Latent Variable Modelling for Complex Observational Health Data

Wendy Jane Harrison

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

The University of Leeds
School of Medicine

October 2016

Intellectual Property Statement

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 3 contains work based on the following publication:

1. Harrison, W.J., Gilthorpe, M.S., Downing, A., Baxter, P.D. 2013. Multilevel latent class modelling of colorectal cancer survival status at three years and socioeconomic background whilst incorporating stage of disease. *International Journal of Statistics and Probability*; **2**(3), pp.85-95.

   Wendy J. Harrison conceived and developed the idea, performed all analyses and drafted the manuscript. The other authors made contributions to the draft manuscript.

Chapter 4 contains work based on the following publication:

2. Gilthorpe, M.S., Harrison, W.J., Downing, A., Forman, D., West, R.M. 2011. Multilevel latent class casemix modelling: a novel approach to accommodate patient casemix. *BMC Health Services Research*. **11**(53).

   As an earlier investigation of greater complexity, the idea and plan was conceived by Professor Mark S. Gilthorpe. Wendy J. Harrison developed the idea, performed all analyses and drafted the manuscript. The other authors made contributions to the draft manuscript.

Chapters 2, 3 and 4 contain work based on the following publication:

3. Harrison, W.J., West, R.M., Downing, A., Gilthorpe, M.S. 2012. Chapter 7: Multilevel Latent Class Modelling. In: Greenwood, D.C. and Tu, Y-K. eds. *Modern Methods for Epidemiology*. London: Springer, pp.117-140.

   A book chapter containing elements of the previous two publications; Wendy J. Harrison performed all additional analyses and drafted the chapter. The other authors made contributions to the draft chapter.

Dedication

For Dad

Acknowledgements

I would like to thank my supervisors Professor Mark S. Gilthorpe and Dr Paul D. Baxter for their advice and encouragement throughout. Without your unwavering support, this thesis would not have been submitted.

Thank you to Dr Graham R. Law and all members of the Division of Epidemiology & Biostatistics for your help and advice through this process.

I would also like to thank the Northern and Yorkshire Cancer Registry and Information Service (NYCRIS) for access to the routinely collected data for the purposes of this research.

Finally, thank you to Mum, and to John, for believing I could do this.

Abstract

Observational health data are a rich resource that present modelling challenges due to data complexity. If inappropriate analytical methods are used to make comparisons amongst either patients or healthcare providers, inaccurate results may generate misleading interpretations that may affect patient care. Traditional approaches cannot fully accommodate the complexity of the data; untenable assumptions may be made, bias may be introduced, or modelling techniques may be crude and lack generality.

Latent variable methodologies are proposed to address the data challenges, while answering a range of research questions within a single, overarching framework. Precise model configurations and parameterisations are constructed for each question, and features are utilised that may minimise bias and ensure that covariate relationships are appropriately modelled for correct inference. Fundamental to the approach is the ability to exploit the heterogeneity of the data by partitioning modelling approaches across a hierarchy, thus separating modelling for causal inference and for prediction.

In research question (1), data are modelled to determine the association between a health exposure and outcome at the patient level. The latent variable approach provides a better interpretation of the data, while appropriately modelling complex covariate relationships at the patient level. In research questions (2) and (3), data are modelled in order to permit performance comparison at the provider level. Differences in patient characteristics are constrained to be balanced across provider-level latent classes, thus accommodating the ‘casemix’ of patients and ensuring that any differences in patient outcome are instead due to organisational factors that may influence provider performance.

Latent variable techniques are thus successfully applied, and can be extended to incorporate patient pathways through the healthcare system, although observational health datasets may not be the most appropriate context within which to develop these methods.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIC ……	Akaike Information Criterion
APACHE ……	Acute Physiology and Chronic Health Evaluation
BIC ……	Bayesian Information Criterion
BMI ……..	Body Mass Index
CCG ……	Clinical Commissioning Group
CE ……..	Classification Error
CI ……..	Confidence Interval
COSD ……	Cancer Outcomes and Services Dataset
DAG ……	Directed Acyclic Graph
DCO ……	Death Certificate Only
EM ……..	Expectation-maximisation
GAM ……	Generalised Additive Models
GP ……..	General Practitioner
ICC ……	Intraclass Correlation Coefficient
ICD-10 ……..	10th revision of the International Classification of Diseases
IPW ……	Inverse Probability Weighting
IRT ……	Item Response Theory
ITT ……	Intention to Treat
LCA ……	Latent Class Analysis
LL ……..	Log-likelihood
LSOA …….	Lower Super Output Area
MeSH …….	Medical Subject Heading
MI ……..	Multiple Imputation
ML ……..	Maximum Likelihood
MLLC …….	Multilevel Latent Class
MLM ……	Multilevel Model / Multilevel Modelling

MPT	…….	Multiple Primary Tumour
MRC	…….	Medical Research Council
NCIN	…….	National Cancer Intelligence Network
NHS	…….	National Health Service
NYCRIS	…….	Northern and Yorkshire Cancer Registry and Information Service
ONS	…….	Office for National Statistics
OR	……..	Odds Ratio
SD	……..	Standard Deviation
SES	…….	Socioeconomic Status
SMR	…….	Standardised Mortality Ratio
SOA	…….	Super Output Area
TDI	…….	Townsend Deprivation Index

Preface

Prior to commencing this research activity in 2007, I was employed within the National Health Service (NHS) as a medical statistician at the Northern and Yorkshire Cancer Registry and Information Service (NYCRIS), where I was responsible for the provision of specialist statistical input relating to all cancer information outputs for the local geographical region. I became involved in a project to investigate the changing effect of socioeconomic deprivation on colorectal cancer incidence rates, which highlighted concerns about the potential introduction or exacerbation of bias when using regression techniques to model imprecise or incomplete covariates.

My honorary contract with the University of Leeds allowed me to collaborate with academic statisticians on the exploration and development of statistical methodologies, thus I became aware of Latent Class Analysis (LCA) as an emerging new method that may be able to address these covariate issues. I worked with Professor Mark S. Gilthorpe to assess the utility of LCA to model cancer registry data, and initial results were promising, potentially showing the approach to be unbiased in estimating the impact of key covariates, when compared with regression analysis.

I commenced my career at the University of Leeds in 2007, but maintained links with NYCRIS, as I was keen to continue to explore and develop innovative approaches to the analysis of cancer registry data specifically, and of routinely collected observational health service data in general. My PhD studies have thus allowed me to apply my developing knowledge of latent variable modelling approaches in an attempt to provide empirical answers to important health service questions, using the cancer registry data as an exemplar. Over time, my interest has expanded to consider also the latent variable approaches from a methodological perspective, and to reflect upon the use of observational datasets as a context within which to develop such methods. Chapters 3 and 4 of this thesis contain earlier, application-based work while Chapter 5 offers a more methodological approach.

A list of detailed definitions of key terms, phrases and abbreviations follows, to set the scene for their later use within this thesis. They are presented in order of their consideration within the abstract and main thesis.

Observational health data. Within this thesis, I classify 'observational health data' as any set of data that is generated by observing patients as they progress through the healthcare system. This progression is termed the 'patient journey' (described in section 2.3.5). These datasets are commonly generated by routine data collection i.e. healthcare organisations record events (e.g. diagnoses, deaths or treatment received) and these events may then be linked together using appropriate identification codes. Thus, observational health datasets may be generated from multiple data sources. Their structure is inherently complex, as described in section 1.2.1.

Observational also refers to the healthcare setting or framework, i.e. data are not obtained from within a clinical trial setting, where patients are typically allocated to treatments based on randomisation; rather, each patient receives care specific to their individual circumstances. It is anticipated that this care may vary by patient, disease group and / or healthcare provider.

Traditional. I use this term to reflect the type of analyses that are most commonly used to examine observational health data within healthcare organisations. Typically, for instance, regression analyses are employed (see section 3.2.2), which may be single level (see section 1.3.1) or multilevel (see section 1.3.2). Traditional strategies are also often utilised to accommodate differences in patient characteristics across healthcare providers (termed 'casemix'; see sections 1.2.3 and 1.3.4). Other terms were considered (e.g. 'standard' or 'established'), but as none seemed ideal, I chose 'traditional' as a general term to describe the more conventional approach to these type of data evaluation approaches.

I do not suppose that other analytical methods are not available, feasible, or utilised; rather, that 'traditional' approaches are those most often employed in the healthcare environment, which may thus benefit from comparison with less commonly adopted or novel analytical approaches.

Methodology. I use this term, defined as “a system of methods and principles used in a particular discipline” (Hanks et al., 1986), when referring to a broad spectrum of methods. For example, latent variable ‘methodology’ or traditional ‘methodology’, may each comprise many possible methods.

The latent variable ‘methodology’ (also ‘approach’, or ‘technique’) may thus refer to any modelling that is performed within a latent variable framework, whether with continuous or discrete observed or latent variables. Specific methods, such as latent class analysis (LCA) or multilevel modelling (MLM), lie within the framework of latent variable methodologies.

Patient pathway. As part of the ‘patient journey’ (see section 2.3.5), the patient pathway reflects the progress of a patient through the healthcare system and may include, for example, tests, medication or surgery. This pathway may be influenced by characteristics of the patient and / or of the disease, and may also be affected by processes within the healthcare organisation(s) attended. Therefore, the patient pathway may differ for each patient, and may differ for two patients of identical socio-demographic backgrounds with identical health conditions when entering the healthcare system.

Causal framework. This term refers to a framework within which modelling for causal inference (see section 1.4.2) is performed. As discussed in section 1.2.1, questions posed within healthcare research commonly relate to causal factors (upon which one might intervene), necessitating a causal inference perspective, rather than merely invoking a predictive modelling approach. Research questions (1) and (3) (see section 1.2.3) explicitly consider causal effects at the patient and provider level respectively.

Patient casemix. This term refers to differences in patient characteristics across healthcare providers. As raised in research question (2) (see section 1.2.3), patient characteristics may vary geographically and hence may vary with respect to circumstances affecting (or even driving) their health status, reflecting different patient combinations across providers that are situated in different geographical locations. In order to make a fair comparison across healthcare providers, patient casemix should be accommodated. Traditional approaches to patient casemix adjustment are discussed in section 1.3.4.

MLM. As described in section 1.3.2, this term refers to the extension of single level regression modelling, where lower-level observations are clustered within higher-level groups. This approach is defined initially within this thesis as a ‘traditional’ technique, as it is commonly utilised within the healthcare environment (see section 3.2.2). It can, however, also be considered as a simple example of a latent variable model, as indicated in section 1.4.1, with homogeneous subgroups at each level of the hierarchy.

LCA. As described in section 1.4.3, I use this term to refer to any statistical analysis where the model allows for parameters to differ across latent subgroups, following the definition offered by Vermunt and Magidson (Vermunt and Magidson, 2003). The single level LCA approach is explained in detail in section 2.2.1.

Multilevel latent class (MLLC) analysis. Also described as MLLC modelling, I use this term as an extension of LCA, to refer to any statistical analysis where discrete latent variables may be incorporated at multiple levels of a hierarchy. The MLLC approach is explained in detail in section 2.2.2.

Chapter 1

Introduction

1.1 Introduction

This thesis explores the utility of unexploited, novel statistical techniques to analyse complex observational health data. Latent variable approaches lie within an overarching causal framework, where modelling may be performed either to adjust for confounding factors and hence make causal inference (i.e. to determine the effect (and magnitude of effect) of an independent variable as an assumed cause of a dependent variable), or to account for differential selection (i.e. to accommodate differences in characteristics (commonly within patients) and thus improve estimates of effect). There is much scope to model complex data configurations, with latent variable methodologies able to account for generic data challenges, such as non-homogeneity, measurement error and causal relationships between covariates, while maintaining a framework that may also be utilised to model patient pathways through the healthcare system, including treatment effects and other institutional characteristics. While traditional methodologies may be appropriate to address some of the fundamental challenges within observational health data, there is no other current methodology available that is able to provide such a comprehensive approach. Further, no other applications have, as yet, similarly exploited the capabilities of the techniques to be addressed within this thesis (evidenced in section 1.4.5).

In order to demonstrate the utility of the latent variable approach, three research questions are considered, representing questions that may be asked about differing aspects of the patient pathway through the healthcare system. An example of a clinical dataset is utilised. Multilevel latent class (MLLC) models are constructed, with model parameterisations tailored to be specific to each research question, yet standard in approach. Where

feasible, approaches are contrasted with traditional modelling techniques to demonstrate proof of principle and to either illustrate comparable results, or generate improved estimates (due, perhaps, to appropriate model construction), and hence an enhanced interpretation.

