clinical Trials Race-, Sex-, and Age-Based Disparities. *JAMA*. 2004;291(22):2720-2726.

51. He J, Morales DR, Guthrie B. Exclusion rates in randomized controlled trials of treatments for physical conditions: a systematic review. *Trials*. 2020;21(1):228. doi:10.1186/s13063-020-4139-0

52. Jin S, Pazdur R, Sridhara R. Re-Evaluating Eligibility Criteria for Oncology Clinical Trials: Analysis of Investigational New Drug Applications in 2015. *J Clin Oncol*. 2017;35(33):3745-3752. doi:10.1200/JCO.2017.73.4186

53. Denicoff AM, McCaskill-Stevens W, Grubbs SS, et al. The National Cancer Institute-American Society of Clinical Oncology Cancer Trial Accrual Symposium: summary and recommendations. *J Oncol Pract Am Soc Clin Oncol*. 2013;9(6):267-276. doi:10.1200/JOP.2013.001119

54. Kukull WA, Ganguli M. Generalizability. *Neurology*. 2012;78(23):1886-1891. doi:10.1212/WNL.0b013e318258f812

55. Logviss K, Krievins D, Purvina S. Characteristics of clinical trials in rare vs. common diseases: A register-based Latvian study. *PLoS ONE*. 2018;13(4). doi:10.1371/journal.pone.0194494

56. Kasenda B, Schandelmaier S, Sun X, et al. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications. *BMJ*. 2014;349. doi:10.1136/bmj.g4539

57. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *Jama*. 2018;02115. doi:10.1001/jama.2017.18391

58. Yang W, Zilov A, Soewondo P, Bech OM, Sekkal F, Home PD. Observational studies: Going beyond the boundaries of randomized controlled trials. *Diabetes Res Clin Pract*. 2010;88(SUPPL. 1):3-9. doi:10.1016/S0168-8227(10)70002-4

59. Brennan P, Croft P. Interpreting the results of observational research: chance is not such a fine thing. *BMJ*. 1994;309(6956):727-730.

60. Carlson MDA, Morrison RS. Study Design, Precision, and Validity in Observational Studies. *J Palliat Med*. 2009;12(1):77-82. doi:10.1089/jpm.2008.9690

61. Ludwig DA. Use and misuse of p-values in designed and observational studies: guide for researchers and reviewers. *Aviat Space Environ Med*. 2005;76(7):675-680.

62. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Routinely collected data and comparative effectiveness evidence: Promises and limitations. *Cmaj*. 2016;188(8):E158-E164. doi:10.1503/cmaj.150653

63. Mc Cord KA, Al-Shahi Salman R, Treweek S, et al. Routinely collected data for randomized trials: Promises, barriers, and implications. *Trials*. 2018;19(1). doi:10.1186/s13063-017-2394-5

64. Kong H-J. Managing Unstructured Big Data in Healthcare System. *Healthc Inform Res*. 2019;25(1):1-2. doi:10.4258/hir.2019.25.1.1

65. Koutkias V. From Data Silos to Standardized, Linked, and FAIR Data for Pharmacovigilance: Current Advances and Challenges with Observational Healthcare Data. *Drug Saf*. 2019;42(5):583-586. doi:10.1007/s40264-018-00793-z

66. Miller AR, Tucker C. Health information exchange, system size and information silos. *J Health Econ*. 2014;33:28-42. doi:10.1016/j.jhealeco.2013.10.004

67. Caldicott F. The Information Governance Review. *Inf Gov Rev*. 2013;(March):139.

68. The European Parliament, The European Council. General Data Protection Regulation. *Off J Eur Union*. 2016;2014(October 1995):20-30. doi:http://eur-lex.europa.eu/pri/en/oj/dat/2003/l_285/l_28520031101en00330037.pdf

69. Chen F, Jiang X, Wang S, et al. Perfectly Secure and Efficient Two-party Electronic Health Record Linkage. *IEEE Internet Comput*. 2018;(April):32-41. doi:10.1109/MIC.2018.112102542

70. Boyd KM. Ethnicity and the ethics of data linkage. *BMC Public Health*. 2007;7:318. doi:10.1186/1471-2458-7-318

71. Kwakkenbos L, Imran M, McCall SJ, et al. CONSORT extension for the reporting of randomised controlled trials conducted using cohorts and routinely collected data (CONSORT-ROUTINE): checklist with explanation and elaboration. *BMJ*. 2021;373:n857. doi:10.1136/bmj.n857

72. Hemkens LG. How Routinely Collected Data for Randomized Trials Provide Long-term Randomized Real-World Evidence. *JAMA Netw Open*. 2018;1(8):e186014. doi:10.1001/jamanetworkopen.2018.6014

73. Tannock IF, Amir E, Booth CM, et al. Relevance of randomised controlled trials in oncology. *Lancet Oncol*. 2016;17(12):e560-e567. doi:10.1016/S1470-2045(16)30572-1

74. Moore TJ, Heyward J, Anderson G, Alexander GC. Variation in the estimated costs of pivotal clinical benefit trials supporting the US approval of new therapeutic agents, 2015–2017: a cross-sectional study. *BMJ Open*. 2020;10(6):e038863. doi:10.1136/bmjopen-2020-038863

75. Cord KAM, Ewald H, Agarwal A, et al. Treatment effects in randomised trials using routinely collected data for outcome assessment versus traditional trials: meta-research study. *BMJ*. 2021;372:n450. doi:10.1136/bmj.n450

76. Lensen S, Macnair A, Love SB, et al. Access to routinely collected health data for clinical trials – review of successful data requests to UK registries. *Trials*. 2020;21(1):398. doi:10.1186/s13063-020-04329-8

77. International Association of Cancer Registries. Published 2021. Accessed June 16, 2021. http://www.iacr.com.fr/index.php?option=com_content&view=article&id=111&Itemid=435

78. Minor LB. Harnessing the Power of Data in Health. *Stanf Med Health Trends Rep*. 2017;(June).

79. Brynjolfsson E, McAfee A. Big Data : The Management Revolution. *Harv Bus Rev*. 2012;(October).

80. Gantz J, Reinsel D, Shadows BD. The Digital Universe in 2020. *IDC IView Big Data Bigger Digit Shad Biggest Growth Far East*. 2012;2007(December 2012):1-16.

81. Reinsel D, Gantz J, Rydning J. Data Age 2025 - The Evolution of Data to Life-Critical: Don ' t Focus on Big Data ; Focus on the Data That ' s Big. *IDC White Pap*. 2017;(April):1-25.

82. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. *J Big Data*. 2015;2(1):1-21. doi:10.1186/s40537-014-0007-7

83. Emmanuel I, Stanier C. Defining Big Data. In: *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*. BDAW '16. Association for Computing Machinery; 2016:1-6. doi:10.1145/3010089.3010090

84. Computational Statistics. In: *Encyclopedia of Measurement and Statistics*. Sage Publications, Inc.; 2007. doi:10.4135/9781412952644.n97

85. Principles of Data Mining | SpringerLink. Accessed November 30, 2020. https://link.springer.com/article/10.2165/00002018-200730070-00010

86. What is Data Science? Published June 3, 2020. Accessed November 30, 2020. https://www.ibm.com/cloud/learn/data-science-introduction

87. Statistics vs Data Science: What's the Difference? Displayr. Published August 7, 2018. Accessed November 30, 2020. https://www.displayr.com/statistics-vs-data-science-whats-the-difference/

88. Data Science vs Statistics | Know Top 5 Beneficial Comparisons. EDUCBA. Published March 14, 2018. Accessed November 30, 2020. https://www.educba.com/data-science-vs-statistics/

89. Leek JT, Peng RD. What is the question? *Science*. 20/032015;347(6228):1314-1315.

90. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods*. 2018;15(4):233-234. doi:10.1038/nmeth.4642

91. Callahan A, Shah NH. Chapter 19 - Machine Learning in Healthcare. In: Sheikh A, Cresswell KM, Wright A, Bates DW, eds. *Key Advances in Clinical Informatics*. Academic Press; 2017:279-291. doi:10.1016/B978-0-12-809523-2.00019-4

92. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*. 2019;19(1):64. doi:10.1186/s12874-019-0681-4

93. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. *RadioGraphics*. 2017;37(2):505-515. doi:10.1148/rg.2017160130

94. Libbrecht M, Noble W. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;16:321-332.

95. Thomsen K, Iversen L, Titlestad TL, Winther O. Systematic review of machine learning for diagnosis and prognosis in dermatology. *J Dermatol Treat*. 2020;31(5):496-510. doi:10.1080/09546634.2019.1682500

96. Fauw JD, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342-1350. doi:10.1038/s41591-018-0107-6

97. Panch T, Mattie H, Celi LA. The “inconvenient truth” about AI in healthcare. *Npj Digit Med*. 2019;2(1):1-3. doi:10.1038/s41746-019-0155-4

98. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. Published online 2017:3-9. doi:1711.05225

99. Lukeoakdenrayner ~. CheXNet: an in-depth review. Luke Oakden-Rayner. Published January 24, 2018. Accessed November 30, 2020. https://lukeoakdenrayner.wordpress.com/2018/01/24/chexnet-an-in-depth-review/

100. Lukeoakdenrayner ~. Exploring the ChestXray14 dataset: problems. Luke Oakden-Rayner. Published December 18, 2017. Accessed November 30, 2020. https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/

101. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *CSBJ*. 2015;13:8-17. doi:10.1016/j.csbj.2014.11.005

102. Bibault J-E, Giraud P, Burgun A. Big Data and machine learning in radiation oncology: State of the art and future prospects. *Cancer Lett*. 2016;382(1):110-117. doi:10.1016/j.canlet.2016.05.033

103. Nagy M, Radakovich N, Nazha A. Machine Learning in Oncology: What Should Clinicians Know? *JCO Clin Cancer Inform*. 2020;4:799-810.