Chapter 1 introduces all key aspects of the thesis, including background and rationale, and establishes the context for the following chapters.

Section 1.2 examines observational health data, considering its inherent structural complexity and generic data challenges. The three research questions and the example dataset are introduced.

Section 1.3 describes the traditional modelling approaches, with a focus on regression analysis as the most commonly used method, and the ability of these techniques to respond either to the generic data challenges, or to account for differential selection.

Section 1.4 introduces the latent variable methodologies, with consideration of their use within causal inference modelling, a brief history of the techniques and a literature review. The applicable statistical software is introduced.

Section 1.5 details the content of the following chapters.

1.2 Observational health data

1.2.1 Structural complexity

Observational health data are a rich resource, commonly collected as part of an ongoing process of data collection by a healthcare provider, such as in an audit, rather than in a more structured manner as would be seen in a clinical trial setting, for example. Sources are numerous: for example, the National Health Service (NHS) holds national datasets compiled from records of patient care (NHS Digital, 2017), which may be used to support commissioning services or service planning, and disease specific datasets are available based on registrations (e.g. Cancer Outcomes and Services Dataset (COSD) (National Cancer Intelligence Network, 2010)). Much related information may be collected together, such as patient characteristics, disease onset and progression, treatment and care pathways, and attendance at one or more healthcare provider locations for diagnosis, treatment or specialist opinion. There are therefore connections between the different aspects of information collected, i.e. the patients, the treatment or care received, and the healthcare provider attended, and the relationships between these aspects may be complex. Figure 1.1 shows how these relationships may be perceived graphically.

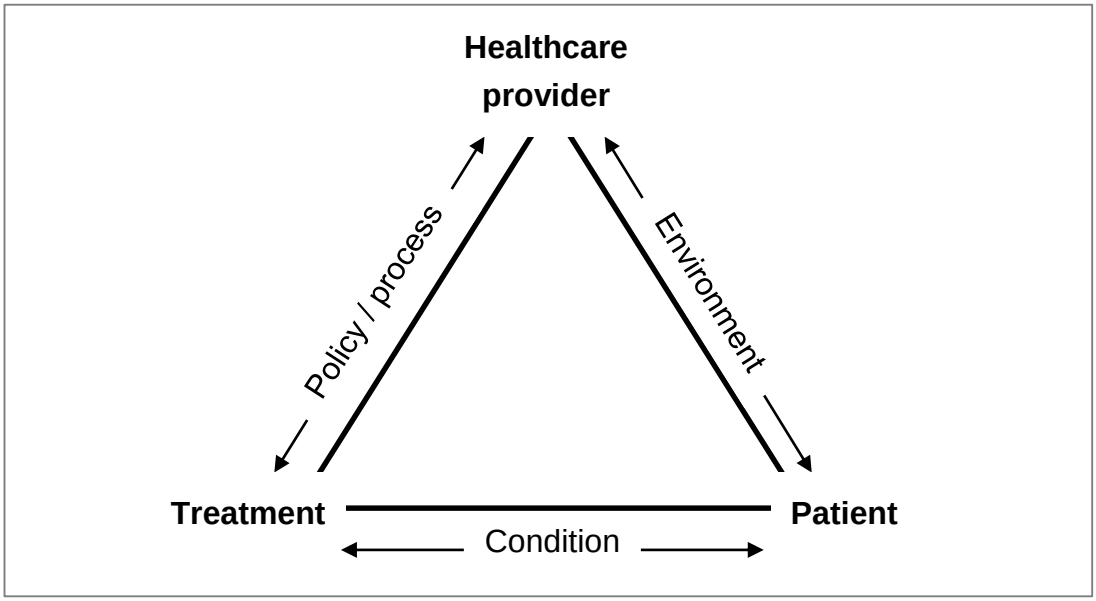


Figure 1.1 Graphical perceived relationships between patient, treatment and healthcare provider

Examples of patient characteristics may be age at diagnosis, sex and stage of disease, while examples of treatment characteristics may be the type of treatment received, and time from diagnosis to treatment. The holistic provision of healthcare means that variation in any part of the patient care pathway may impact on patient outcomes. For example, if patients from a homogeneous group receive the same treatment for the same condition, they may still respond differently if they are attending different healthcare providers, possibly due to organisational characteristics such as surgeon specialism or available beds. In an observational setting, it is not feasible to standardise all aspects of care. Different diseases are also not identically managed, therefore there is inherent heterogeneity surrounding patient entry to the healthcare system.

While the assessment of treatment effects is usually the domain of clinical trials, many research questions may be asked of observational health data that are of interest to the patient (e.g. what factors affect disease survival?), or to the healthcare provider (e.g. what constitutes good practice?). Any analysis performed must fully accommodate the complexity of the data and the healthcare environment in order to ensure correct inference, as inappropriate interpretation may have a direct impact on patient care. In an observational health framework, these questions may not be set at the start of the data collection process, meaning that when they are posed later, there may be data challenges in addition to the structural complexity that must be addressed prior to, or as part of, the analysis. Challenges generic to many observational health datasets are discussed in section 1.2.2.

“Big data” is becoming widely recognised as an all-encompassing term used to describe extremely large and complex datasets that are stored and analysed digitally (Boyd and Crawford, 2012), and it has been designated as one of the “eight great technologies that support UK science strengths and business capabilities” (Department for Business, Innovation and Skills, 2013). These datasets may be linked, further extending their size and scope. Research applications are therefore extensive, with involvement from many academic and research institutions, including the University of Leeds (University of Leeds, 2017).

There is much promise inherent in being able to access and examine such expansive data, however these datasets suffer from the same data challenges seen in observational health data, and the use of automated techniques that do not correctly address these challenges may result in results being based on predictive modelling (i.e. predicting the value of a dependent variable, based on values of independent variables) rather than causal inference. While a predictive modelling approach may be appropriate in many research areas, healthcare research is an inherently causal framework as questions commonly relate to 'causes' either of disease, or of relief from disease.

1.2.2 Generic data challenges

In addition to structural complexity, observational health data suffer from many data challenges generic to routinely collected data. Appropriate accommodation of each is essential in order to make correct inference, and avoid misleading interpretations, perhaps due to biased results. The generic data challenges are introduced here, and the consequences of inappropriate modelling with respect to these challenges are examined in section 1.3.3.

(i) Structure. Observational health data are commonly structured in a hierarchical manner, for example, patients living in the same geographical area may attend the same clinic or hospital for treatment, and these patients may attend multiple times. Therefore, patients can be said to be 'clustered' within the relevant healthcare provider, and measurements may also be 'clustered' within the patient.

(ii) Non-homogeneity. There may be differences within the population studied at any level of the hierarchy, i.e. samples may be heterogeneous at any level. For example, patients may vary in their characteristics (Office for National Statistics, 2016b) or in their response to treatment (Roden and George Jr, 2002), and healthcare providers may utilise different resources dependent on the route of admission (Simmonds et al., 2014), potentially leading to differences in the level of care received by the patient.

(iii) Measurement error. Measurements may also be imprecise, perhaps due to variations within a patient during their care; increased levels of anxiety may increase blood pressure (Sheppard et al., 2016), for example. Measurements taken by different clinicians (Wallis et al., 2015), or using different equipment (Wiesel et al., 2014), may also be interpreted differently.

(iv) Unmeasured variables. Despite recording a large amount of information, there may remain variables that are not included in a dataset, perhaps because their association is unknown, but which may affect the outcome. Because they are unidentified, any effect due to their exclusion is unknown.

(v) Complex observed relationships. The observed variables collected on patients as part of routine data collection, such as age, sex and stage, may have complex relationships with each other within a population, and these relationships may differ across populations. For any given research question where inference of an exposure (independent variable) on an outcome (dependent variable) is required, there may be any number of other variables that may either confound, mediate, or moderate, this relationship.

(vi) Missing data. There may be missing data, and the data that are not recorded may be related to some quality of the population to be studied. For example, basic measurements such as height, weight and blood pressure are commonly taken when a patient is admitted to hospital (Evans and Best, 2014), but if a patient is very ill on admission, it may not be feasible to take these measurements. Missing data may therefore be predictive of an underlying health state.

(vii) Area-based measurements. Individual measures of deprivation are rarely available, especially when using routine data. Indices of socioeconomic status (SES) such as the Townsend Deprivation Index (TDI) (Townsend et al., 1987) or the Index of Multiple Deprivation (Noble et al., 2004) are all that are routinely available. These indices are measured at the small-area level, such as electoral ward or super output area (SOA). Their use can lead to the ecological fallacy (Robinson, 1950) if area-based findings are extrapolated to individuals living within each area.

1.2.3 Three research questions

Three research questions are posed. These questions, although conceived to demonstrate the latent variable approach, represent typical enquiries commonly made of observational health data. Each question concerns a different aspect of the patient journey through the healthcare system.

(1) What is the relationship between a health exposure and outcome, and what other factors affect this relationship?

This is an example within epidemiology where interest lies in determining the association between a health exposure or risk factor (e.g. SES) and an outcome (e.g. survival), where it is difficult, if not impossible, to conduct a randomised controlled trial. Causal inference is sought within a multilevel framework where focus is on the patient level and variation at all other levels is effectively 'nuisance', i.e. upper-level variation must be accounted for, but inference is not required.

An intractable problem within causal inference modelling is also raised, where a potential interaction (e.g. between SES and stage) may be of interest, but may introduce bias if not sought carefully. Inappropriate adjustment of alleged confounders that may lie on the causal path between exposure and outcome can invoke bias (Kirkwood and Sterne, 2003), known as the reversal paradox (Stigler, 1999), and this bias has been shown to be a potentially serious problem in epidemiology (Hernández-Díaz et al., 2006; Tu et al., 2005). A Directed Acyclic Graph (DAG) (Pearl, 2000) is essential to assess covariate relationships.

(2) How does the performance of a healthcare provider vary after accommodating patient differences?

One area of interest in healthcare provision is performance monitoring, where indicators are used to measure, and compare, outcomes at an area level, for example by NHS Trust (Raleigh et al., 2012; Abel et al., 2014; Gomes et al., 2016). Different patient characteristics (e.g. age, sex and stage) may, however, lead to different outcomes (e.g. survival from disease), therefore these characteristics should be balanced across providers to

ensure a fair comparison of performance. Patient population characteristics may vary geographically (e.g. by age and sex (Office for National Statistics, 2016b)), and as patients commonly attend a healthcare provider close to their geographical location (Dixon et al., 2010), patient characteristics may therefore vary across healthcare providers, which is termed ‘casemix’, and thus leads to differential access to care, a form of differential selection.

This is a major topic of interest, as there are few strategies that can overcome the uncertainties associated with patient casemix differences (see section 1.3.4). To establish this approach within a framework that can extend to accommodate patient pathways (e.g. treatment effects) is challenging, and original. Initially, no provider-level covariates are examined in answering this question, in order to make comparison to existing methods. This extension is possible, however, as explored in research question (3).

(3) Can causal provider-level covariate effects be identified, after accommodating patient differences?

This is an extension of research question (2), where modelling for prediction at the patient level (i.e. accounting for casemix differences), is separated from causal inference at the provider level, in order to examine organisational factors (e.g. surgeon specialism or available beds) that may affect patient outcomes. A deliberate limitation at this stage is not considering multivariable DAGs at the provider level in order to first establish the principle that a single provider-level causal effect can be recovered.

This novel application demonstrates the flexibility of a methodological framework that must account for a hierarchical data structure, accommodate uncertainty due to both measured and unmeasured variables, adjust for patient casemix, and exploit the complexity (i.e. heterogeneity) of the data in order to partition prediction and causal inference.

1.2.4 Example dataset

To investigate the three research questions, an example dataset is utilised, containing routinely collected data for patients diagnosed with colorectal cancer between 1998 and 2004. The dataset is thoroughly described in

Chapter 2. It is an example of a dataset available within observational health data, and is utilised here to demonstrate the latent variable techniques; many other observational health datasets could be analysed in the same manner, with appropriate compensation for data-specific challenges. This dataset is utilised to answer research questions (1) and (2), using available patient-level covariates.

For research question (3), however, the example dataset cannot be used as it does not contain any provider-level covariates. Furthermore, a real-world dataset would not be amenable to evaluation of the effectiveness of the proposed techniques, as simulation is ideally required to assess the effectiveness of the approach before evaluating the latent variable methodology in practice. Simulations are therefore undertaken to explore the proof of principle for the inclusion of provider-level covariates, which is essential to evaluate the robustness of the proposed strategy of analysis.

1.2.5 Simplifications

Certain deliberate simplifications are made to the data for the purposes of analysis. They are described here and their implications are explored further in Chapter 6.

Missing data

Not addressed within any of the research questions is how to accommodate missing data. Within the example dataset, stage at diagnosis suffers from missing data, with 13.1% of patients having missing values for stage. As only a minor concern within this dataset, data are therefore simplified by generating a separate category for the missing values; thus all stage data are included in the analysis. In general, however, methods to address missing data should be employed, which is feasible as a separate extension that could then be combined with latent variable modelling approaches. There are methodological challenges, however, as the tools are not yet available to impute missing values within a multilevel framework. This is not the focus of this thesis, and the simplified approach is thus considered sufficient to demonstrate proof of principle for each of the research questions. Missing data challenges are discussed further in section 6.3.2.

Health outcome

The outcome is selected to be whether or not the patient survives at three years following diagnosis, as this is clinically meaningful and facilitates ready comparison with other studies. It is, however, a simplification for what is potentially a survival measure. While survival analysis may be a desirable alternative, it would not be as comparable to other literature, hence the binary outcome is selected instead. There are also methodological challenges. The methods and associated principles proposed within this thesis will extend to a survival analysis context, although only in a different statistical software package to that used throughout for the latent variable modelling. This extension is discussed further in section 6.3.2.

Area-based measurements

SES is measured at the small-area level, but is attributed to individuals, which may provoke the ecological fallacy as described in section 1.2.2. For this reason, another level should ideally be introduced into any model – the small-area level – and this would be cross-classified with healthcare providers, i.e. patients from one small area might attend different providers, and similarly patients attending one provider may be drawn from different small areas of residence. Theoretically, it is possible to conduct a cross-classified latent variable model, where small areas may also be grouped into latent classes, although this is not a currently supported option within the statistical software used.

Of primary interest, however, is the illustration of the latent variable methodology, and the primary research questions also pertain to the population or sub-population (i.e. latent classes), not individuals. The simplified approach of attributing small-area scores of SES to individual patients is therefore adopted, omitting the cross-classified small-area level completely. Alternative modelling approaches are discussed in section 6.3.2.

1.3 Traditional methodologies

1.3.1 Single level regression analysis

Traditional modelling techniques, such as regression analysis, are commonly used to analyse observational health datasets (see section 3.2.2). Regression (linear and logistic) is a well-documented approach (Normand et al., 2005) where the relationship between an outcome and one or more exposures is modelled, effectively to 'adjust' the predicted outcome in relation to the likely influences of these factors. For example, the exposures may be patient characteristics such as age, sex and stage of disease, while the outcome may be survival from a disease. These covariates, however, also modify the estimated coefficient effects of each other, which may not always be appropriate, dependent on the research question. Regression analysis identifies a 'best-fit' model where the effect of covariates is the same over the whole sample, however individual measurements will vary, giving residual error. A linear regression model is traditionally identified using the 'least squares' approach, where the differences between the regression line and each observation are squared and minimised over all observations. An equation is then generated for the regression line in the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where Y is the outcome, β_0 is the intercept, and β_j (for j=*1* to *n*) is the slope for each of the X_j model covariates. This equation may be used for prediction, to predict the likely outcome for a specific set of observations. In a regression model, the intercept and slope parameters are estimated as fixed, giving a fixed-effects model.

1.3.2 Multilevel modelling (MLM)

Regression modelling is often extended to a multilevel framework in order to incorporate differences across healthcare providers (Leyland and Goldstein, 2001; Leyland and Groenewegen, 2003). This approach is utilised for hierarchical data, where lower-level observations (e.g. patients) are 'clustered' within higher-level groups (e.g. healthcare providers). In contrast

to single level regression modelling, the ‘best-fit’ model may differ dependent on which upper-level group is considered, and so both the intercept and slope are allowed to vary across these groups, giving a random-effects model. The variations of the intercepts and slopes are assumed to be normally distributed about a mean of zero and these variations are also assumed to be independent of the variation in the individual measurements.

1.3.3 Traditional approach to the data challenges

Neither single level regression analysis nor MLM are able to address all of the data challenges inherent within observational health data, as introduced in section 1.2.2. Specific challenges that remain unaccounted for by use of these techniques are described here.

Structure and non-homogeneity

Single level regression analysis does not take into account any hierarchical structure of the data, and homogeneity is assumed at the single level. Maintenance of the data structure during analysis is important in order to correctly estimate standard errors associated with estimates of effect, as underestimation of standard errors leads to overestimation of statistical significance, i.e. a type I error (Normand et al., 2005). Non-homogeneity generates residual error which may also increase standard errors. As both challenges are inherent within observational health data, modelling with single level regression analysis is not appropriate.

While MLM does account for a hierarchical data structure, and provides improved estimates compared with regression (Cohen et al., 2009; Damman et al., 2009), the assumptions of normality and independence may not be valid in observational health data, as patients are not randomly assigned to healthcare providers, and providers are not randomly allocated geographically. MLM also assumes that a study sample is homogeneous at every level of the hierarchy, i.e. the same model would be applied to all members of the sample and the effects of covariates would be the same throughout. This may not be valid in observational health data due to differential selection, as raised in section 1.2.3, relating specifically to research question (2).

Observed and unobserved variation

Regression techniques only allow for variation in the outcome, not in the model covariates, hence they are a poor choice to incorporate uncertainty in any of the observed variables, for example due to imprecise measurements. Studies have shown that statistical analyses using regression modelling (single level or multilevel) may yield biased results where model covariates have measurement error (Greenwood, 2012) or missing values (Carroll et al., 2006; Fuller, 1987). Furthermore, as regression is performed using observed covariates only, no adjustment can be made for unmeasured differences across the observations.

Complex observed relationships

Within a regression model, statistical adjustment is commonly sought for all potential confounders (i.e. variables that may affect both the exposure and the outcome) in order to assess the impact of an exposure on an outcome. Inclusion of a covariate that is instead an effect mediator (i.e. it potentially lies on the causal path between exposure and outcome) may introduce bias due to the reversal paradox, however, as introduced in section 1.2.3. For confounders that are also potential effect modifiers (i.e. they exhibit an interaction with the main exposure), product interaction terms are commonly included. If this confounder is also measured with error, however, or has missing values, bias may be exacerbated (Greenwood et al., 2006).

As the use of multiple covariates within regression modelling modifies both the predicted outcome and the coefficient effects of these covariates, regression is therefore best placed for use within predictive modelling, where the focus remains on the predicted outcome. In causal inference modelling, where there is a primary exposure of interest, for example to answer research question (1), it is the modification of model coefficients that is the focus and, in the circumstances just described, traditional regression approaches cannot fully model the complex relationships within the data.

1.3.4 Casemix adjustment strategies

There are a number of alternative strategies that adjust for differential selection, each effective within their own constraints, but none that are adaptable to analysis within an extended modelling framework. Measurement uncertainty within observed covariates cannot be accommodated, potentially untestable assumptions may be made, and they cannot accommodate provider-level variation. Patient variation is accommodated through measured covariates only, which is crude, as models ought to reflect the uncertainty associated with patient casemix characteristics. Further, no casemix-adjustment strategy will eliminate all bias, due to unmeasured differences amongst patients (Nicholl, 2007), and some procedures increase bias (Deeks et al., 2003).

Well-established techniques include matching (Rothman et al., 1986), stratification (Normand et al., 2005) and regression analysis.

In matching, pairs of subjects are matched, based on their observed characteristics, to generate subgroups for comparison (e.g. by treatment); unmatched patients are excluded from analysis. Both identifying and recording all factors that are required for matching to be effective is challenging, and near impossible in the area of routine data collection. Measurement error cannot be accounted for, and differential selection cannot be addressed due to the limited variables available.

With stratification, homogeneous subgroups are identified using strata defined from observed covariates. Each patient is assigned to one stratum, and no patients are excluded. Distributions of covariates are thus balanced across subgroups, and analysis is performed within the defined strata. Similar to the challenges described for matching, it is not realistic to expect to stratify on all relevant factors. Further, stratification on numerous variables can lead to small numbers within strata, which introduces increased uncertainty that is not directly compensated for in any way. Bias due to differential selection is not explicitly addressed.

Regression techniques, as described in sections 1.3.1 and 1.3.2, are commonly used to model variables relating to patient characteristics (see

section 3.2.2), thus adjusting the outcome with respect to these factors. Whilst care is required to model appropriately complex relationships in order to make causal inference, and to minimise bias due to the reversal paradox, no such concern is necessary when modelling for prediction. Viewing differential selection as a prediction problem (as indicated in section 1.2.3), therefore indicates the utility of regression techniques in casemix adjustment. They cannot, however, be used to model together both causal inference and purely prediction.

A balancing score, such as the propensity score (Rosenbaum and Rubin, 1983; Rosenbaum and Rubin, 1984), may be utilised. Such a score is calculated from all observed covariates; patients with the same propensity score will thus have approximately the same distributions of their observed covariates. It is commonly used in combination with matching, stratification or regression to increase precision and produce unbiased estimates (Rosenbaum and Rubin, 1983). Within matching, the propensity score may be utilised as the matching variable, while for stratification, equal sized subgroups may be defined by values of the propensity score (D'Agostino, 1998). In regression modelling, the propensity score may be modelled alone for a more parsimonious model, or in combination with a subset of observed covariates (Rubin, 1979). Complete case data are usually required when calculating the propensity score, as missing values cannot be included and their exclusion may bias the calculated score. The propensity score may also not be appropriate for subgroups with very different prediction covariates, for example across different disease groups. Fundamentally, however, propensity score analysis conflates confounding (i.e. for causal inference modelling) and differential selection (i.e. within predictive modelling), and there remains the possibility that this technique may actually introduce bias in some instances (Pearl, 2009; Pearl, 2011).

Alternative composite risk scores may be utilised, and disease specific scores are commonly available (e.g. Acute Physiology and Chronic Health Evaluation II (APACHE II) (Knaus et al., 1981); calculated from age plus twelve routine physiological measurements). These scores may also be utilised within matching, stratification or regression analysis techniques.

Direct or indirect standardisation approaches may also be employed (Rothman et al., 1986; Breslow and Day, 1987). These techniques essentially adjust the outcome using a reference (standard) population, thus enabling direct comparison across populations with differing casemix structures. The direct approach uses a population distribution (e.g. an age distribution) as the standard, while the indirect approach uses a common set of specific rates. Both methods compare the number of expected events (e.g. deaths) calculated from the standard population with those observed. Indirect standardisation utilises a standardised mortality ratio (SMR), which will be considered as the traditional comparison to a latent variable approach in Chapter 4. Direct standardisation is widely used (National Cancer Institute, 2016; International Agency for Research on Cancer, 2010) as it preserves consistency between populations, although it cannot be performed if standard population distribution figures are unknown. Indirect standardisation may be used without these figures, but the associated weightings reflect the casemix of the local population, meaning that it may not be appropriate to directly compare populations with very different patient casemix.

1.4 Introduction to latent variable methodologies

1.4.1 Latent variable framework

The latent variable terminology and framework are introduced here, together with the key advantages of using these methodologies. Comprehensive exploration of the specific modelling approaches and features, suitability of the methodologies to address the generic data challenges, and parameterisations appropriate to answer the three research questions are presented in Chapter 2.

Latent variable modelling is an inclusive term covering the identification of an unobserved (latent) structure within observational data. Underpinning the theory is the concept that observations can be grouped within latent variables, which may be continuous and distributed as per a standard cumulative distribution function (typically assumed to be normal; latent variable), discrete (latent classes), or a mixture of both, and that there is a mathematical relationship between the observed values and the latent structure. Early analyses separated latent variables and latent classes, while contemporary modelling allows combinations of both (Vermunt and Magidson, 2003).

As discussed in section 1.3.4, regression methods cannot separate modelling for prediction from causal inference, as is ultimately required when accounting for both patient casemix and potentially causal factors at the provider level. Use of a latent variable approach can separate the prediction focus (i.e. accommodation of differential selection) at the patient level, and the causal inference focus at the provider level, which serves to overcome this potential conflict between two distinctly separate analytical strategies. This is a fundamental advantage over traditional techniques, and ensures that the overarching methodology can be retained when answering a wide range of research questions.

The traditional MLM is a simple example of a latent variable model, with a single, homogeneous group at the lower level and a continuous, normally distributed, latent variable at the upper level. Incorporating discrete latent classes extends the utility, as described in detail in sections 2.2.1 and 2.2.2.

There are a number of features unique to latent variable approaches that can be utilised to model appropriately complex observational data. Class membership models, introduced in section 2.2.3, allow for variables to be modelled separately to the main exposure-outcome relationship, thus minimising bias due to measurement error or the reversal paradox, for example. Class-dependent and class-independent features, introduced in section 2.2.4.3, are vital to ensure precise configuration of the model based on context and research question. The overarching approach, however, remains unchanged and hence similar in all its merits to address the data challenges.