104. Panahiazar M, Taslimitehrani V, Pereira N, Pathak J. Using EHRs and Machine Learning for Heart Failure Survival Analysis. *Stud Health Technol Inform*. 2015;216:40-44.

105. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS ONE*. 2018;13(8):1-20. doi:10.1371/journal.pone.0202344

106. Wang P, Li Y, Reddy CK. Machine Learning for Survival Analysis: A Survey. *ACM Comput Surv*. 2019;51(6):110:1-110:36. doi:10.1145/3214306

107. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *J Am Stat Assoc*. 1958;53(282):457-481. doi:10.1080/01621459.1958.10501452

108. Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B Methodol*. 1972;34(2):187-202. doi:10.1111/j.2517-6161.1972.tb00899.x

109. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2(3):841-860. doi:10.1214/08-AOAS169

110. Gelman A. Discussion: difficulties in making inferences about scientific truth from distributions of published p-values. In: *Biostatistics 15:18–23 DOI 10.1093/Biostatistics/Kxt034*. ; 2014.

111. Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*. 2014;15(1):1-12. doi:10.1093/biostatistics/kxt007

112. Ioannidis JPA. Why most published research findings are false. *Get Good Res Integr Biomed Sci*. 2018;2(8):2-8. doi:10.1371/journal.pmed.0020124

113. Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology*. 2008;19(5):640-648. doi:10.1097/EDE.0b013e31818131e7

114. Leek JT, Peng RD. What is the question? *Science*. 2015;347(6228):1314-1315. doi:10.1126/science.aaa6146

115. What is Data Tampering | IGI Global. Accessed December 10, 2020. https://www.igi-global.com/dictionary/data-tampering/6834

116. Wang MQ, Yan AF, Katz RV. Researcher Requests for Inappropriate Analysis and Reporting: A U.S. Survey of Consulting Biostatisticians. *Ann Intern Med*. 2018;169(8):554-558. doi:10.7326/M18-1230

117. Localio AR, Stack CB, Meibohm AR, et al. Inappropriate Statistical Analysis and Reporting in Medical Research: Perverse Incentives and Institutional Solutions. *Ann Intern Med*. 2018;169(8):577-578. doi:10.7326/M18-2516

118. Hernán MA, Takkouche B, Caamaño-Isorna F, Gestal-Otero JJ. A meta-analysis of coffee drinking, cigarette smoking, and the risk of Parkinson's disease. *Ann Neurol*. 2002;52(3):276-284. doi:10.1002/ana.10277

119. Hernán MA. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. *Am J Public Health*. 2018;108(5):616-619. doi:10.2105/AJPH.2018.304337

120. Listl S, Jürges H, Watt RG. Causal inference from observational data. *Community Dent Oral Epidemiol*. 2016;44(5):409-415. doi:10.1111/cdoe.12231

121. Nichols A. Causal Inference with Observational Data. *Stata J*. 2007;7(4):507-541. doi:10.1177/1536867X0800700403

122. SNOMED - Home | SNOMED International. Accessed September 25, 2020. http://www.snomed.org/

123. NHS Digital. Read Codes. Accessed June 26, 2018. https://digital.nhs.uk/services/terminology-and-classifications/read-codes

124. NHS Digital. Read Code Browser. Accessed June 26, 2018. https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/9

125. World Health Organisation. International Disease Classification. Published 2020. Accessed April 28, 2021. http://www.who.int/classifications/icd/en/

126. NHS Digital. Hospital Episode Statistics. Published November 19, 2020. Accessed April 28, 2021. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics

127. January 2013 14. An introduction to clinical coding. Health Service Journal. Accessed September 30, 2020. https://www.hsj.co.uk/technology-and-innovation/an-introduction-to-clinical-coding/5052917.article

128. Public Health England. Guidance: National Cancer Registration and Analysis Service (NCRAS). Published 2016. Accessed November 21, 2019. https://www.gov.uk/guidance/national-cancer-registration-and-analysis-service-ncras

129. Brierley JD, Gospodarowicz MK, Wittekind C. *TNM Classification of Malignant Tumours*. John Wiley & Sons; 2017.

130. Guo W, Polich ED, Su J, et al. Ageing and the epidemiology of multimorbidity Miguel. *Cell Rep*. 2015;11(10):1651-1666. doi:10.1080/10937404.2015.1051611.INHALATION

131. Hubbard RE, Peel NM, Samanta M, Gray LC, Mitnitski A, Rockwood K. Frailty status at admission to hospital predicts multiple adverse outcomes. *Age Ageing*. 2017;46(5):801-806. doi:10.1093/ageing/afx081

132. Olaroiu M, Ghinescu M, Naumov V, Brinza I, Den Heuvel WV. Does Frailty Predict Health Care Utilization in Community-Living Older Romanians? *Curr Gerontol Geriatr Res*. 2016;2016. doi:10.1155/2016/6851768

133. Terret C, Castel-Kremer E, Albrand G, Droz JP. Effects of comorbidity on screening and early diagnosis of cancer in elderly people. *Lancet Oncol*. 2009;10(1):80-87. doi:10.1016/S1470-2045(08)70336-X

134. Piccirillo JF, Creech C, Zequeira R, Anderson S, Johnson A. Inclusion of Comorbidity into Oncology Data Registries. *J Regist Manag*. 1999;26(2):66-70.

135. Piccirillo JF, Tierney RM, Costas I, Grove L, Spitznagel EL. Prognostic importance of comorbidity in a hospital-based cancer registry. *J Am Med Assoc*. 2004;291(20):2441-2447. doi:10.1001/jama.291.20.2441

136. Abdel-Rahman O. Impact of diabetes comorbidity on the efficacy and safety of FOLFOX first-line chemotherapy among patients with metastatic colorectal cancer: a pooled analysis of two phase-III studies. *Clin Transl Oncol Off Publ Fed Span Oncol Soc Natl Cancer Inst Mex*. 2019;21(4):512-518.

137. Mehta HB, Dimou F, Adhikari D, et al. Comparison of Comorbidity Scores in Predicting Surgical Outcomes. *Med Care*. 2016;54(2):180-187. doi:10.1097/MLR.0000000000000465

138. Daver N, Naqvi K, Jabbour E, et al. Impact of comorbidities by ACE-27 in the revised-IPSS for patients with myelodysplastic syndromes. *Am J Hematol*. 2014;89(5):509-516. doi:10.1002/ajh.23675

139. Fleming ST, Sabatino SA, Kimmick G, et al. Developing a Claim-based Version of the ACE-27 Comorbidity Index: A Comparison With Medical Record Review. *Med Care*. 2011;49(8):752-760.

140. Kallogjeri D, Piccirillo JF, Spitznagel EL, Steyerberg EW. Comparison of scoring methods for ACE-27: Simpler is better. *J Geriatr Oncol*. 2012;3(3):238-245. doi:10.1016/j.jgo.2012.01.006

141. Monteiro AR, Garcia AR, Pereira TC, et al. ACE-27 as a prognostic tool of severe acute toxicities in patients with head and neck cancer treated with chemoradiotherapy: a real-world, prospective, observational study. *Support Care Cancer*. Published online August 13, 2020. doi:10.1007/s00520-020-05679-4

142. Rogers SN, Aziz A, Lowe D, Husband DJ. Feasibility study of the retrospective use of the Adult Comorbidity Evaluation index (ACE-27) in patients with cancer of the head and neck who had radiotherapy. *Br J Oral Maxillofac Surg*. 2006;44(4):283-288. doi:10.1016/j.bjoms.2005.06.025

143. Hall G, The Kings Fund. Embedding technology in health and social care: PPM+ and the Leeds Care Record. Accessed June 26, 2018. https://www.kingsfund.org.uk/sites/default/files/media/Geoff_Hall.pdf

144. Matsuo K, Machida H, Mandelbaum RS, Konishi I, Mikami M. Validation of the 2018 FIGO cervical cancer staging system. *Gynecol Oncol*. 2019;152(1):87-93. doi:10.1016/j.ygyno.2018.10.026

145. Information NC for B, Pike USNL of M 8600 R, MD B, Usa 20894. *Glycated Haemoglobin (HbA1c) for the Diagnosis of Diabetes*. World Health Organization; 2011. Accessed September 28, 2020. https://www.ncbi.nlm.nih.gov/books/NBK304271/

146. Diagnostic criteria for diabetes. Diabetes UK. Accessed December 10, 2020. https://www.diabetes.org.uk/professionals/position-statements-reports/diagnosis-ongoing-management-monitoring/new_diagnostic_criteria_for_diabetes

147. Diabetes UK. The HbA1c test. Diabetes. Published January 15, 2019. Accessed December 10, 2020. https://www.diabetes.co.uk/hba1c-test.html

148. Body Mass Index - an overview | ScienceDirect Topics. Accessed September 28, 2020. https://www.sciencedirect.com/topics/medicine-and-dentistry/body-mass-index

149. The Office For National Statistics. English indices of deprivation 2019. Published 2019. Accessed July 24, 2020. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

150. Ministry of Housing C and LG. English Indices of Deprivation 2015 - LSOA Level. Accessed June 26, 2018. https://data.gov.uk/dataset/8f601edb-6974-417e-9c9d-85832dd2bbf2/english-indices-of-deprivation-2015-lsoa-level

151. Clelland D, Hill C. Deprivation, policy and rurality: The limitations and applications of area-based deprivation indices in Scotland. *Local Econ*. 2019;34(1):33-50. doi:10.1177/0269094219827893

152. Deas I, Robson B, Wong C, Bradford M. Measuring Neighbourhood Deprivation: A Critique of the Index of Multiple Deprivation. *Environ Plan C Gov Policy*. 2003;21(6):883-903. doi:10.1068/c0240

153. Cheah Y-W, Plale B. Provenance analysis: Towards quality provenance. In: *2012 IEEE 8th International Conference on E-Science*. ; 2012:1-8. doi:10.1109/eScience.2012.6404480