Sections 1.4.3 and 1.4.4 provide a brief history of the latent variable techniques for context, while section 1.4.5 provides a review of the literature to determine how these novel modelling approaches are being used in practice with observational data.

1.4.2 Modelling for causal inference

As indicated in section 1.4.1, latent variable modelling for either prediction or causal inference can be partitioned across levels of a hierarchy, which is fundamental to the applications addressed within this thesis. The approach to causal inference modelling incorporated here is well established within epidemiology (Greenland et al., 1999), where causal diagrams are explored as a method to identify relationships between modelled variables. The concept was formalised by Pearl (2000), in the use of DAGs (introduced in section 1.2.3) to display covariate relationships. These diagrams make explicit the causal assumptions made between model covariates, thus formally identifying confounders and other variables that may either mediate or modify the effect of an exposure on an outcome. These features were discussed briefly with respect to traditional modelling approaches in section 1.3.3, and will be explored further within the latent variable framework in section 2.2.4.1.

In more contemporary publications, Vanderweele (2015) emphasises methods to define and assess mediation and covariate interactions, while Pearl et al. (2016) provide a comprehensive introduction to causality.

1.4.3 A brief history of latent variable techniques

This is not intended to be a systematic, or comprehensive, review of the historical literature, but rather a consideration of important publications within the subject area, to put the methodology into context.

Latent variable approaches were first used in the field of social psychology, initially with a focus on interpretation rather than on statistical concerns. 'Factor analysis' (Spearman, 1904; Thurstone, 1947) first utilised continuous variables for both the observed and latent variables, while 'latent structure analysis' (Lazarsfeld, 1950; Lazarsfeld, 1959), incorporated discrete classes for both. It was hypothesised that, instead of correlations between individual responses, the population studied was in fact heterogeneous, with different underlying latent groups (Lazarsfeld, 1950). Related approaches, termed 'latent trait analysis' (Lord, 1952) (utilising continuous latent variables with discrete observed variables), and 'latent profile analysis' (Gibson, 1959) (utilising discrete latent classes with continuous observed variables) were soon introduced in practice. Latent trait analysis is commonly used in educational testing (Lord and Novick, 1968), as 'item response theory' (IRT).

Parameter estimation procedures (commonly using matrix algebra) were developed by many, notably Green (1951), Anderson (1954; 1959), Gibson (1955; 1962), and Lazarsfeld and Henry (Lazarsfeld and Henry, 1968). However, Goodman (1974a; 1974b; 1979) formalised the methodologies by his development of the maximum likelihood (ML) parameter estimation algorithm, methods to determine whether estimated parameters are identifiable, and consideration of how well the models fit the data. Goodman also extended the analysis to include nominal observed variables (Goodman, 1974b), which was followed by further extensions for ordinal observed variables (Muthén, 1984) and for longitudinal data (Hagenaars, 1990; Vermunt, 1997).

There was a gradual move towards a more generalised approach, with 'mixture modelling' allowing for models to contain mixtures of both continuous and discrete latent variables, explored for example in Anderson (1959), Bartholomew (1980), Muthén (1984; 2002), Arminger and Küsters (1989), and Skrondal and Rabe-Hesketh (2004). This, together with the

availability of more powerful computers towards the end of the 20th Century, made the methodologies widely available. Latent class analysis (LCA) became known as any statistical analysis where the model allows for parameters to differ across latent subgroups (Vermunt and Magidson, 2003), regardless of terminology or data type. Extensions to a multilevel framework were introduced by Vermunt (Vermunt, 2003; Vermunt, 2008a), which further widened the methodological scope, and effectively brings this historical summary up to date.

1.4.4 Structural equation modelling (SEM)

For completeness, SEM is addressed. This is also a latent variable methodology, with SEM itself incorporating many other approaches, such as 'confirmatory factor analysis', 'canonical correlation analysis', and 'latent growth curve models'. A full history of this broad methodology is not attempted here; two useful sources are Kaplan (2009) and Matsueda (2012), with Hox and Bechger (1998) providing a useful non-technical introduction to the techniques. There are two components to a SEM: the structural element that establishes a causal framework between observed variables, and the measurement element that specifies relationships between latent variables. Variation due to measurement error, for example, may be incorporated within the latent structures, while the causal framework is appropriate to address research questions commonly examined within observational health data. It may not, however, be as useful when considering differential selection i.e. to account for patient casemix, as path diagrams are designed to reflect all covariate relationships rather than to separate causal inference from prediction. It also becomes increasingly complex in a multilevel framework (Hox, 2013). SEM will not be utilised within this thesis. Rather, a more comprehensive approach is sought that addresses the generic data challenges, while distinguishing between causal inference and prediction within an overarching framework.

1.4.5 Review of comparable latent variable approaches

Consideration is also given to the current usage of similar approaches as presented within this thesis, i.e. where latent variables (continuous or discrete) are used at multiple levels of a hierarchy to account for data challenges, such as heterogeneity, while modelling causal relationships between covariates at any level. Comparable articles are identified and summarised, without full critical appraisal as the primary intent is to recognise the scope of the published material, rather than to assess the strengths and weaknesses of each application. The strengths and weaknesses of latent variable approaches are discussed in Chapter 6.

There is much differing terminology used across applications, which makes it complex to identify similar uses of the latent variable approach; the methods have evolved in isolation and terminology has developed independently within each context. A broad literature search is therefore performed initially, with all abstracts reviewed for their relevance. As applications within the social sciences are common, and those within the field of medicine are most relevant to the research questions, two databases are utilised: Medline and PsycINFO. Full search strategies can be seen in Appendix A. Book chapters are excluded as they generally focus on principles, rather than applications. The review spans ten years; as the techniques are still adapting, anything older than ten years is likely to have been superseded methodologically.

A total of 174 results are found initially across the two databases. Duplicates (N=31) are excluded, together with irrelevant articles (e.g. teaching notes; N=16), and single level latent variable approaches (N=20). Two further articles are excluded: that detailing the research performed for Chapter 4 of this thesis (Gilthorpe et al., 2011), and the methodological-based paper cited in section 1.4.3 (Vermunt, 2008a). Four additional articles are sourced from other citations, thus a total of 109 results are available for consideration.

Thirty-four results utilise multilevel factor analysis, for example Bostan et al. (2015) and Koch et al. (2016). As introduced in section 1.4.3, factor analysis is an early latent variable approach where continuous latent variables are used throughout. Although extended to incorporate a multilevel structure, latent classes are not permitted and as such, this technique is not

comparable to that adopted here. Of note, however, is an article by Varriale and Vermunt (2012), where multilevel factor models are extended to consider latent classes at the upper level, termed the multilevel mixture factor model. Although not discussed in detail here, as continuous latent variables remain at the lower level, this article highlights the overlap and emerging terminology across latent variable applications.

A further 32 articles utilise SEM as the primary analytical technique, for example Geiser et al. (2015) and Preacher et al. (2016). As explained in section 1.4.4, the SEM approach is not appropriate for model extensions to include differential selection, for example, so is not considered here.

Twenty-two results employ analyses to model longitudinal data, including investigation of 'trajectories', for example Mumford et al. (2013) and Sanfelix-Gimeno et al. (2015), and of 'growth', for example Tu et al. (2013), Smith et al. (2014) and Burns et al. (2015). These are special cases within the latent variable methodology; any longitudinal application will inherently be multilevel, but it is not analogous to the approach taken within this thesis.

Three articles discuss the use of latent variables within multiple imputation, for example He at al. (2014).

Of the remaining 18 papers, 10 can be considered to be simpler applications of the approach adopted for this thesis. Termed latent class 'cluster' analysis, this technique involves the profiling of attributes or characteristics into latent classes, or clusters, that may have utility. Measurement uncertainty is accommodated within the latent framework, while the multilevel approach accounts for the complexity of the data. Within these articles, the regression part of the model i.e. the relationship between exposure(s) and outcome, is not incorporated within the assignment to classes. Some employ follow up analysis to determine associations based on class membership, but none examine causal inference. There are applications in alcohol use (Rindskopf, 2006; van Lettow et al., 2013), with the former focusing on individual alcohol use within geographical sites, while the latter considers the classification of descriptive terms within groups of survey respondents. Van Horn et al. (2008) illustrate the techniques using an example in substance use, considering problem behaviours within

individuals and schools. Zhang et al. (2012) investigate the proportions of individuals who are obese within classes distinguished by individual consumption within the fast food environment. Three levels of analysis are considered within a study focusing on the family-level subtypes of patients with schizophrenia (Derks et al., 2012). There are applications within education, considering students' attitudes within University groups (Mutz and Daniel, 2013), or pupils' examination responses within teaching groups (Auer et al., 2016). A hierarchical approach to social exclusion (Pirani, 2013) considers classes of both individuals and geographical regions, while the perceptions of the causes of poverty within individuals and countries are similarly investigated by da Costa and Dias (2014; 2015).

Finally, 8 papers are identified as containing analysis that can be considered comparable to that adopted within this thesis. Three are primarily methodological, and are not examined in detail. Two incorporate simulation studies; one assesses model selection (Yu and Park, 2014), while the other investigates the performance of methods of parameterisation (Finch and French, 2014). The third methodological paper explores model specification using an illustrative application within education testing (Vermunt, 2008b).

Thus, 5 articles remain; each focuses on the application of multilevel latent variable approaches, using comparable techniques to those employed within this thesis, although none adopt exactly the same approach. They span a variety of disciplines, covering social science (Kalmijn and Vermunt, 2007), healthcare (Downing et al., 2010), behavioural research (Henry and Muthén, 2010), political science (Morselli and Passini, 2012), and education (Bennink et al., 2014).

In the earliest application, Kalmijn and Vermunt (2007) present an application investigating the homogeneity of social networks, where the age and marital status of individuals (the 'ego' level) are considered as a joint dependent variable, and are modelled to identify the association with individuals' network contacts (the 'alter' level). A non-parametric specification is used at the upper level, i.e. ego-level latent classes are identified, and a single latent class is used at the lower level. The principle is thus similar to that adopted here, although different in consideration at the

lower level, and the context is wildly different. The latent class approach is seen to improve model fit and aid interpretation, with network contacts identified as more similar than would be expected based on the ego class. This finding may be partly explained by the accommodation of uncertainty due to unmeasured covariates within the latent constructs.

In my collaborative work with Downing et al. (2010), a multilevel latent class approach is utilised to model the association between socioeconomic deprivation and breast cancer survival status at five years. Of primary interest is the utility of the approach to model appropriately stage at diagnosis, identified as a mediator of the deprivation-survival relationship. A continuous latent variable is utilised at the upper level for model fit and parsimony, and two patient-level classes are identified. Model fit improves when stage is excluded from the regression part of the model, and latent classes are clearly distinguished by disease severity. The research performed in Chapter 3 advances this work, by consideration of causal circularity and the inclusion of discrete latent classes at the upper level.

Henry and Muthén (2010) identify typologies of adolescent smoking status, using data from 10,772 European females within one of 206 rural communities. Parametric and non-parametric approaches are investigated, with the parametric approach providing the best fit to the data, although the non-parametric approach allows community-level classes to be identified. The selected outcome variable is latent, rather than observed, and covariates are included at both levels. The probability of membership of the individual-level classes may vary across the upper-level communities, thus allowing for interpretation of upper-level classes by proportions of lower-level typologies. No accommodation is made for differential selection, however.

In an unfamiliar context, that of political science, Morselli and Passini (2012) utilise the multilevel latent variable approach to model individuals within countries, in order to classify different types of political movement and protest. Unusually, the selection of four lower-level classes is based on an a priori hypothesis. The model is a good fit to the data, however, and classes are highly interpretable, with the inclusion of covariates again allowing investigation of class membership by characteristics. A non-parametric

approach is adopted at the upper level, and two classes are identified based on similarities in the lower-level class distribution. The latent class approach thus shows utility in the identification of differing attitudes towards democracy across protesters. Cross-country differences, however, are defined solely by protester characteristics.

Finally, in an application to education, Bennink et al. (2014) investigate students nested within school groups, with 3,458 students from 60 schools classified based on their responses to a 24-item multiple-choice test. Focus is on a performance comparison at the school level, with adjustment for student ability using a continuous latent variable at the lower level. Both continuous and discrete latent variables are utilised at the upper level, and classes thus identify a small minority of schools where performance is poor. With a single latent variable at the student level, however, no explicit accommodation is made for differential selection, as student ability is assumed to be homogeneous. With a heterogeneous patient group, as considered within this thesis, accommodation for casemix must be modelled explicitly. Nevertheless, the mathematical framework is comparable.