154. Buneman P, Khanna S, Wang-Chiew T. Why and Where: A Characterization of Data Provenance. In: Van den Bussche J, Vianu V, eds. *Database Theory — ICDT 2001*. Lecture Notes in Computer Science. Springer; 2001:316-330. doi:10.1007/3-540-44503-X_20

155. Mackintosh J. Beware lessons of history when dealing with quirky indices. Published August 24, 2015. Accessed September 23, 2020. https://www-ft-com.libezproxy.open.ac.uk/content/40e78992-481e-11e5-af2f-4d6e0e5eda22

156. Digital N. Quality Outcomes Framework. Accessed December 12, 2017. http://content.digital.nhs.uk/qof

157. Mukaka M, White SA, Terlouw DJ, Mwapasa V, Kalilani-Phiri L, Faragher EB. Is using multiple imputation better than complete case analysis for estimating a prevalence (risk) difference in randomized controlled trials when binary outcome observations are missing? *Trials*. 2016;17. doi:10.1186/s13063-016-1473-3

158. Soley-Bori M. Dealing with missing data: Key assumptions and methods for applied analysis. https://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf

159. Lee KJ, Carlin JB. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *Am J Epidemiol*. Published online 2010. doi:10.1093/aje/kwp425

160. Sullivan TR, Lee KJ, Ryan P, Salter AB. Multiple imputation for handling missing outcome data when estimating the relative risk. *BMC Med Res Methodol*. Published online 2017. doi:10.1186/s12874-017-0414-5

161. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? *Int J Methods Psychiatr Res*. Published online 2011. doi:10.1002/mpr.329

162. Dong Y, Peng C-YJ. Principled missing data methods for researchers. *SpringerPlus*. 2013;2. doi:10.1186/2193-1801-2-222

163. Dormann CF, Elith J, Bacher S, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*. 2013;36(1):27-46. doi:10.1111/j.1600-0587.2012.07348.x

164. Spearman's Rank-Order Correlation - A guide to when to use it, what it does and what the assumptions are. Accessed December 10, 2020. https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

165. Skelly AC, Dettori JR, Brodt ED. Assessing bias: the importance of considering confounding. *Evid-Based Spine-Care J*. 2012;3(1):9-12. doi:10.1055/s-0031-1298595

166. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection Bias and Information Bias in Clinical Research. *Nephron Clin Pract*. 2010;115(2):c94-c99. doi:10.1159/000312871

167. Marino MJ. How often should we expect to be wrong? Statistical power, P values, and the expected prevalence of false discoveries. *Biochem Pharmacol*. 2018;151:226-233. doi:10.1016/j.bcp.2017.12.011

168. Armstrong R. When to use the Bonferroni correction - PubMed. *Ophthalmic Physiol Opt*. 2014;34:502-508. doi:https://doi.org/10.1111/opo.12131

169. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567(7748):305-307. doi:10.1038/d41586-019-00857-9

170. Lovell DP. Biological Importance and Statistical Significance. *J Agric Food Chem*. 2013;61(35):8340-8348. doi:10.1021/jf401124y

171. Alfò M, Maruotti A. Two-part regression models for longitudinal zero-inflated count data. *Can J Stat*. 2010;38(2):197-216. doi:10.1002/cjs.10056

172. Hu M-C, Pavlicova M, Nunes EV. Zero-Inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. *Am J Drug Alcohol Abuse*. 2011;37(5):367-375. doi:10.3109/00952990.2011.597280

173. Desmarais BA, Harden JJ. Testing for Zero Inflation in Count Models: Bias Correction for the Vuong Test. *Stata J*. 2013;13(4):810-835. doi:10.1177/1536867X1301300408

174. Rich J, Neely G, Paniello R, Voelka C, Nussenbaum B, Wang E. A Practical Guide to Understanding Kaplan-Meier Curves. *Otolaryngol Head Neck Surg*. 2010;143(3):331-336. doi:10.1016/j.otohns.2010.05.007

175. Ranganathan P, Pramesh CS. Censoring in survival analysis: Potential for bias. *Perspect Clin Res*. 2012;3(1):40. doi:10.4103/2229-3485.92307

176. Prinja S, Gupta N, Verma R. Censoring in Clinical Trials: Review of Survival Analysis Techniques. *Indian J Community Med Off Publ Indian Assoc Prev Soc Med*. 2010;35(2):217-221. doi:10.4103/0970-0218.66859

177. Xue Y, Schifano ED. Diagnostics for the Cox model. *Commun Stat Appl Methods*. 2017;24(6):583-604. doi:10.29220/CSAM.2017.24.6.583

178. Amini Z. Log-linearity for Cox's regression model. undefined. Published 2015. Accessed September 28, 2020. /paper/Log-linearity-for-Cox%27s-regression-model-Amini/7eb27a3c3316bc3097c6e04b31e8f9a5e92cf138

179. Chappell R. Competing Risk Analyses: How Are They Different and Why Should You Care? *Clin Cancer Res*. 2012;18(8):2127-2129. doi:10.1158/1078-0432.CCR-12-0455

180. Dignam JJ, Zhang Q, Kocherginsky MN. The use and interpretation of competing risks regression models. *Clin Cancer Res*. 2012;18(8):2301-2308. doi:10.1158/1078-0432.CCR-11-2097.The

181. Pintilie M. An Introduction to Competing Risks Analysis. *Rev Esp Cardiol Engl Ed*. 2011;64(7):599-605. doi:10.1016/j.rec.2011.03.016

182. Rodríguez G. Competing Risks. Published online 2005. Accessed December 16, 2020. https://data.princeton.edu/pop509/competingrisks.pdf

183. Completing a medical certificate of cause of death (MCCD). GOV.UK. Accessed September 29, 2020. https://www.gov.uk/government/publications/guidance-notes-for-completing-a-medical-certificate-of-cause-of-death

184. Ederer F, Geisser MS, Mongin SJ, Church TR, Mandel JS. Colorectal cancer deaths as determined by expert committee and from death certificate: A comparison. The Minnesota study. *J Clin Epidemiol*. 1999;52(5):447-452. doi:10.1016/S0895-4356(99)00016-5

185. Begg CB, Schrag D. Attribution of deaths following cancer treatment. *J Natl Cancer Inst*. 2002;94(14):1044-1045. doi:10.1093/jnci/94.14.1044

186. Welch HG, Black WC. Are deaths within 1 month of cancer-directed surgery attributed to cancer? *J Natl Cancer Inst*. 2002;94(14):1066-1070. doi:10.1093/jnci/94.14.1066

187. Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. *Biostatistics*. 2014;15(4):757-773. doi:10.1093/biostatistics/kxu010

188. Breiman L. Random Forests. *Mach Learn*. 2001;(45):5-32.

189. Probst P, Wright MN, Boulesteix A-L. Hyperparameters and tuning strategies for random forest. *WIREs Data Min Knowl Discov*. 2019;9(3):e1301. doi:10.1002/widm.1301

190. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*. 2007;8. doi:10.1186/1471-2105-8-25

191. Guestrin MTR Sameer Singh, Carlos. Local Interpretable Model-Agnostic Explanations (LIME): An Introduction. O'Reilly Media. Published August 12, 2016. Accessed September 28, 2020. https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

192. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. *JAMA J Am Med Assoc*. Published online 1982. doi:10.1001/jama.1982.03320430047030

193. Han X, Zhang Y, Shao Y. On comparing two correlated C indices with censored survival data. *Stat Med*. 2017;36(25):4041-4049. doi:10.1002/sim.7414

194. Heagerty PJ, Lumley T, Pepe MS. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics*. 2000;56(2):337-344. doi:10.1111/j.0006-341X.2000.00337.x

195. Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med Res Methodol*. 2017;17(1):53. doi:10.1186/s12874-017-0332-6

196. Kronek L-P, Reddy A. Logical analysis of survival data: prognostic survival models by detecting high-degree interactions in right-censored data. *Bioinformatics*. 2008;24(16):i248-i253. doi:10.1093/bioinformatics/btn265

197. Haider H, Hoehn B, Davis S, Greiner R. Effective Ways to Build and Evaluate Individual Survival Distributions. *ArXiv181111347 Cs Stat*. Published online November 27, 2018. Accessed October 1, 2020. http://arxiv.org/abs/1811.11347

198. Hemingway H, Croft P, Perel P, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ*. 2013;346:e5595. doi:10.1136/bmj.e5595

199. Riley RD, Hayden JA, Steyerberg EW, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med*. 2013;10(2):e1001380. doi:10.1371/journal.pmed.1001380

200. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381. doi:10.1371/journal.pmed.1001381

201. Hingorani AD, Windt DA van der, Riley RD, et al. Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ*. 2013;346. doi:10.1136/bmj.e5793

202. Fundamental Rights Agency. Handbook of European Data Protection Law. Published 2018. Accessed October 7, 2019. https://fra.europa.eu/en/publication/2018/handbook-european-data-protection-law

203. Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland - Office for National Statistics. Accessed October 5, 2020. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland

204. Mellitus D. Definition , Diagnosis and Classification of Diabetes Mellitus and its Complications Part 1 : Diagnosis and Classification of. *World Health*. Published online 1999. doi:10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S

205. English Housing Survey data on new households and recent movers. GOV.UK. Accessed October 7, 2020. https://www.gov.uk/government/statistical-data-sets/new-households-and-recent-movers

206. Aunan JR, Cho WC, Søreide K. The Biology of Aging and Cancer: A Brief Overview of Shared and Divergent Molecular Hallmarks. *Aging Dis*. 2017;8(5):628-642. doi:10.14336/AD.2017.0103

207. Thakkar JP, Villano JL, McCarthy BJ. Age-Specific Cancer Incidence Rates Increase Through the Oldest Age Groups. *Am J Med Sci*. 2014;348(1):65-70. doi:10.1097/MAJ.0000000000000281

208. George RL, McGwin GJ, Metzger J, Chaudry IH, Rue LWI. The Association between Gender and Mortality among Trauma Patients as Modified by Age. *J Trauma Acute Care Surg*. 2003;54(3):464-471. doi:10.1097/01.TA.0000051939.95039.E6

209. Lee C, Joseph L, Colosimo A, Dasgupta K. Mortality in diabetes compared with previous cardiovascular disease: A gender-specific meta-analysis. *Diabetes Metab*. 2012;38(5):420-427. doi:10.1016/j.diabet.2012.04.002

210. MacIntyre K, Stewart S, Capewell S, et al. Gender and survival: a population-based study of 201,114 men and women following a first acute myocardial infarction. *J Am Coll Cardiol*. 2001;38(3):729-735. doi:10.1016/S0735-1097(01)01465-6

211. Vaartjes I, Reitsma JB, Sijl MB, Bots ML. Gender Differences in Mortality after Hospital Admission for Stroke. *Cerebrovasc Dis*. 2009;28(6):564-571. doi:10.1159/000247600

212. Wang Y, Wang B, Yan S, et al. Type 2 diabetes and gender differences in liver cancer by considering different confounding factors: a meta-analysis of cohort studies. *Ann Epidemiol*. 2016;26(11):764-772.