1.4.6 Statistical software

The software Stata (StataCorp, 2015) is used for all data management operations on the example dataset, including data manipulation, summary statistics, and the production of tables and charts. It is also used to perform the data simulations, collation of results and linear regression analyses for the simulated data used in Chapter 5.

The statistical software Latent GOLD (Vermunt and Magidson, 2005; Vermunt and Magidson, 2013) is used for all latent variable models. Technical specifications are set at a level where consistent results can be achieved without unduly extending analysis time due to computational requirements.

The software R (2010) is used to identify threshold values for covariates in the example dataset, discussed specifically in relation to research question (2) in Chapter 4. Although Stata could also have been utilised in this situation, R was chosen to gain experience of its approach.

1.5 Content of following chapters

The following chapters contain methods, data, results and interpretation that demonstrate the utility of the multilevel latent variable approach to answer a range of research questions. The overarching methodological framework is maintained throughout, while specific strategies and parameterisations are explored for each research question.

Chapter 2 fully explores key aspects of the latent variable methodological approaches that may be utilised to model appropriately complex observational health data. This includes discrete latent classes, covariate modelling based on complex relationships, and model features vital to the construction of detailed model configurations. Modelling approaches are provided with respect to the research questions. The example dataset is described in full, with context specific data challenges identified with reference to the generic challenges introduced in section 1.2.2.

Chapters 3, 4 and 5 contain methods, results and interpretation as required to answer research questions (1), (2), and (3) respectively.

Chapter 3 uses multilevel latent class (MLLC) modelling to answer research question (1) using the example dataset, and directly compares results with those from a traditional multilevel modelling (MLM) approach. Latent classes are identified at both the patient and provider levels, with modelling for causal inference at the patient level and adjustment for heterogeneity at the provider level. The focus is on patients, and consideration is given to the context specific data challenges discussed in section 2.4.2. Provider-level classes are also interpreted to contrast the latent class approach with the use of a continuous latent variable at the upper level in MLM.

Chapter 4 uses MLLC modelling to answer research question (2) using the example dataset, comparing performance rankings at the provider level with those generated by calculation of the standardised mortality ratio (SMR). Latent classes are again identified at both the patient and provider levels, however modelling techniques are partitioned across levels of the hierarchy, with the accommodation of differential selection at the patient level and causal inference at the provider level. Provider classes are thus ‘adjusted’

for patient casemix. This approach provides a foundation for the extension of MLLC models to incorporate patient pathway and process characteristics.

Chapter 5 extends the MLLC modelling approach established in Chapter 4 to answer research question (3) by incorporating provider-level covariates. The utility of the latent variable approach is demonstrated by accounting for differential selection at the patient level, while modelling for causal inference at the provider level. Data are simulated, including both binary and continuous provider-level covariates and both binary and continuous outcomes. As there is no appropriate comparison with a traditional approach, assessment is made of the ability of the MLLC models to recover simulated values of the provider-level covariate.

Chapter 6 unites the approaches utilised in Chapters 3, 4 and 5. Methods are reviewed, and comprehensive suggestions for future development are included.

Chapter 2

Latent Variable Methodologies and Example Dataset

2.1 Introduction

Chapter 1 introduced complex observational data, exploring the linked aspects of patient, treatment and healthcare provider, and also discussed the generic data challenges commonly seen in such datasets. Three research questions, typical of common enquiries, were posed with consideration of their utility in an observational data context, and the example dataset was introduced. The traditional methodologies, and their limitations, were described with reference to the data challenges, and specific strategies were examined that have traditionally been used to account for differential selection (i.e. patient casemix).

The latent variable framework was then introduced as an overarching causal framework that allows modelling for both inference and prediction, with fundamental advantages over traditional methodologies; latent variable features may be used both to address the generic data challenges and to account for differential selection, while the framework has the capacity to extend beyond the scope of this thesis to incorporate patient pathways through the healthcare system. A literature search in section 1.4.5 demonstrated that there are few other applications utilising the capabilities of this in-depth methodology.

Chapter 2 considers the latent variable methodologies in depth, with focus on the use of discrete latent classes and their potential application to observational health data. Aspects of the latent variable methodologies are explored in detail, including appropriate adjustment for variables that may have a complex observed relationship with the exposure or outcome, such as those that may confound, modify or mediate the exposure-outcome relationship. Specific features that are utilised to precisely configure the

modelling approaches are introduced and described in detail, together with consideration of errors in classification and a suggested approach to model construction. Appropriate modelling approaches for the research questions are considered, utilising the overarching causal framework, and exploring their utility in addressing the generic data challenges. Broad modelling strategies and detailed parameterisations are included. The example dataset is described in detail, from source to summary statistics, and specific data challenges relevant to the data are included.

Section 2.2 explores the latent class approaches in detail, introducing class membership models, key modelling features, errors in class assignment and optimum model construction.

Section 2.3 revisits the example dataset to describe how the dataset is obtained and adapted for use within this research activity.

Section 2.4 considers the modelling approach in detail, exploring the appropriate analytical methods with discussion of the data challenges, broad modelling strategies and detailed parameterisations.

This chapter contains work based on a publication (Harrison et al., 2012).

2.2 Methods and features

2.2.1 Latent Class Analysis (LCA)

LCA, also known as ‘discrete latent variable modelling’, or ‘mixture modelling’ (Goodman, 1974b; Magidson and Vermunt, 2004), is well established within single level regression analysis. In LCA, a number of discrete latent variables (i.e. latent classes, or subgroups), are identified, the optimum choice of which is selected by the researcher, typically informed by log-likelihood (LL) statistics. The Bayesian Information Criterion (BIC) (Schwarz, 1978), the Akaike Information Criterion (AIC) (Akaike, 1974), and changes in LL are commonly used as model-evaluation indicators, though models may also be selected on the basis of interpretation (Gilthorpe et al., 2009). Both the BIC and AIC incorporate a sense of model parsimony by accommodating the varying number of model parameters (Vermunt and Magidson, 2016), while the LL does not. Model parameters for each latent class are determined empirically, along with their contribution to the outcome distribution.

Observations are probabilistically assigned to latent classes, i.e. each observation has a probability of belonging to each latent class, which sums to one, as each observation must be fully assigned across all classes. This assignment is based on similarities in characteristics; latent classes are therefore homogeneous, with similar effects of each covariate on observations in the same latent class, although covariate effects may differ across the classes. The relationship between outcome and associated risk factors can thus be determined within each latent class, rather than over all observations. As with single level regression, the intercept and slope within each class are fixed, so no distributional assumptions are required.

Uncertainty surrounding class membership is incorporated within the latent classes, since observations may belong to all classes, with probabilities determined empirically. LCA thus manages the uncertainty associated with use of a limited number of predictors when determining subtypes of outcomes. Although accommodated implicitly within the latent framework, very few analytical research strategies seek clearly to exploit this aspect.

2.2.2 Multilevel Latent Class (MLLC) modelling

MLLC models (Vermunt, 2003; Vermunt, 2008a) are an extension of LCA that incorporate discrete latent variables at all levels of a hierarchy. Latent classes are thus determined at more than one level, with the choice, as in LCA, informed by model-evaluation statistics, or based on interpretation. An optimum solution is sought for all classes at all levels simultaneously using ML estimation (Goodman, 1974b) obtained by an adapted expectation-maximisation (EM) algorithm (Vermunt, 2003).

Observations are probabilistically assigned to latent classes at all levels, i.e. they have a probability of belonging to each lower- and upper-level class, which sum to one at each level to reflect full assignment of each observation across all latent classes at all levels. Assignment to classes at the lower level is based on similarities in characteristics (Skrondal and Rabe-Hesketh, 2004), while latent classes at the upper level may be based on either similarities or differences, dependent on model specification and research question. Latent classes at the lower level are thus homogeneous, while latent classes at the upper level may be either homogeneous or heterogeneous. Covariates can be included at any level and, as with single level LCA, their effect is the same within each latent class, but may differ across the classes (if deemed appropriate). The relationship between outcome and associated risk factors is again determined within each latent class, rather than over all observations and, if intercepts and slopes are fixed within the classes at all levels, no distributional assumptions are required.

A richer mixture model can thus be represented, with latent classes adopted at any or all levels of the hierarchy. As the number of classes at the upper level is increased, a MLLC model can also be viewed as the traditional MLM but with a relaxation of normality assumptions, i.e. many upper-level latent classes can be viewed as a discrete approximation of a continuous latent variable that need not necessarily be normally distributed.

The use of latent classes within a multilevel structure permits several complex model configurations, where each configuration may be designed to address a specific research question, within a specific context. Each parameterisation relates to different assumptions, and leads to different

interpretations, with the modelling choices lying with the researcher. Many potential models, however, would be meaningless in many instances, as some of these parameterisations may not be identifiable, and some identifiable models may not be interpretable. Careful consideration must therefore be used when specifying the model; decisions regarding model configuration must be justified and the implications of each parameterisation fully considered. These issues only become more complex in a multilevel setting. Analysis within this thesis illustrates how such complexity can be exploited to address otherwise challenging or even intractable problems, using novel analytical strategies to address a range of research questions.

2.2.3 Class membership models

Covariates can be entered into a latent class model as within a traditional regression model, i.e. as ‘predictors’ of the outcome variation. The same covariates may also enter the model as ‘predictors’ of the latent class structure; termed the ‘class membership model’. In either scenario, causality should not be inferred. The term ‘predictor’ is unfortunate in that it may mislead users to infer causality when it only implies an association between outcome variation and the covariate in question. With this in mind, the favoured nomenclature within this thesis is that covariates in the regression part of the model are referred to as ‘covariates’, whilst the covariates (same or different) in the class membership part of the model are referred to as ‘class predictors’, though causality is not to be inferred. For MLLC models, covariates and class predictors may operate at any level.

If variables are included both as covariates and as class predictors, their association with the outcome should not then be allowed to vary across the latent classes, as a model where a variable is both predicting class membership and where its effect differs across classes would be difficult to interpret. If a variable is included only as a class predictor, however, then the resultant latent classes at that level will have a graduated outcome analogous to that observed for different values of the class predictor, and the relationship between the outcome and associated risk factors can thus be explored across these classes.

2.2.4 Modelling features

2.2.4.1 Confounding, effect modification and mediation

Class predictors may be utilised to ‘remove’ variables from the regression part of the model, which may improve model precision due to measurement error or missing variables, or potentially minimise bias due to the reversal paradox. This has been shown explicitly in collaboration undertaken as part of my research activities (Downing et al., 2010), and is discussed in section 1.4.5.

Modelling a confounder that is also a potential effect modifier (e.g. alcohol consumption may modify the effect of smoking on cancer mortality) as a class predictor yields an implicit interaction, since the exposure-outcome relationship may vary across latent classes. This averts the need to include an explicit confounder-exposure product term in the regression part of the model, which would otherwise exacerbate any bias introduced if the confounder is measured with error or has missing values, as discussed in section 1.3.3 with respect to the traditional regression analytical approaches to the data challenges. Modelling effect modification this way minimises bias; uncertainty associated with confounder values is explicitly accommodated via the latent class part of the model.

If an alleged confounder lies on the causal path between an exposure and an outcome, it is termed a ‘mediator’ (e.g. diet may mediate the effect of maternal deprivation on birthweight); statistical adjustment that includes a mediator as a covariate within a regression model may introduce bias due to the reversal paradox, as also discussed in section 1.3.3. It would then be wise to discard the mediator as a model covariate. This does not preclude the mediator becoming a class predictor however, though some implicit bias may remain. Modelling a mediator as a class predictor yields the potential for implicit interaction, as before, where the exposure-outcome relationship may vary across latent classes. The exposure may cause the mediator, which in turn part determines the latent class structure, within which the exposure-outcome relationship may vary. Circularity thus arises in the causal interplay of exposure, mediator and outcome. This can be avoided if the exposure-

outcome relationship is not allowed to vary across latent classes. In such instances, only the intercept varies across each latent class, not the exposure-outcome slope. The latent classes generated during modelling will then have a gradient that corresponds to specific patient features that can be labelled post-hoc according to outcome (e.g. ‘high’ or ‘low’ birthweight) or to class predictors (e.g. ‘good’ or ‘poor’ diet).

2.2.4.2 Inactive covariates

Covariates may also be included as ‘inactive’ within the model, whereby their distribution across the latent classes may be identified and interpreted, but they are not allowed to affect either the relationship between exposure and outcome, or class membership. This may be useful as a crude solution to enable the inclusion of covariates containing substantial amounts of missing data (e.g. treatment data where not all patients receive treatment). Their inclusion as inactive covariates aids interpretability of the model results, but ensures that additional bias is not introduced due to the missing data.