213. Collins PF, Stratton RJ, Kurukulaaratchy RJ, Elia M. Influence of deprivation on health care use, health care costs, and mortality in COPD. *Int J Chron Obstruct Pulmon Dis*. 2018;13:1289-1296. doi:10.2147/COPD.S157594

214. McCormick J, Chen R. Impact of socioeconomic deprivation on mortality in people with haemorrhagic stroke: a population-based cohort study. *Postgrad Med J*. 2016;92(1091):501-505. doi:10.1136/postgradmedj-2015-133663

215. McCartney G, Popham F, Katikireddi SV, Walsh D, Schofield L. How do trends in mortality inequalities by deprivation and education in Scotland and England & Wales compare? A repeat cross-sectional study. *BMJ Open*. 2017;7(7):e017590. doi:10.1136/bmjopen-2017-017590

216. Graaf MA de, Jager KJ, Zoccali C, Dekker FW. Matching, an Appealing Method to Avoid Confounding? *Nephron Clin Pract*. 2011;118(4):c315-c318. doi:10.1159/000323136

217. Nguyen CD, Carlin JB, Lee KJ. Model checking in multiple imputation: an overview and case study. *Emerg Themes Epidemiol*. 2017;14. doi:10.1186/s12982-017-0062-6

218. Pham A, Cummings M, Lindeman C, Drummond N, Williamson T. Recognizing misclassification bias in research and medical practice. *Fam Pract*. 2019;36(6):804-807. doi:10.1093/fampra/cmy130

219. Bhangu A, Nepogodiev D, Taylor C, Durkin N, Patel R. Accuracy of clinical coding from 1210 appendicectomies in a British district general hospital. *Int J Surg*. 2012;10(3):144-147. doi:10.1016/j.ijsu.2012.01.007

220. Burns EM, Rigby E, Mamidanna R, et al. Systematic review of discharge coding accuracy. *J Public Health*. 2012;34(1):138-148. doi:10.1093/pubmed/fdr054

221. Daultrey H, Gooday E, Dhatariya K. Increased length of inpatient stay and poor clinical coding: audit of patients with diabetes. *JRSM Short Rep*. 2011;2(11):1-6. doi:10.1258/shorts.2011.011100

222. Dixon J, Sanderson C, Elliott P, Walls P, Jones J, Petticrew M. Assessment of the reproducibility of clinical coding in routinely collected hospital activity data: A study in two hospitals. *J Public Health U K*. 1998;20(1):63-69. doi:10.1093/oxfordjournals.pubmed.a024721

223. Connor AE, Visvanathan K, Baumgartner KB, et al. Ethnic differences in the relationships between diabetes, early age adiposity and mortality among breast cancer survivors: the Breast Cancer Health Disparities Study. *Breast Cancer Res Treat*. 2016;157(1):167-178.

224. Quiñones AR, Botoseneanu A, Markwardt S, et al. Racial/ethnic differences in multimorbidity development and chronic disease accumulation for middle-aged adults. *PLoS ONE*. 2019;14(6). doi:10.1371/journal.pone.0218462

225. Amshoff Y, Maskarinec G, Shvetsov YB, et al. Type 2 diabetes and colorectal cancer survival: The multiethnic cohort. *Int J Cancer*. 2018;143(2):263-268.

226. Johnson-Lawrence V, Zajacova A, Sneed R. Education, race/ethnicity, and multimorbidity among adults aged 30–64 in the National Health Interview Survey. *SSM - Popul Health*. 2017;3:366-372. doi:10.1016/j.ssmph.2017.03.007

227. Glashan RW, Cartwright RA. Occupational Bladder Cancer and Cigarette Smoking in West Yorkshire. *Br J Urol*. 1981;53(6):602-604. doi:10.1111/j.1464-410X.1981.tb03270.x

228. McKinney PA, Roberts BE, O'Brien C, et al. Chronic Myeloid Leukaemia in Yorkshire: A Case Control Study. *Acta Haematol*. 1990;83(1):35-38. doi:10.1159/000205160

229. Sorahan T, Harrington JM. Lung cancer in Yorkshire chrome platers, 1972–97. *Occup Environ Med*. 2000;57(6):385-389. doi:10.1136/oem.57.6.385

230. Marti A, Moreno-Aliaga MJ, Hebebrand J, Martínez JA. Genes, lifestyles and obesity. *Int J Obes*. 2004;28(3):S29-S36. doi:10.1038/sj.ijo.0802808

231. Quality and Outcomes Framework, 2019-20. NHS Digital. Accessed October 5, 2020. https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data/2019-20

232. Chen Z, Wu Z, Ning W. Advances in Molecular Mechanisms and Treatment of Radiation-Induced Pulmonary Fibrosis. *Transl Oncol*. 2018;12(1):162-169. doi:10.1016/j.tranon.2018.09.009

233. Lowenthal RM, Eaton K. Toxicity of chemotherapy. *Hematol Clin*. 1996;10(4):967-990. doi:10.1016/S0889-8588(05)70378-6

234. Epidemiology and pathophysiology of cancer-associated thrombosis | British Journal of Cancer. Accessed October 7, 2020. https://www.nature.com/articles/6605599

235. Cantiello F, Cicione A, Salonia A, et al. Association between metabolic syndrome, obesity, diabetes mellitus and oncological outcomes of bladder cancer: a systematic review. *Int J Urol*. 2015;22(1):22-32.

236. Meyer CC, Calis KA, Burke LB, Walawander CA, Grasela TH. Symptomatic Cardiotoxicity Associated with 5-Fluorouracil. *Pharmacother J Hum Pharmacol Drug Ther*. 1997;17(4):729-736. doi:10.1002/j.1875-9114.1997.tb03748.x

237. Mohan N, Jiang J, Dokmanovic M, Wu WJ. Trastuzumab-mediated cardiotoxicity: current understanding, challenges, and frontiers. *Antib Ther*. 2018;1(1):13-17. doi:10.1093/abt/tby003

238. Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. *Int J Ayurveda Res*. 2010;1(4):274-278. doi:10.4103/0974-7788.76794

239. Sarfati D, Koczwara B, Jackson C. The impact of comorbidity on cancer and its treatments. *CA Cancer J Clin*. 2016;66(4):337-350. doi:10.3322/caac.21342.

240. Extermann M. Interaction between comorbidity and cancer. *Cancer Control*. 2007;14(1):13-22. doi:10.1177/107327480701400103

241. Devane C. *Living after Cancer: Median Cancer Survival Times*. Macmillan Cancer Support; 2011. Accessed October 8, 2020. https://www.macmillan.org.uk/documents/aboutus/newsroom/livingaftercancermediancancersurvivaltimes.pdf

242. Cancer survival in England - adults diagnosed - Office for National Statistics. Accessed January 13, 2021. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/cancersurvivalratescancersurvivalinenglandadultsdiagnosed

243. Tang DH, Chang SS. Management of carcinoma in situ of the bladder: best practice and recent developments. *Ther Adv Urol*. 2015;7(6):351-364. doi:10.1177/1756287215599694

244. *Clinical Radiotherapy Guidelines for Carcinoma of the Bladder.* London Cancer; 2013. Accessed October 8, 2020. http://www.londoncancer.org/media/84322/london-cancer-bladder-radiotherapy-guidelines-2013-v1-0.pdf

245. Shim M, Kelly B, Hornik R. Cancer Information Scanning and Seeking Behavior is Associated with Knowledge, Lifestyle Choices, and Screening. *J Health Commun*. 2006;11(sup001):157-172. doi:10.1080/10810730600637475

246. Møller H, Coupland VH, Tataru D, et al. Geographical variations in the use of cancer treatments are associated with survival of lung cancer patients. *Thorax*. 2018;73(6):530-537. doi:10.1136/thoraxjnl-2017-210710

247. Gilbert SM, Pow-Sang JM, Xiao H. Geographical factors associated with health disparities in prostate cancer. *Cancer Control*. 2016;23(4):401-408. doi:10.1177/107327481602300411

248. Kavecansky J. Beyond Checkpoint Inhibitors: The Next Generation of Immunotherapy in Oncology. *Am J Hematol Oncol*. 2017;13(2). Accessed October 8, 2020. https://www.gotoper.com/publications/ajho/2017/2017feb/beyond-checkpoint-inhibitors-the-next-generation-of-immunotherapy-in-oncology

249. Postmus PE, Kerr KM, Oudkerk M, et al. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2017;28:iv1-iv21. doi:10.1093/annonc/mdx222

250. Klaiber U, Probst P, Büchler MW, Hackert T. Pylorus preservation pancreatectomy or not. *Transl Gastroenterol Hepatol*. 2017;2. doi:10.21037/tgh.2017.11.15

251. Goldsmith C, Plowman PN, Green MM, Dale RG, Price PM. Stereotactic ablative radiotherapy (SABR) as primary, adjuvant, consolidation and re-treatment option in pancreatic cancer: scope for dose escalation and lessons for toxicity. *Radiat Oncol*. 2018;13(1):204. doi:10.1186/s13014-018-1138-3

252. Tantiworawit A, Rattanathammethee T, Chai-Adisaksopha C, Rattarittamrong E, Norasetthada L. Outcomes of adult acute lymphoblastic leukemia in the era of pediatric-inspired regimens: a single-center experience. *Int J Hematol*. 2019;110(3):295-305. doi:10.1007/s12185-019-02678-y

253. Boulos DL, Groome PA, Brundage MD, et al. Predictive validity of five comorbidity indices in prostate carcinoma patients treated with curative intent. *Cancer*. 2006;106(8):1804-1814. doi:10.1002/cncr.21813

254. Lange JR, Kang S, Balch CM. Melanoma in the older patient: Measuring frailty as an index of survival. *Ann Surg Oncol*. 2011;18(13):3531-3532. doi:10.1245/s10434-011-2015-6

255. Caspi RR. Immunotherapy of autoimmunity and cancer: the penalty for success. *Nat Rev Immunol*. 2008;8(12):970-976. doi:10.1038/nri2438

256. Dooley MJ, Poole SG, Rischin D. Dosing of cytotoxic chemotherapy: impact of renal function estimates on dose. *Ann Oncol*. 2013;24(11):2746-2752. doi:10.1093/annonc/mdt300

257. Eklund JW, Trifilio S, Mulcahy MF. Chemotherapy dosing in the setting of liver dysfunction. *Oncol Williston Park N*. 2005;19(8):1057-1063; discussion 1063-1064, 1069.

258. Lavelle K, Todd C, Moran A, Howell A, Bundred N, Campbell M. Non-standard management of breast cancer increases with age in the UK: a population based cohort of women X65 years. *Br J Cancer*. 2007;96:1197-1203. doi:10.1038/sj.bjc.6603709

259. Mounce LTA, Price S, Valderas JM, Hamilton W. Comorbid conditions delay diagnosis of colorectal cancer: a cohort study using electronic primary care records. *Br J Cancer*. 2017;116(12):1536-1543. doi:10.1038/bjc.2017.127

260. Olesen F, Hansen RP, Vedsted P. Delay in diagnosis: the experience in Denmark. *Br J Cancer*. Published online 2009. doi:10.1038/sj.bjc.6605383

261. Lai S, Chen P, Liao K, Muo C, Lin C. Risk of Hepatocellular Carcinoma in Diabetic Patients and Risk Reduction Associated With Anti-Diabetic Therapy : A Population-Based Cohort Study. *Am J Gastroenterol*. 2011;107(1):46-52. doi:10.1038/ajg.2011.384

262. Schütte K, Bornschein J, Malfertheiner P. Hepatocellular carcinoma-epidemiological trends and risk factors. *Dig Dis*. 2009;27(2):80-92. doi:10.1159/000218339

263. Davila JA, Morgan RO, Shaib Y, Mcglynn KA. Diabetes increases the risk of hepatocellular carcinoma in the United States: a population based case control study. *Liver*. 2005;54(4):533-539. doi:10.1136/gut.2004.052167

264. Li J, Han T, Xu L, Luan X. Diabetes mellitus and the risk of cholangiocarcinoma: An updated meta-analysis. *Przeglad Gastroenterol*. 2015;10(2):108-117. doi:10.5114/pg.2015.49004

265. Tseng CH. Obesity paradox: differential effects on cancer and noncancer mortality in patients with type 2 diabetes mellitus. *Atherosclerosis*. 2013;226(1):186-192.

266. Thorat MA, Cuzick J. Role of aspirin in cancer prevention. *Curr Oncol Rep*. 2013;15(6):533-540. doi:10.1007/s11912-013-0351-3

267. Patrignani P, Patrono C. Aspirin, platelet inhibition and cancer prevention. *Platelets*. 2018;29(8):779-785. doi:10.1080/09537104.2018.1492105

268. Gronich N, Rennert G. Beyond aspirin - Cancer prevention with statins, metformin and bisphosphonates. *Nat Rev Clin Oncol*. 2013;10(11):625-642. doi:10.1038/nrclinonc.2013.169

269. Zanders MM, van Herk-Sukel MP, Vissers PA, Herings RM, Haak HR, van de Poll-Franse LV. Are metformin, statin and aspirin use still associated with overall mortality among colorectal cancer patients with diabetes if adjusted for one another? *Br J Cancer*. 2015;113(3):403-410.

270. A. D, M. P, P. G, et al. Metformin and cancer risk in diabetic patients: A systematic review and meta-analysis. *Cancer Prev Res (Phila Pa)*. 2010;3(11):1451-1461. doi:10.1158/1940-6207.CAPR-10-0157

271. Chong RW, Vasudevan V, Zuber J, Solomon SS. Metformin Has a Positive Therapeutic Effect on Prostate Cancer in Patients With Type 2 Diabetes Mellitus. *Am J Med Sci*. Published online 2016. doi:10.1016/j.amjms.2016.01.013

272. Zhou XL, Xue WH, Ding XF, et al. Association between metformin and the risk of gastric cancer in patients with type 2 diabetes mellitus: a meta-analysis of cohort studies. *Oncotarget*. 2017;8(33):55622-55631. doi:10.18632/oncotarget.16973

273. Alimova IN, Liu B, Fan Z, et al. Metformin inhibits breast cancer cell growth, colony formation and induces cell cycle arrest in vitro. *Cell Cycle*. 2009;8(6):909-915. doi:10.4161/cc.8.6.7933

274. Sahra I Ben, Laurent K, Loubat A, et al. The antidiabetic drug metformin exerts an antitumoral effect in vitro and in vivo through a decrease of cyclin D1 level. *Oncogene*. 2008;27(25):3576-3586. doi:10.1038/sj.onc.1211024

275. Jansen RJ, Alexander BH, Anderson KE, Church TR. Quantifying lead-time bias in risk factor studies of cancer through simulation. *Ann Epidemiol*. 2013;23(11):735-741.e1. doi:10.1016/j.annepidem.2013.07.021

276. Ravindrarajah R, Hazra NC, Hamada S, et al. Systolic Blood Pressure Trajectory, Frailty, and All-Cause Mortality >80 Years of Age. *Circulation*. 2017;135(24):2357-2368. doi:10.1161/CIRCULATIONAHA.116.026687

277. Flegal KM, Graubard BI, Williamson DF, Cooper RS. Reverse Causation and Illness-related Weight Loss in Observational Studies of Body Weight and Mortality. *Am J Epidemiol*. 2011;173(1):1-9. doi:10.1093/aje/kwq341

278. Edwards BK, Noone AM, Mariotto AB, et al. Annual Report to the Nation on the status of cancer, 1975-2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. *Cancer*. 2014;120(9):1290-1314. doi:10.1002/cncr.28509

279. Kendal WS. Dying with cancer: The influence of age, comorbidity, and cancer site. *Cancer*. 2008;112(6):1354-1362. doi:10.1002/cncr.23315

280. Janssen-Heijnen MLG, Houterman S, Lemmens VEPP, Louwman MWJ, Maas HAAM, Coebergh JWW. Prognostic impact of increasing age and co-morbidity in cancer patients: A population-based approach. *Crit Rev Oncol Hematol*. 2005;55(3):231-240. doi:10.1016/j.critrevonc.2005.04.008

281. Piccirillo JF, Tierney RM, Costas I, Grove L, Spitznagel EL. Prognostic importance of comorbidity in a hospital-based cancer registry. *JAMA*. 2004;291(20):2441-2447. doi:10.1001/jama.291.20.2441

282. Read WL, Tierney RM, Page NC, et al. Differential prognostic impact of comorbidity. *J Clin Oncol*. 2004;22(15):3099-3103. doi:10.1200/JCO.2004.08.040

283. Archie JP. Mathematic coupling of data: a common source of error. *Ann Surg*. 1981;193(3):296-303. doi:10.1097/00000658-198103000-00008

284. Tu Y-K, Maddick IH, Griffiths GS, Gilthorpe MS. Mathematical coupling can undermine the statistical assessment of clinical research: illustration from the treatment of guided tissue regeneration. *J Dent*. 2004;32(2):133-142. doi:10.1016/j.jdent.2003.10.001

285. Pearson K. Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc R Soc Lond*. 1897;60(359-367):489-498. doi:10.1098/rspl.1896.0076

286. Nouraei SAR, Hudovsky A, Frampton AE, et al. A study of clinical coding accuracy in surgery: Implications for the use of administrative big data for outcomes management. *Ann Surg*. 2015;261(6):1096-1107. doi:10.1097/SLA.0000000000000851

287. Santos S, Murphy G, Baxter K, Robinson KM. Organisational factors affecting the quality of hospital clinical coding. *Health Inf Manag J*. 2008;37(1):25-37. doi:10.1177/183335830803700103

288. Fox J, Weisberg S. *Cox Proportional-Hazards Regression for Survival Data in R*. Statistical tools for high-throughput data analysis; 2018. Accessed January 12, 2020. https://socialsciences.mcmaster.ca/jfox/Books/Companion-1E/appendix-cox-regression.pdf

289. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons; 2005.