2.2.4.3 Class-dependent and class-independent features

Within latent variable models, parameter restrictions may be applied such that more (or less) parsimonious models may be estimated as required. This is achieved using ‘class-dependent’ or ‘class-independent’ features, thus determining how parameters at a lower level are set in relation to class structures at higher levels. A different interpretation is seen for each, and the choice of configuration is driven by both the context and research question. These features are what enables the flexibility that is exploited in the novel approaches proposed in this thesis, and their complexity therefore warrants a detailed exposition.

Parameter restrictions may be set for intercepts, covariate effects, class sizes and error variances (where there is a continuous outcome); the class-independent option applies the constraints, while the class-dependent option relaxes them. The technical detail is given here, while the practical use of these features is explored in section 2.4.4 with specific relevance to latent classes at patient and provider levels.

Intercepts

Within a MLLC structure, latent class intercepts at the lower level may vary across latent classes at the upper levels. This is very broadly analogous to random intercepts within a traditional MLM. Latent class intercepts at a lower level, however, may be either class dependent or class independent in relation to class structures at higher levels, thus they may exhibit relative differences that are either identical or different within each upper-level class. In both cases, upper-level classes may differ in their overall outcome.

Where relative differences are identical, lower-level class intercepts differ by the same degree, irrespective of which upper-level class the observations are assigned to; intercepts are thus class independent. This configuration therefore enables identical contrasts to be made among lower-level classes, within each upper-level class.

Where lower-level class intercepts vary by different degrees across upper-level classes, the intercepts are class dependent. This configuration indicates that lower-level differences can mean different things according to which upper-level class is being considered.

Covariate effects

Covariate effects can apply at any level of a latent class structure. Similar to the concept surrounding intercepts, covariate effects at the lower level may be modelled as either class dependent or class independent, in relation to the upper-level classes. In traditional MLMs, for example, the lower-level covariates could have estimated regression coefficients that remain fixed across the upper-level classes (hence the term ‘fixed-effects’). Alternatively, these covariates could be allowed to vary randomly across the upper-level classes, thereby yielding ‘random-effects’, sometimes referred to as random slopes.

In a MLLC model, each lower-level covariate may be constrained to have identical estimated parameter values for each of the upper-level classes. This is the class-independent configuration. Alternatively, this constraint may be relaxed so that the covariate parameters may have different estimated values for each upper-level class. This is the class-dependent configuration. This configuration is akin to random slopes in the traditional MLM, but where

the random effects (represented by a continuous latent variable) are effectively categorised and multiple fixed-effects parameter values are estimated for each upper-level latent class.

Not all covariate effects would necessarily be modelled in this way, so the number of lower-level covariates that are upper-level class dependent could be fewer than the total number of lower-level covariates. This can be much less parsimonious than the traditional MLM, since for the latter, only one continuous latent variable variance is estimated per covariate random slope, as opposed to multiple fixed-effects parameter values for each upper-level class. This is an example of why it becomes necessary to consider carefully the pros and cons of class-dependent versus class-independent covariate effects.

Class sizes

Lower-level class sizes may also be class dependent or class independent, with respect to upper-level classes. The number of lower-level classes per upper-level class is fixed during modelling, but the proportions of each may be either identical (class independent) or different (class dependent) within each of the upper-level classes.

In the class-dependent configuration, some lower-level classes may contain no observations at all, indicating that some upper-level classes might actually favour fewer lower-level classes. This is a discretised version of the traditional MLM approach, with cluster imbalance.

Alternatively, it is possible to constrain class sizes such that the proportion of each lower-level class remains the same for each upper-level class; the class-independent configuration. The total number of observations per upper-level class can still vary, but this configuration forces the upper-level classes to represent the entire spectrum; thus accommodating a structure that, in the specific circumstances considered in this thesis, accounts for differential selection.

Error variance

Appropriate for continuous outcomes only, error variance within lower-level classes may also be class dependent or class independent, in relation to

upper-level classes. The choice should be based both on the distribution of the data used, and on the concepts that are to be explored.

The class-independent configuration constrains the estimates to be the same within each upper-level class, thus ensuring that the variance is held constant across all lower-level classes, i.e. homoscedasticity.

The class-dependent configuration allows for different estimates of the variance of the outcome within each upper-level class, hence permitting unequal variance across the lower-level classes, i.e. heteroscedasticity.

2.2.5 Classification error (CE)

As introduced in sections 2.2.1 and 2.2.2, observations are probabilistically assigned to latent classes at all levels; termed ‘probabilistic assignment’. An alternative way of assigning observations to latent classes is known as ‘modal assignment’, where an observation is allocated to a latent class at each level according to the highest membership probability. The CE is the proportion of observations that are estimated to be misclassified by their modal assignment, and this is usually expressed as a percentage. A CE value is thus observed at both the lower and upper levels.

A small CE indicates that the latent classes are more ‘real’, i.e. they correspond to groups where upper- or lower-level observations are almost entirely assigned to a single class. A smaller CE may be favoured where it results in greater interpretability of the latent classes at any level.

In contrast, a large CE indicates that the latent classes are more ‘virtual’, i.e. a construct of probabilistic assignment only, as they differ substantially from modal assignment. This may result in the identification of additional latent classes that reflect outliers, or unusual but minority (potentially latent) features.

It therefore depends upon the context and purpose of the model as to whether or not one worries about CE values, low or high. It is important to be mindful of the magnitude of CEs, and in some instances models may be preferred where they are not too large, or not too close to zero.

2.2.6 The optimum model

There is no single optimum model; rather the preferred model will differ dependent on context and research question. As introduced in section 2.2.1, model-evaluation criteria (such as the AIC and BIC) are commonly used to aid more parsimonious model selection, as these are penalised versions of the log-likelihood (LL) criterion. Lower values indicate better models but reflect better parsimony compared to LL. For the analyses described in this thesis, all model-evaluation criteria are used for guidance only, with optimum model selection based also on interpretability. Modelling the same data for different research questions may therefore yield differing optimum models. Being able to interpret meaningfully the latent class structure is crucial, as latent class model selection should not be determined solely on likelihood-based statistics. There is, however, a general approach, described here, that may be utilised. Optimum model construction is discussed specifically in relation to each research question in sections 3.2.5, 4.2.4 and 5.3.4.

To consider the construction of an optimum MLLC model using an initially simple approach, the latent class structures may be considered to be built one level at a time. Lower-level observations may first be assigned to latent classes, generating an optimum number of lower-level latent classes as selected by the researcher. Conditional on belonging to a given lower-level class, the upper-level observations may then also be assigned to latent classes based on model configuration, as driven by context and research question.

Within the estimation process, there is no sense of ordering in terms of which level of latent classes are formed ahead of other levels, because this all happens simultaneously. Models are an optimum solution for all classes, at all levels, conditional on covariates considered in the model. Estimation procedures hence seek to maximise the likelihood function in a single process.

In practice, however, a continuous latent variable may be adopted initially at the upper level as an approximation, while the latent class structure is explored for the lower level. Once the optimum number of lower-level classes is determined, the continuous latent variable at the upper level may

be switched to categorical and the optimum latent class structure determined at the upper level. Latent class models are commonly explored where the number of latent classes at all levels are sequentially increased from one to identify the required optimum model, with reference to model-evaluation criteria, parsimony, interpretability and CE.

2.3 The example dataset

2.3.1 Source and extraction

The colorectal cancer data were extracted from the Northern and Yorkshire Cancer Registry and Information Service (NYCRIS) database in November 2006, using geographical boundaries defined at the 2001 census. Figure 2.1 shows the boundaries of the regional cancer registries at this time point; Northern and Yorkshire marks the boundary for these data. At the 2001 census, the total NYCRIS population was around 6.7 million, approximately 12.4% of the total population of England and Wales (NYCRIS, 2007).

Figure 2.1 Locations of the regional cancer registries at the 2001 Census

Each regional cancer registry is responsible for collecting data on all cancer registrations within their geographical boundary, ensuring data are accurate and complete, and analysing and interpreting data for regional reference reports. The registration process is complex, with data obtained from multiple sources including medical records, tumour registers and death certificates, and so there is a time delay between the cancer diagnosis date and the availability of complete data. Prior to 2009, cancer registries were required to provide their registration data to the Office for National Statistics (ONS) within eighteen months following the end of the calendar year (Office for National Statistics, 2016a). Therefore, at November 2006, the latest available calendar year of registration data was 2004.

Diagnosed cases of colorectal cancer (10th revision of the International Classification of Diseases (ICD-10) (World Health Organisation, 2005) codes C18, C19 and C20) were initially identified where the date of diagnosis was between 1 January 1991 and 31 December 2004, and the patient was resident in the Northern and Yorkshire regions. Due to issues concerning the completeness of these data prior to 1998, when NYCRIS merged two smaller cancer registries, diagnoses from 1 January 1998 only were retained.

Data extraction and initial processing to obtain a non-identifiable dataset were performed in advance of commencement of this research activity and are therefore not described here.

Variables available after initial processing were: age at diagnosis, sex, tumour site (either colon (C18), rectosigmoid junction (C19) or rectum (C20)), stage at diagnosis (using Dukes classification (Dukes, 1949); ranging from stage A (early stage) to stage D (late stage)), lower super output area (LSOA) (used to derive SES using the TDI (as described in section 1.2.2), recorded at the 2001 census), whether or not the patient was treated curatively, which hospital(s) were attended and whether the patient was alive or dead at the latest data download date.

The laterality of the tumour was also determined from the tumour site, with rectosigmoid junction and rectal tumours identified on the left side of the body, while colon tumours may present on either side of the body.

2.3.2 Exclusions

Figure 2.2 details all exclusions made to the original extracted dataset containing diagnosis dates from 1991 to 2004.

Three years of follow up data are required, as the outcome is survival status at three years following diagnosis. A new download of death dates was obtained on 30 June 2007, meaning that diagnosis dates only up to 30 June 2004 could be included. As such, patients who are diagnosed after this date are excluded; 1,982 patients are excluded.

Townsend deprivation scores are imported into the dataset by matching to the LSOA of residence of the patient. One patient does not have a recorded LSOA, indicating that they are resident outside the NYCRIS area; this patient is therefore excluded from the dataset.

Multiple tumours are excluded. Clinical information is not available in order to identify whether an additional recorded tumour for a patient is due to spread or recurrence of the original tumour, or if the patient has been diagnosed with a multiple primary tumour (MPT). A MPT occurs where more than one histologically distinct tumour is found in the same patient, and treatment may differ for these differing types of multiple tumour. There are 540 patients diagnosed with between two and four tumours; 561 multiple tumours are therefore excluded.

Patients with a recorded age at diagnosis of more than one hundred years are excluded; 7 patients are excluded.

Patients identified by death certificate only (DCO) are excluded. This occurs where the death certificate has provided the only tumour notification, thus there is no information available regarding, for example, hospital visits or treatment received. DCO registrations are commonly used to measure data completeness (Hill, 1995). Where registration is initially provided by death certificate, the cancer registry attempts to add missing information from other sources, such as GP or hospital records, but if none can be found, registration is classified as DCO. 364 patients are excluded.

Exclusions related to identification of the diagnostic centre are discussed in section 2.3.3.