290. Lin M, Lucas HC, Shmueli G. Research Commentary—Too Big to Fail: Large Samples and the p-Value Problem. *Inf Syst Res*. 2013;24(4):906-917. doi:10.1287/isre.2013.0480

291. Owrang M, Copeland LR, Ricks-Santi JL, et al. Breast Cancer Prognosis for Young Patients. *In Vivo*. 2017;31(4):661-668. doi:10.21873/invivo.11109

292. Rohrer JM. Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Adv Methods Pract Psychol Sci*. 2018;1(1):27-42. doi:10.1177/2515245917745629

293. Neufeld E, Kristtorn S. Does non-correlation imply non-causation? *Intern J Approx Reason*. 2006;46(2007):257-273.

294. Textor J, van der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: The R package "dagitty." *Int J Epidemiol*. 2016;45(6):1887-1894. doi:10.1093/ije/dyw341

295. Weng HY, Hsueh YH, Messam LLM, Hertz-Picciotto I. Methods of covariate selection: Directed acyclic graphs and the change-in-estimate procedure. *Am J Epidemiol*. 2009;169(10):1182-1190. doi:10.1093/aje/kwp035

296. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol*. 2018;47(1):226-235. doi:10.1093/ije/dyx206

297. Fewell Z, Davey Smith G, Sterne JAC. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am J Epidemiol*. 2007;166(6):646-655. doi:10.1093/aje/kwm165

298. Anders H-J, Huber TB, Isermann B, Schiffer M. CKD in diabetes: diabetic kidney disease versus nondiabetic kidney disease. *Nat Rev Nephrol*. 2018;14(6):361-377. doi:10.1038/s41581-018-0001-y

299. Hwang JL, Weiss RE. Steroid-induced diabetes: a clinical and molecular approach to understanding and treatment. *Diabetes Metab Res Rev*. 2014;30(2):96-102. doi:10.1002/dmrr.2486

300. Babyak MA. Understanding confounding and mediation. *Evid Based Ment Health*. 2009;12(3):68-71. doi:10.1136/ebmh.12.3.68

301. Dahabreh IJ, Kent DM. Index event bias: an explanation for the paradoxes of recurrence risk research. *JAMA J Am Med Assoc*. 2011;305(8):822-823. doi:10.1001/jama.2011.163

302. Hall WH, Jani AB, Ryu JK, Narayan S, Vijayakumar S. The impact of age and comorbidity on survival outcomes and treatment patterns in prostate cancer. *Prostate Cancer Prostatic Dis*. 2005;8(1):22-30. doi:10.1038/sj.pcan.4500772

303. Rieker R, Hammer E, Eisele R, Schmid E, Hogel J. The impact of comorbidity on the overall survival and the cause of death in patients after colorectal cancer resection | SpringerLink. *Langenbecks Arch Surg*. 2002;387(2):72-76.

304. Earle C, Tsai J, Gelber R, Weinstein M, Neumann P, Weeks J. Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrumental variable and propensity analysis. *J Clin Oncol Off J Am Soc Clin Oncol*. 2001;19(4):1064-1070. doi:10.1200/jco.2001.19.4.1064

305. Lee L, Cheung WY, Atkinson E, Krzyzanowska MK. Impact of comorbidity on chemotherapy use and outcomes in solid tumors: a systematic review. *J Clin Oncol Off J Am Soc Clin Oncol*. 2011;29(1):106-117. doi:10.1200/JCO.2010.31.3049

306. Geifman N, Butte A. Do cancer clinica trial populations truly represent cancer patients? *Pac Symp Biocomput*. 2016;21:309-320.

307. Vaeth PA, Satariano WA, Ragland DR. Limiting comorbid conditions and breast cancer stage at diagnosis. *J Gerontol A Biol Sci Med Sci*. 2000;55(10):M593-600. doi:10.1093/gerona/55.10.m593

308. Zafar SY, Abernethy AP, Abbott DH, et al. Comorbidity, age, race and stage at diagnosis in colorectal cancer: a retrospective, parallel analysis of two health systems. *BMC Cancer*. 2008;8(1):345. doi:10.1186/1471-2407-8-345

309. Walter L, Lindquist K, Nugent S, et al. Impact of Age and Comorbidity on Colorectal Cancer Screening Among Older Veterans. *Ann Intern Med*. 2013;150(7):465-473.

310. Yasmeen S, Chlebowski R, Xing G, Morris C, Romano P. Severity of comorbid conditions and early-stage breast cancer therapy: linked SEER-medicare data from 1993 to 2005. *Cancer Med*. 2013;2(4):526-536. doi:10.1002/cam4.66

311. Fleming ST, McDavid K, Pearce K, Pavlov D. Comorbidities and the Risk of Late-Stage Prostate Cancer. *Sci World J*. 2006;6:2460-2470. doi:10.1100/tsw.2006.383

312. Handforth C, Clegg A, Young C, et al. The prevalence and outcomes of frailty in older cancer patients: a systematic review. *Ann Oncol*. 2015;26(6):1091-1101. doi:10.1093/annonc/mdu540

313. Creager M, Luscher T. Diabetes and Vascular Disease. *Circulation*. 2003;108(12):1527-1532.

314. Hassan M, Curley S, Li D, et al. Association of diabetes duration and diabetes treatment with the risk of hepatocellular carcinoma. *Cancer*. 2010;116(8):1938-1946. doi:10.1016/j.cgh.2008.07.016.Cytokeratin

315. Klip A, Marette A, Dimitrakoudis D, et al. Effect of diabetes on glucoregulation. From glucose transporters to glucose metabolism in vivo. *Diabetes Care*. 1992;15(11):1747-1766. doi:10.2337/diacare.15.11.1747

316. Pinter M, Trauner M, Peck-Radosavljevic M, Sieghart W. Cancer and liver cirrhosis: implications on prognosis and management. *ESMO Open*. 2016;1(2). Accessed October 22, 2020. https://esmoopen.bmj.com/content/1/2/e000042

317. Pang Y, Kartsonaki C, Turnbull I, et al. Diabetes, Plasma Glucose, and Incidence of Fatty Liver, Cirrhosis, and Liver Cancer: A Prospective Study of 0.5 Million People. *Hepatology*. 2018;68(4):1308-1318.

318. Durham AL, Adcock IM. The relationship between COPD and lung cancer. *Lung Cancer*. 2015;90(2):121-127. doi:10.1016/j.lungcan.2015.08.017

319. Young RP, Hopkins RJ. Link between COPD and lung cancer. *Respir Med*. 2010;104(5):758-759. doi:10.1016/j.rmed.2009.11.025

320. Sekine Y, Katsura H, Koh E, Hiroshima K, Fujisawa T. Early detection of COPD is important for lung cancer surveillance. *Eur Respir J*. 2012;39(5):1230-1240. doi:10.1183/09031936.00126011

321. Van Gestel YRBM, Hoeks SE, Sin DD, et al. COPD and cancer mortality: The influence of statins. *Thorax*. 2009;64(11):963-967. doi:10.1136/thx.2009.116731

322. Parimon T, Chien JW, Bryson CL, McDonell MB, Udris EM, Au DH. Inhaled corticosteroids and risk of lung cancer among patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2007;175(7):712-719. doi:10.1164/rccm.200608-1125OC

323. Brusselle GG, Joos GF, Bracke KR. New insights into the immunology of chronic obstructive pulmonary disease. *The Lancet*. 2011;378(9795):1015-1026. doi:10.1016/S0140-6736(11)60988-4

324. Harman D. Free radical theory of aging: an update: increasing the functional life span. *Ann N Y Acad Sci*. 2006;1067:10-21. doi:10.1196/annals.1354.003

325. Kirkham PA, Caramori G, Casolari P, et al. Oxidative stress-induced antibodies to carbonyl-modified protein correlate with severity of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2011;184(7):796-802. doi:10.1164/rccm.201010-1605OC

326. Adcock I, Camori G, Barnes P. Chronic Obstructive Pulmonary Disease and Lung Cancer: New Molecular Insights. *Respiration*. 2011;81(4):265-284.

327. Savale L, Chaouat A, Bastuji-Garin S, et al. Shortened telomeres in circulating leukocytes of patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2009;179(7):566-571. doi:10.1164/rccm.200809-1398OC

328. Joost O, Wilk JB, Cupples LA, et al. Genetic loci influencing lung function: a genome-wide scan in the Framingham Study. *Am J Respir Crit Care Med*. 2002;165(6):795-799. doi:10.1164/ajrccm.165.6.2102057

329. Schwartz AG, Ruckdeschel JC. Familial Lung Cancer. *Am J Respir Crit Care Med*. 2006;173(1):16-22. doi:10.1164/rccm.200502-235PP

330. Yang IA, Holloway JW, Fong KM. Genetic susceptibility to lung cancer and co-morbidities. *J Thorac Dis*. 2013;5(Suppl 5):S454-S462. doi:10.3978/j.issn.2072-1439.2013.08.06

331. Franco R, Schoneveld O, Georgakilas AG, Panayiotidis MI. Oxidative stress, DNA methylation and carcinogenesis. *Cancer Lett*. 2008;266(1):6-11. doi:10.1016/j.canlet.2008.02.026

332. Kabesch M, Adcock IM. Epigenetics in asthma and COPD. *Biochimie*. 2012;94(11):2231-2241. doi:10.1016/j.biochi.2012.07.017

333. Tessema M, Yingling C, Picchi M, Wu G, Liu Y. Epigenetic Repression of CCDC37 and MAP1B Links Chronic Obstructive Pulmonary Disease to Lung Cancer. *J Thorac Oncol*. 2015;10(8):1181-1188.

334. Sacco RL, Kasner SE, Broderick JP, et al. An Updated Definition of Stroke for the 21st Century. *Stroke*. Published online July 2013. doi:10.1161/STR.0b013e318296aeca

335. Mandybur T. Intracranial hemorrhage caused by metastatic tumors | Neurology. *Neurology*. 1977;27(7). Accessed October 21, 2020. https://n.neurology.org/content/27/7/650

336. Velander AJ, DeAngelis LM, Navi BB. Intracranial hemorrhage in patients with cancer. *Curr Atheroscler Rep*. 2012;14(4):373-381. doi:10.1007/s11883-012-0250-3

337. Lefkovits J, Davis S, Rossiter S, et al. Acute stroke outcome: effects of stroke type and risk factors - Lefkovits - 1992 - Australian and New Zealand Journal of Medicine - Wiley Online Library. *Aust N Z J Med*. 1992;22(1). Accessed October 21, 2020. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1445-5994.1992.tb01705.x?casa_token=h-LBjwp5398AAAAA:OCTIri-rF4tGR3dCq86vFS0ss-ci75WHzA52Pxk8QPPIhIoGjKMHh7GKubSTPFE0C4cErXgAtb3lLw

338. Stefan O, Vera N, Otto B, Heinz L, Wolfgang G. Stroke in cancer patients: a risk factor analysis. *J Neurooncol*. 2009;94(2):221-226. doi:10.1007/s11060-009-9818-3

339. Grisold W, Oberndorfer S, Struhal W. Stroke and cancer: a review. *Acta Neurol Scand*. 2009;119(1):1-16. doi:10.1111/j.1600-0404.2008.01059.x

340. King D, Wittenberg R, Patel A, Quayyum Z, Berdunov V, Knapp M. The future incidence, prevalence and costs of stroke in the UK. *Age Ageing*. 2020;49(2):277-282. doi:10.1093/ageing/afz163

341. Kwon H-M, Kang BS, Yoon B-W. Stroke as the first manifestation of concealed cancer. *J Neurol Sci*. 2007;258(1):80-83. doi:10.1016/j.jns.2007.02.035

342. Navi BB, Iadecola C. Ischemic stroke in cancer patients: A review of an underappreciated pathology. *Ann Neurol*. 2018;83(5):873-883. doi:10.1002/ana.25227

343. Banke A, Schou M, Videbæk L, et al. Incidence of cancer in patients with chronic heart failure: A long-term follow-up study. *Eur J Heart Fail*. 2016;18(3):260-266. doi:10.1002/ejhf.472

344. Boffetta P, Malhotra J. Impact of Heart Failure on Cancer Incidence: A Complicated Question ∗. *J Am Coll Cardiol*. 2018;71(14):1511-1512. doi:10.1016/j.jacc.2018.02.015

345. Hasin T, Gerber Y, McNallan SM, et al. Patients with heart failure have an increased risk of incident cancer. *J Am Coll Cardiol*. Published online 2013. doi:10.1016/j.jacc.2013.04.088

346. Meijers WC, De Boer RA. Common risk factors for heart failure and cancer. *Cardiovasc Res*. 2019;115(5):844-853. doi:10.1093/cvr/cvz035

347. Meijers WC, Maglione M, Bakker SJL, et al. Heart failure stimulates tumor growth by circulating factors. *Circulation*. Published online 2018. doi:10.1161/CIRCULATIONAHA.117.030816

348. Malmborg M, Christiansen CB, Schmiegelow MD, Torp-Pedersen C, Gislason G, Schou M. Incidence of new onset cancer in patients with a myocardial infarction – a nationwide cohort study. *BMC Cardiovasc Disord*. 2018;18(1):198. doi:10.1186/s12872-018-0932-z

349. Itzhaki Ben Zadok O, Hasdai D, Gottlieb S, et al. Characteristics and outcomes of patients with cancer presenting with acute myocardial infarction. *Coron Artery Dis*. 2019;30(5):332-338. doi:10.1097/MCA.0000000000000733

350. Koelwyn GJ, Newman AAC, Afonso MS, et al. Myocardial infarction accelerates breast cancer via innate immune reprogramming. *Nat Med*. 2020;26(9):1452-1458. doi:10.1038/s41591-020-0964-7

351. Mela CF, Kopalle PK. The impact of collinearity on regression analysis : the asymmetric eVect of negative and positive correlations. *Appl Econ*. 2002;34:667±677.

352. Vernon SW, Tilley BC, Neale AV, Steinfeldt L. Ethnicity, survival, and delay in seeking treatment for symptoms of breast cancer. *Cancer*. 1985;55(7):1563-1571. doi:10.1002/1097-0142(19850401)55:7<1563::AID-CNCR2820550726>3.0.CO;2-1

353. Hollingsworth SJ, Tang CB, Dialynas M, Barker SGE. Varicose veins: Loss of release of vascular endothelial growth factor and reduced plasma nitric oxide. *Eur J Vasc Endovasc Surg*. 2001;22(6):551-556. doi:10.1053/ejvs.2001.1520

354. Hollingsworth SJ, Powell GL, Barker SGE, Cooper DG. Primary varicose veins: Altered transcription of VEGF and its receptors (KDR, flt-1, soluble flt-1) with sapheno-femoral junction incompetence. *Eur J Vasc Endovasc Surg*. 2004;27(3):259-268. doi:10.1016/j.ejvs.2003.12.015

355. Alevizakos M, Kaltsas S, Syrigos KN. The VEGF pathway in lung cancer. *Cancer Chemother Pharmacol*. 2013;72(6):1169-1181. doi:10.1007/s00280-013-2298-3

356. Bendardaf R, Buhmeida A, Hilska M, et al. VEGF-1 expression in colorectal cancer is associated with disease localization, stage, and long-term disease-specific survival. *Anticancer Res*. 2008;28(6 B):3865-3870.

357. Hurwitz H. Integrating the Anti–VEGF-A Humanized Monoclonal Antibody Bevacizumab with Chemotherapy in Advanced Colorectal Cancer. *Clin Colorectal Cancer*. 2004;4:S62-S68. doi:10.3816/CCC.2004.s.010

358. Wang C-A, Harrell JC, Iwanaga R, Jedlicka P, Ford HL. Vascular endothelial growth factor C promotes breast cancer progression via a novel antioxidant mechanism that involves regulation of superoxide dismutase 3. *Breast Cancer Res*. 2014;16(5):462. doi:10.1186/s13058-014-0462-2

359. Adams J, Carder PJ, Downey S, et al. Vascular endothelial growth factor (VEGF) in breast cancer: comparison of plasma, serum, and tissue VEGF and microvessel density and effects of tamoxifen. *Cancer Res*. 2000;60(11):2898-2905.

360. Lantz B. The large sample size fallacy. *Scand J Caring Sci*. 2013;27(2):487-492. doi:10.1111/j.1471-6712.2012.01052.x

361. Holmberg L, Robinson D, Sandin F, et al. A comparison of prostate cancer survival in England, Norway and Sweden: A population-based study. *Cancer Epidemiol*. Published online 2012. doi:10.1016/j.canep.2011.08.001

362. Lung Cancer Survival Rates | 5-Year Survival Rates for Lung Cancer. Accessed November 6, 2020. https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/survival-rates.html

363. Downing A, Harrison WJ, West RM, Forman D, Gilthorpe MS. Latent class modelling of the association between socioeconomic background and breast cancer survival status at 5 years incorporating stage of disease. *J Epidemiol Community Health*. 2010;64(9):772-776. doi:10.1136/jech.2008.085852

364. Haller B, Schmidt G, Ulm K. Applying competing risks regression models: An overview. *Lifetime Data Anal*. Published online 2012. doi:10.1007/s10985-012-9230-8

365. Berglund A, Wigertz A, Adolfsson J, et al. Impact of comorbidity on management and mortality in women diagnosed with breast cancer. *Breast Cancer Res Treat*. 2012;135(1):281-289. doi:10.1007/s10549-012-2176-4

366. Jørgensen TL, Hallas J, Friis S, Herrstedt J. Comorbidity in elderly cancer patients in relation to overall and cancer-specific mortality. *Br J Cancer*. 2012;106(7):1353-1360. doi:10.1038/bjc.2012.46

367. Land LH, Dalton SO, Jensen M-B, Ewertz M. Impact of comorbidity on mortality: a cohort study of 62,591 Danish women diagnosed with early breast cancer, 1990-2008. *Breast Cancer Res Treat*. 2012;131(3):1013-1020. doi:10.1007/s10549-011-1819-1

368. Govindarajulu US, Spiegelman D, Thurston SW, Ganguli B, Eisen EA. Comparing smoothing techniques in Cox models for exposure–response relationships. *Stat Med*. 2007;26(20):3735-3752. doi:https://doi.org/10.1002/sim.2848

369. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. Vol 103. Springer New York; 2013. doi:10.1007/978-1-4614-7138-7

370. Gauthier J, Wu QV, Gooley TA. Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians. *Bone Marrow Transplant*. 2020;55(4):675-680. doi:10.1038/s41409-019-0679-x

371. Accuracy of Death Certificates and Assessment of Factors for Misclassification of Underlying Cause of Death. *J Epidemiol*. 2016;26(4):191-198. doi:10.2188/jea.JE20150010

372. Messite J, Stellman SD. Accuracy of Death Certificate Completion: The Need for Formalized Physician Training. *JAMA*. 1996;275(10):794-796. doi:10.1001/jama.1996.03530340058030

373. Washirasaksiri C, Raksasagulwong P, Chouriyagune C, Phisalprapa P, Srivanichakorn W. Accuracy and the factors influencing the accuracy of death certificates completed by first-year general practitioners in Thailand. *BMC Health Serv Res*. 2018;18(1):478. doi:10.1186/s12913-018-3289-1

374. Turner EL, Metcalfe C, Donovan JL, et al. Contemporary accuracy of death certificates for coding prostate cancer as a cause of death : Is reliance on death certification good enough? A comparison with blinded review by an independent cause of death evaluation committee. *Br J Cancer*. 2016;115(1):90-94.

375. Sarfati D, Hill S, Blakely T, et al. The effect of comorbidity on the use of adjuvant chemotherapy and survival from colon cancer: a retrospective cohort study. *BMC Cancer*. 2009;9:116. doi:10.1186/1471-2407-9-116

376. van de Poll-Franse LV, Haak HR, Coebergh JWW, Janssen-Heijnen MLG, Lemmens VEPP. Disease-specific mortality among stage I-III colorectal cancer patients with diabetes: a large population-based analysis. *Diabetologia*. 2012;55(8):2163-2172. doi:10.1007/s00125-012-2555-8

377. Seigneurin A, Delafosse P, Trétarre B, et al. Are comorbidities associated with long-term survival of lung cancer? A population-based cohort study from French cancer registries. *BMC Cancer*. 2018;18(1):1091. doi:10.1186/s12885-018-5000-7

378. Miao F, Cai Y-P, Zhang Y-X, Li Y, Zhang Y-T. Risk Prediction of One-Year Mortality in Patients with Cardiac Arrhythmias Using Random Survival Forest. *Comput Math Methods Med*. 2015;215:e303250. doi:https://doi.org/10.1155/2015/303250

379. Dietrich S, Floegel A, Troll M, et al. Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis. *Int J Epidemiol*. 2016;45(5):1406-1420. doi:10.1093/ije/dyw145

380. Oxford RM, Daniel LG. Basic Cross-Validation: Using the “Holdout” Method To Assess the Generalizability of Results. *Res Sch*. 2001;8(1):83-89.