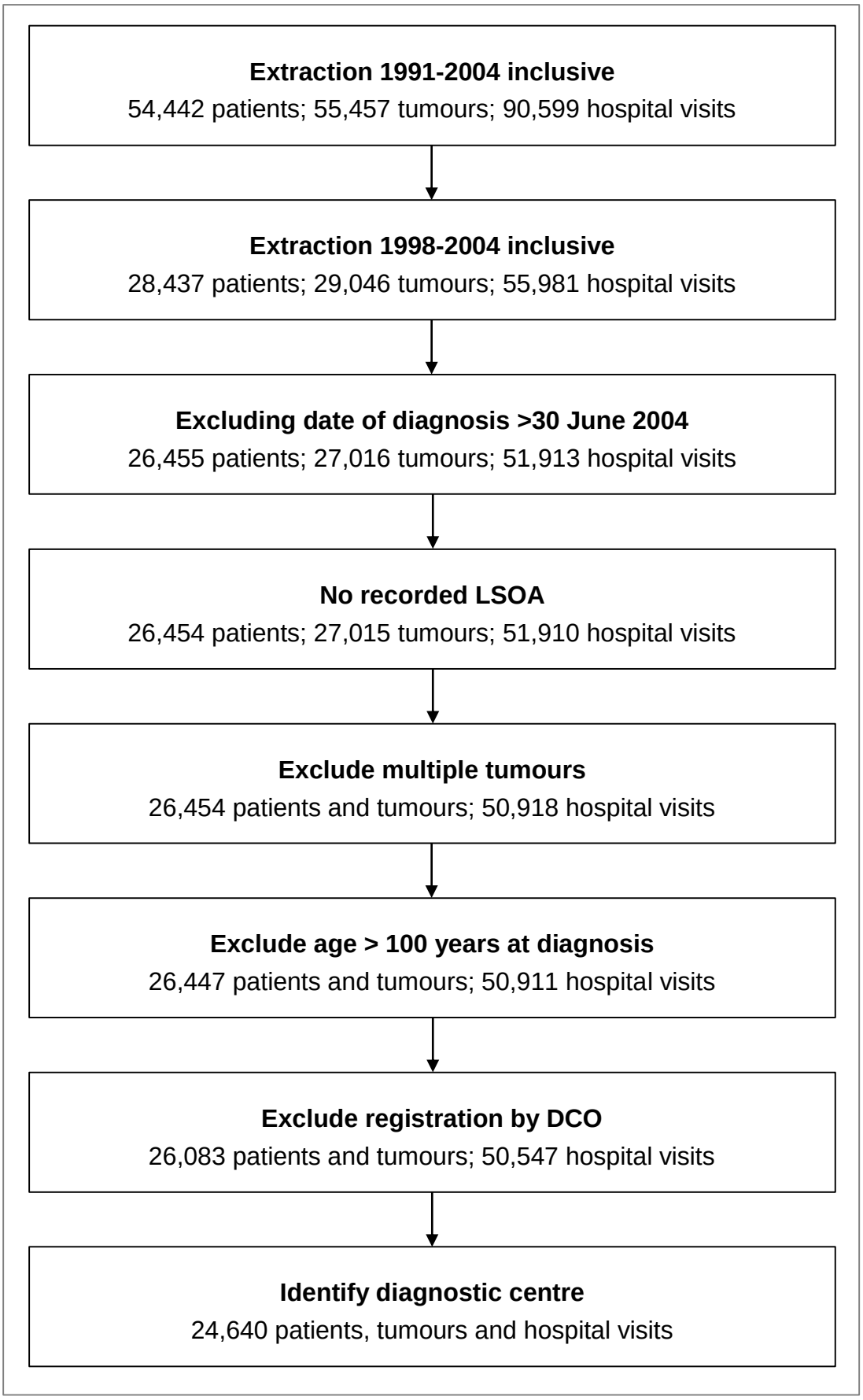


Figure 2.2 Flowchart showing exclusions from the original extracted dataset

2.3.3 Identification of diagnostic centre

While interest may lie in investigating potential treatment centre characteristics associated with colorectal cancer survival, this can be complex to assess, as patients may attend multiple hospital visits at different hospitals throughout their care, and thus may receive treatment at more than one hospital. For these data, patients attend up to seven hospital visits for diagnosis, treatment or specialist opinion in total, with 16,549 patients (63.4%) attending more than once.

Table 2.1 shows the number and percentage of hospital visits per patient where treatment is recorded; patients attend between zero and five treatment visits. Of the 26,083 patients currently in the dataset, most have only one treatment visit (14,437; 55.4%), although 7,323 patients (28.1%) have more than one treatment visit. Of the 7,323 patients, only 650 (8.9%) receive treatment at the same hospital each time, while 6,554 (89.5%) receive treatment at two different hospitals, and 119 (1.6%) receive treatment at three different hospitals. 4,323 patients (16.6% of the 26,083) do not receive curative treatment and would therefore be excluded from any analysis using treatment centres.

Table 2.2 shows the number and percentage of hospital visits per patient where diagnosis is recorded. Although patients may also attend multiple diagnosis visits, there is less variability, with patients attending between zero and three diagnosis visits. Most patients have only one diagnosis visit (25,542; 97.9%), 220 patients (0.8%) have no diagnosis visits and 321 patients (1.2%) have more than one diagnosis visit. Of the 321 patients, 228 (71.0%) receive diagnosis at the same hospital each time, 92 (28.7%) receive diagnosis at two different hospitals and 1 (0.3%) receives diagnosis at three different hospitals. Diagnostic centres are therefore used in order to include all patients, whether treated or not, and to limit the variability introduced by attendance at different hospitals.

Table 2.1 Number and percentage of treatment visits per patient

No. hospital visits with treatment recorded	No. patients (% of total patients)	No. different hospitals attended per patient (% of patients with related number of visits)		
		1	2	3
0	4,323 (16.6%)	N/A	N/A	N/A
1	14,437 (55.4%)	14,437 (100.0%)	N/A	N/A
2-5	7,323 (28.1%)	650 (8.9%)	6,554 (89.5%)	119 (1.6%)
	26,083			

Table 2.2 Number and percentage of diagnosis visits per patient

No. hospital visits with diagnosis recorded	No. patients (% of total patients)	No. different hospitals attended per patient (% of patients with related number of visits)		
		1	2	3
0	220 (0.8%)	N/A	N/A	N/A
1	25,542 (97.9%)	25,542 (100.0%)	N/A	N/A
2-3	321 (1.2%)	228 (71.0%)	92 (28.7%)	1 (0.3%)
	26,083			

The diagnostic centre is initially defined as the hospital where the latest staging took place. For those 25,542 patients with only one diagnosis visit, this is straightforward. For those 321 patients attending more than one diagnosis visit, the location of the most recent diagnosis hospital is recorded at the diagnostic centre, as this is considered to provide the latest staging information. For those 220 patients with no diagnosis visit, the location of their first hospital visit is taken as the diagnostic centre. Although they have no recorded diagnosis visit, all of these patients have a recorded ICD-10 diagnosis code of colorectal cancer.

Each hospital is contained within an NHS Trust, and nineteen Trusts are identified in the NYCRIS geographical area. Trust codes are matched to the 154 hospital codes in order to identify diagnostic centres at the Trust level. Of the 21,760 patients who receive treatment (83.4%), 17,598 (80.9%) are treated initially within the same hospital as they are diagnosed, with only 12,879 (59.2%) remaining within this hospital throughout. This contrasts with 19,368 patients (89.0%) who are treated initially within the same Trust as they are diagnosed, with 16,163 (74.3%) remaining within this Trust throughout. As the figures at the Trust level show improvement from those at the hospital level, the choice is made to analyse by Trust of diagnosis instead of by hospital. Increased movement between centres could introduce variability of care for the patient and thus mitigate the diagnostic centre effect.

1,443 patients are found to be diagnosed at Trusts external to the NYCRIS geographical area and are thus excluded, leaving 24,640 patients available for analysis. While the diagnosis visit remains in the dataset, all other Trust visits are also excluded, leaving just one Trust visit per tumour, per patient, as shown in Figure 2.2.

2.3.4 Descriptive statistics

There are 24,640 patients available in the final dataset for analysis; each patient is diagnosed at one of the nineteen NHS Trusts. Due to coding updates, there are minor differences in the number of deaths recorded in the final datasets adopted for each research question, although this has no material impact on the messages of each.

For analyses in Chapter 3, 12,708 patients (51.6%) died within three years of diagnosis, while for analyses in Chapter 4 (which were initially performed before the material seen in Chapter 3), 12,856 patients (52.2%) died within three years of diagnosis. The number of patients available in the dataset remains the same, and the small differences in deaths do not impact on the demonstration of the utility of the MLLC analysis. Table 2.3 provides summary statistics for all explanatory variables available in the dataset, which remain the same across both versions of the dataset. Trusts are ordered alphabetically and allocated numbers one to nineteen. Differences in values across Trusts demonstrate the heterogeneity of the patient casemix.

All variables contain complete data except for stage at diagnosis. As discussed in section 1.2.5, missing values for stage are categorised (overall 3,223; 13.1% missing (coded ‘X’)). Age at diagnosis is centred on the study mean of 71.5 years, to improve model precision.

Table 2.3 Summary statistics for all explanatory variables in the final dataset

Variable		**Overall N=24,640**	**Trust 1 N=538**	**Trust 2 N=1,028**	**Trust 3 N=1,280**	**Trust 4 N=648**
		Mean (SD)				
Deprivation		-0.04 (3.18)	-1.46 (2.81)	1.31 (4.25)	-0.32 (3.03)	-2.32 (1.63)
Age at diagnosis (years)		71.5 (11.6)	71.8 (11.6)	71.7 (11.9)	72.0 (11.7)	72.8 (11.7)
Variable	**Categories**	**Number (%)**				
Sex	Female	10,862 (44.1%)	249 (46.3%)	469 (45.6%)	577 (45.1%)	300 (46.3%)
	Male	13,778 (55.9%)	289 (53.7%)	559 (54.4%)	703 (54.9%)	348 (53.7%)
Stage at diagnosis	A	2,859 (11.6%)	62 (11.5%)	107 (10.4%)	154 (12.0%)	86 (13.3%)
	B	6,784 (27.5%)	143 (26.6%)	291 (28.3%)	353 (27.6%)	206 (31.8%)
	C	6,173 (25.1%)	174 (32.3%)	279 (27.1%)	271 (21.2%)	179 (27.6%)
	D	5,601 (22.7%)	106 (19.7%)	218 (21.2%)	316 (24.7%)	120 (18.5%)
	Missing (X)	3,223 (13.1%)	53 (9.9%)	133 (12.9%)	186 (14.5%)	57 (8.8%)
ICD-10	C18 (colon)	14,510 (58.9%)	312 (58.0%)	596 (58.0%)	701 (54.8%)	388 (59.9%)
	C19 (rectosigmoid junction)	2,585 (10.5%)	43 (8.0%)	137 (13.3%)	130 (10.2%)	75 (11.6%)
	C20 (rectum)	7,545 (30.6%)	183 (34.0%)	295 (28.7%)	449 (35.1%)	185 (28.5%)
Laterality	Left	16,261 (66.0%)	367 (68.2%)	673 (65.5%)	883 (69.0%)	443 (68.4%)
	Right	6,727 (27.3%)	137 (25.5%)	300 (29.2%)	343 (26.8%)	180 (27.8%)
	Split	1,652 (6.7%)	34 (6.3%)	55 (5.4%)	54 (4.2%)	25 (3.9%)
Treated	Y	20,582 (83.5%)	478 (88.8%)	873 (84.9%)	1,054 (82.3%)	555 (85.6%)
	N	4,058 (16.5%)	60 (11.2%)	155 (15.1%)	226 (17.7%)	93 (14.4%)

Deprivation (measured using TDI) is inversely related to social status.

Table 2.3 continued Summary statistics for all explanatory variables in the final dataset

Variable		Trust 5 N=1,832	Trust 6 N=1,187	Trust 7 N=2,239	Trust 8 N=1,716	Trust 9 N=1,193
		Mean (SD)				
Deprivation		-0.10 (2.88)	-1.77 (2.15)	0.07 (3.25)	0.21 (3.59)	-0.64 (2.90)
Age at diagnosis (years)		71.1 (11.6)	71.6 (11.4)	72.3 (12.1)	71.7 (11.7)	71.7 (11.5)
Variable	**Categories**	**Number (%)**				
Sex	Female	794 (43.3%)	508 (42.8%)	1,022 (45.6%)	731 (42.6%)	537 (45.0%)
	Male	1,038 (56.7%)	679 (57.2%)	1,217 (54.4%)	985 (57.4%)	656 (55.0%)
Stage at diagnosis	A	275 (15.0%)	130 (11.0%)	224 (10.0%)	170 (9.9%)	142 (11.9%)
	B	451 (24.6%)	365 (30.7%)	638 (28.5%)	466 (27.2%)	296 (24.8%)
	C	433 (23.6%)	343 (28.9%)	601 (26.8%)	495 (28.8%)	261 (21.9%)
	D	447 (24.4%)	239 (20.1%)	520 (23.2%)	387 (22.6%)	303 (25.4%)
	Missing (X)	226 (12.3%)	110 (9.3%)	256 (11.4%)	198 (11.5%)	191 (16.0%)
ICD-10	C18 (colon)	1,016 (55.5%)	678 (57.1%)	1,363 (60.9%)	1,035 (60.3%)	654 (54.8%)
	C19 (rectosigmoid junction)	247 (13.5%)	138 (11.6%)	245 (10.9%)	91 (5.3%)	99 (8.3%)
	C20 (rectum)	569 (31.1%)	371 (31.3%)	631 (28.2%)	590 (34.4%)	440 (36.9%)
Laterality	Left	1,270 (69.3%)	778 (65.5%)	1,471 (65.7%)	1,143 (66.6%)	804 (67.4%)
	Right	490 (26.7%)	353 (29.7%)	643 (28.7%)	492 (28.7%)	332 (27.8%)
	Split	72 (3.9%)	56 (4.7%)	125 (5.6%)	81 (4.7%)	57 (4.8%)
Treated	Y	1,535 (83.8%)	1,028 (86.6%)	1,886 (84.2%)	1,480 (86.2%)	979 (82.1%)
	N	297 (16.2%)	159 (13.4%)	353 (15.8%)	236 (13.8%)	214 (17.9%)

Deprivation (measured using TDI) is inversely related to social status.