381. Hawkins DM, Basak SC, Mills D. Assessing Model Fit by Cross-Validation. *J Chem Inf Comput Sci*. 2003;43(2):579-586. doi:10.1021/ci025626i

382. Breiman L. Out-of-bag estimation. *Berkley Univ*. Published online 1996. Accessed November 13, 2020. ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps

383. Borg I, Groenen PJF. Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *J Stat Softw*. 2005;14(September).

384. Ishwaran H. Variable importance in binary regression trees and forests. *Electron J Stat*. 2007;1:519-537. doi:10.1214/07-EJS039

385. Greenwell B M. pdp: An R Package for Constructing Partial Dependence Plots. *R J*. 2017;9(1):421. doi:10.32614/RJ-2017-016

386. Miao F, Cai Y-P, Zhang Y-T, Li C-Y. Is Random Survival Forest an Alternative to Cox Proportional Model on Predicting Cardiovascular Disease? In: Lacković I, Vasic D, eds. *6th European Conference of the International Federation for Medical and Biological Engineering*. IFMBE Proceedings. Springer International Publishing; 2015:740-743. doi:10.1007/978-3-319-11128-5_184

387. Kurt Omurlu I, Ture M, Tokatli F. The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer. *Expert Syst Appl*. 2009;36(4):8582-8588. doi:10.1016/j.eswa.2008.10.023

388. Nasejje JB, Mwambi H, Dheda K, Lesosky M. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Med Res Methodol*. 2017;17(1):1-17. doi:10.1186/s12874-017-0383-8

389. Giunchiglia E, Nemchenko A, Schaar M. *RNN-SURV: A Deep Recurrent Model for Survival Analysis*.; 2018.

390. Kelly C, Okada K. Variable interaction measures with random forest classifiers. In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2012:154-157. doi:10.1109/ISBI.2012.6235507

391. Hooker G, Mentch L. Please Stop Permuting Features: An Explanation and Alternatives. *ArXiv190503151 Cs Stat*. Published online May 1, 2019. Accessed November 18, 2020. http://arxiv.org/abs/1905.03151

392. Toloşi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*. 2011;27(14):1986-1994. doi:10.1093/bioinformatics/btr300

393. Hawkins DM. The Problem of Overfitting. *J Chem Inf Comput Sci*. 2004;44(1):1-12. doi:10.1021/ci0342472

394. Rosenberg J, Chia YL, Plevritis S. The effect of age, race, tumor size, tumor grade, and disease stage on invasive ductal breast cancer survival in the U.S. SEER database. *Breast Cancer Res Treat*. 2005;89(1):47-54. doi:10.1007/s10549-004-1470-1

395. Brandt J, Garne JP, Tengrup I, Manjer J. Age at diagnosis in relation to survival following breast cancer: a cohort study. *World J Surg Oncol*. 2015;13(1):33. doi:10.1186/s12957-014-0429-x

396. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak*. 2019;19(1):146. doi:10.1186/s12911-019-0874-0

397. Stoppa-Lyonnet D. The biological effects and clinical implications of BRCA mutations: where do we go from here? *Eur J Hum Genet*. 2016;24(Suppl 1):S3-S9. doi:10.1038/ejhg.2016.93

398. Bodmer WF, Bailey CJ, Bodmer J, et al. Localization of the gene for familial adenomatous polyposis on chromosome 5. *Nature*. 1987;328(6131):614-616. doi:10.1038/328614a0

399. Kastrinos F, Syngal S. Inherited Colorectal Cancer Syndromes. *Cancer J Sudbury Mass*. 2011;17(6):405-415. doi:10.1097/PPO.0b013e318237e408

400. Quaglia A, Tavilla A, Shack L, et al. The cancer survival gap between elderly and middle-aged patients in Europe is widening. *Eur J Cancer*. 2009;45(6):1006-1016. doi:10.1016/j.ejca.2008.11.028

401. Lu C-H, Lee S-H, Liu K-H, et al. Older age impacts on survival outcome in patients receiving curative surgery for solid cancer. *Asian J Surg*. 2018;41(4):333-340. doi:10.1016/j.asjsur.2017.02.008

402. Babińska M, Chudek J, Chełmecka E, Janik M, Klimek K, Owczarek A. Limitations of Cox Proportional Hazards Analysis in Mortality Prediction of Patients with Acute Coronary Syndrome. *Stud Log Gramm Rhetor*. 2015;43(1). doi:10.1515/slgr-2015-0040

403. Walters S, Maringe C, Coleman MP, et al. Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004–2007. *Thorax*. 2013;68(6):551-564. doi:10.1136/thoraxjnl-2012-202297

404. Walters S, Maringe C, Butler J, et al. Breast cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK, 2000-2007: a population-based study. *Br J Cancer*. 2013;108(5):1195-1208. doi:10.1038/bjc.2013.6

405. O'Connell JB, Maggard MA, Ko CY. Colon Cancer Survival Rates With the New American Joint Committee on Cancer Sixth Edition Staging. *JNCI J Natl Cancer Inst*. 2004;96(19):1420-1425. doi:10.1093/jnci/djh275

406. Grading (Tumors) - an overview | ScienceDirect Topics. Accessed November 19, 2020. https://www.sciencedirect.com/topics/medicine-and-dentistry/grading-tumors

407. Salehidoost R, Mansouri A, Amini M, Aminorroaya Yamini S, Aminorroaya A. Diabetes and all-cause mortality, a 18-year follow-up study. *Sci Rep*. 2020;10(1):3183. doi:10.1038/s41598-020-60142-y

408. Beg MS, Dwivedi AK, Ahmad SA, Ali S, Olowokure O. Impact of diabetes mellitus on the outcome of pancreatic cancer. *PLoS ONE Electron Resour*. 2014;9(5):e98511.

409. El-Jurdi NH, Saif MW. Diabetes and pancreatic cancer. *Jop J Pancreas Electron Resour*. 2013;14(4):363-366.

410. Magruder JT, Elahi D, Andersen DK. Diabetes and pancreatic cancer: chicken or egg? *Pancreas*. 2011;40(3):339-351. doi:10.1097/MPA.0b013e318209e05d

411. Forbes A, Murrells T, Mulnier H, Sinclair AJ. Mean HbA1c, HbA1c variability, and mortality in people with diabetes aged 70 years and older: a retrospective cohort study. *Lancet Diabetes Endocrinol*. 2018;6(6):476-486. doi:10.1016/S2213-8587(18)30048-2

412. Huang Y, Zheng H, Chen P, et al. An Elevated HbA1c Level Is Associated With Short-Term Adverse Outcomes in Patients With Gastrointestinal Cancer and Type 2 Diabetes Mellitus. *J Clin Med Res*. 2017;9(4):303-309.

413. Penno G, Solini A, Zoppini G, et al. Hemoglobin A1c variability as an independent correlate of cardiovascular disease in patients with type 2 diabetes: a cross-sectional analysis of the renal insufficiency and cardiovascular events (RIACE) Italian multicenter study. *Cardiovasc Diabetol*. 2013;12:98. doi:10.1186/1475-2840-12-98

414. Kendal WS. Dying with cancer: the influence of age, comorbidity, and cancer site. *Cancer*. 2008;112(6):1354-1362. doi:10.1002/cncr.23315

415. Sarfati D, Koczwara B, Jackson C. The impact of comorbidity on cancer and its treatment. *CA Cancer J Clin*. 2016;66(4):337-350. doi:10.3322/caac.21342

416. Abdel-Rahman O. Impact of Diabetes on the Outcomes of Patients With Castration-resistant Prostate Cancer Treated With Docetaxel: A Pooled Analysis of Three Phase III Studies. *Clin Genitourin Cancer*. 2019;17(1):e104-e112.

417. Agache A, Mustatea P, Mihalache O, et al. Diabetes Mellitus as a Risk-factor for Colorectal Cancer Literature Review - Current Situation and Future Perspectives. *Chir Bucuresti*. 2018;113(5):603-610.

418. Bakhru A, Buckanovich RJ, Griggs JJ. The impact of diabetes on survival in women with ovarian cancer. *Gynecol Oncol*. 2011;121(1):106-111.

419. Bar B, Hemphill JC. Charlson comorbidity index adjustment in intracerebral hemorrhage. *Stroke*. 2011;42(10):2944-2946. doi:10.1161/STROKEAHA.111.617639

420. Moro-Sibilot D, Aubert A, Diab S, et al. Comorbidities and Charlson score in resected stage I nonsmall cell lung cancer. *Eur Respir J*. 2005;26(3):480-486. doi:10.1183/09031936.05.00146004

421. Frenkel WJ, Jongerius EJ, Mandjes-van Uitert MJ, van Munster BC, de Rooij SE. Validation of the Charlson Comorbidity Index in acutely hospitalized elderly adults: a prospective cohort study. *J Am Geriatr Soc*. 2014;62(2):342-346. doi:10.1111/jgs.12635

422. Williamson E, Walker A, Goldacre B. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. 2020;584:430-436.

423. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Med Inform*. 2019;7(2). doi:10.2196/12239

424. Alhusain L, Hafez AM. Cluster ensemble based on Random Forests for genetic data. *BioData Min*. 2017;10(1):37. doi:10.1186/s13040-017-0156-2

425. Chao C, Yu Y, Cheng B. Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine , Logistic Regression and Decision Tree. *J Med Syst*. 2014;38(106):1-7. doi:10.1007/s10916-014-0106-1