Table 2.3 continued Summary statistics for all explanatory variables in the final dataset

Variable		Trust 10 N=774	Trust 11 N=661	Trust 12 N=1,567	Trust 13 N=1,937	Trust 14 N=1,258
		Mean (SD)				
Deprivation		-0.97 (2.10)	2.35 (3.18)	-0.03 (3.64)	-0.33 (2.55)	-0.97 (2.45)
Age at diagnosis (years)		72.6 (11.7)	72.0 (10.8)	70.9 (11.4)	72.1 (11.4)	71.6 (11.4)
Variable	**Categories**	Number (%)				
Sex	Female	352 (45.5%)	290 (43.9%)	645 (41.2%)	859 (44.3%)	583 (46.3%)
	Male	422 (54.5%)	371 (56.1%)	922 (58.8%)	1,078 (55.7%)	675 (53.7%)
Stage at diagnosis	A	86 (11.1%)	88 (13.3%)	190 (12.1%)	210 (10.8%)	155 (12.3%)
	B	244 (31.5%)	187 (28.3%)	409 (26.1%)	594 (30.7%)	351 (27.9%)
	C	158 (20.4%)	155 (23.4%)	419 (26.7%)	458 (23.6%)	293 (23.3%)
	D	156 (20.2%)	155 (23.4%)	354 (22.6%)	387 (20.0%)	287 (22.8%)
	Missing (X)	130 (16.8%)	76 (11.5%)	195 (12.4%)	288 (14.9%)	172 (13.7%)
ICD-10	C18 (colon)	481 (62.1%)	374 (56.6%)	883 (56.3%)	1,158 (59.8%)	792 (63.0%)
	C19 (rectosigmoid junction)	88 (11.4%)	90 (13.6%)	183 (11.7%)	239 (12.3%)	131 (10.4%)
	C20 (rectum)	205 (26.5%)	197 (29.8%)	501 (32.0%)	540 (27.9%)	335 (26.6%)
Laterality	Left	496 (64.1%)	466 (70.5%)	1,021 (65.2%)	1,292 (66.7%)	808 (64.2%)
	Right	206 (26.6%)	145 (21.9%)	399 (25.5%)	539 (27.8%)	377 (30.0%)
	Split	72 (9.3%)	50 (7.6%)	147 (9.4%)	106 (5.5%)	73 (5.8%)
Treated	Y	628 (81.1%)	577 (87.3%)	1,307 (83.4%)	1,590 (82.1%)	1,084 (86.2%)
	N	146 (18.9%)	84 (12.7%)	260 (16.6%)	347 (17.9%)	174 (13.8%)

Deprivation (measured using TDI) is inversely related to social status.

Table 2.3 continued Summary statistics for all explanatory variables in the final dataset

Variable		Trust 15 N=1,208	Trust 16 N=2,009	Trust 17 N=1,255	Trust 18 N=771	Trust 19 N=1,539
		Mean (SD)				
Deprivation		1.25 (3.12)	-0.21 (2.35)	0.65 (3.34)	0.89 (3.05)	0.90 (3.63)
Age at diagnosis (years)		70.5 (11.0)	70.9 (11.6)	70.3 (11.4)	72.1 (11.3)	71.3 (11.6)
Variable	**Categories**	**Number (%)**				
Sex	Female	504 (41.7%)	889 (44.3%)	533 (42.5%)	354 (45.9%)	666 (43.3%)
	Male	704 (58.3%)	1,120 (55.7%)	722 (57.5%)	417 (54.1%)	873 (56.7%)
Stage at diagnosis	A	127 (10.5%)	214 (10.7%)	171 (13.6%)	74 (9.6%)	194 (12.6%)
	B	291 (24.1%)	547 (27.2%)	302 (24.1%)	212 (27.5%)	438 (28.5%)
	C	324 (26.8%)	478 (23.8%)	334 (26.6%)	167 (21.7%)	351 (22.8%)
	D	287 (23.8%)	459 (22.8%)	301 (24.0%)	193 (25.0%)	366 (23.8%)
	Missing (X)	179 (14.8%)	311 (15.5%)	147 (11.7%)	125 (16.2%)	190 (12.3%)
ICD-10	C18 (colon)	713 (59.0%)	1,245 (62.0%)	762 (60.7%)	471 (61.1%)	888 (57.7%)
	C19 (rectosigmoid junction)	137 (11.3%)	155 (7.7%)	106 (8.5%)	72 (9.3%)	179 (11.6%)
	C20 (rectum)	358 (29.6%)	609 (30.3%)	387 (30.8%)	228 (29.6%)	472 (30.7%)
Laterality	Left	801 (66.3%)	1,279 (63.7%)	772 (61.5%)	476 (61.7%)	1,018 (66.2%)
	Right	320 (26.5%)	492 (24.5%)	347 (27.6%)	213 (27.6%)	419 (27.2%)
	Split	87 (7.2%)	238 (11.8%)	136 (10.8%)	82 (10.6%)	102 (6.6%)
Treated	Y	969 (80.2%)	1,642 (81.7%)	1,031 (82.2%)	624 (80.9%)	1,262 (82.0%)
	N	239 (19.8%)	367 (18.3%)	224 (17.8%)	147 (19.1%)	277 (18.0%)

Deprivation (measured using TDI) is inversely related to social status.

2.3.5 Patient journey

As defined in the preface to this thesis, the patient journey reflects the progression of a patient through the healthcare system, starting from their first interaction, including all referrals, and ending with completion of their treatment. Figure 2.3 shows a theorised patient journey for patients receiving treatment or care for colorectal cancer, with reference to the clinical pathways produced by the National Institute for Health and Care Excellence (NICE) (NICE, 2017a). Entry points to the healthcare system are outlined in red.

Patients may visit their GP with concern regarding symptoms, and may then be referred for a specialist appointment, which is required to take place within two weeks (NICE, 2017b). Alternatively, patients may receive a positive result following screening. Screening is included to ensure the patient journey reflects current experience, although it was not available in the United Kingdom until 2006, as discussed in section 3.2.2, thus none of the patients in the example dataset described were diagnosed via the screening route. Emergency admission of patients with worsening symptoms may lead to immediate surgery, which may be a risk factor for poor outcomes, as also discussed in section 3.2.2.

Once diagnosed, patients may undergo further investigation to establish the spread of the disease, i.e. to determine the stage of the tumour. Treatment options are then discussed, and a combination of different treatments may be performed, based on both patient and tumour characteristics. Patients are followed up to either monitor tumour recurrence (where curative surgery was performed), or to receive palliative care (where the tumour was considered inoperable).

Not all aspects of the illustrated patient journey are reflected in the example dataset. In this dataset, patients are identified at their diagnosis visit and followed up as they receive specialist advice or treatment. Detailed treatment information is not included, nor do the data reflect ongoing monitoring or palliative care received.

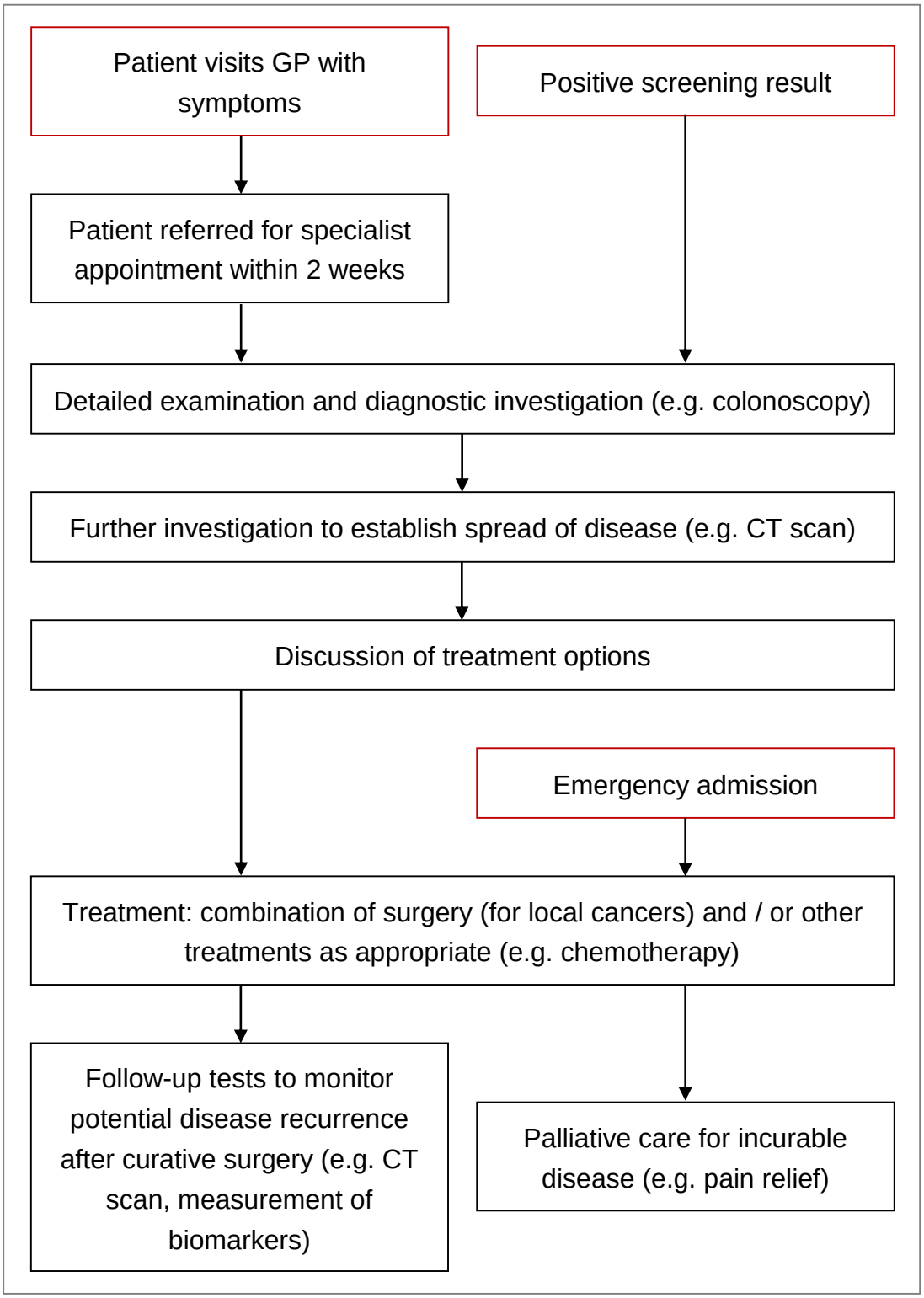


Figure 2.3 Theorised patient journey through the healthcare system for patients receiving treatment or care for colorectal cancer; entry points outlined in red

2.4 Modelling approach to the research questions

2.4.1 Appropriate analytical method

Three research questions were posed in section 1.2.3, representing common enquiries that may be made of observational health data, with each question relating to a different aspect of the patient journey within the healthcare system. The questions are:

(1) What is the relationship between a health exposure and outcome, and what other factors affect this relationship?

(2) How does the performance of a healthcare provider vary after accommodating patient differences?

(3) Can causal provider-level covariate effects be identified, after accommodating patient differences?

Two-level MLLC modelling is utilised to answer these questions, with patients at the lower level of the hierarchy and healthcare providers (i.e. NHS Trusts) at the upper level. Multiple discrete latent classes are therefore identified at both levels. MLLC analysis is supported by an overarching latent variable framework, as introduced in section 1.4.1, that is inherently causal, and that can accommodate the separation of modelling for causal inference and modelling for prediction (i.e. differential selection). This comprehensive approach can thus be retained to answer all three research questions. As introduced in section 2.2.2, patient classes are determined according to similarities in characteristics, while Trust classes are determined either according to similarities or differences, dependent on the research question. This leads to two broad modelling strategies as described in section 2.4.3.

The use of unique features within latent variable methodologies, as described in section 2.2.4.3, can be exploited to ensure a precise model configuration is adopted for each research question. Detailed parameterisations are described in section 2.4.4.

2.4.2 Data challenges

Generic data challenges inherent within observational health data were introduced in section 1.2.2, and the traditional approach to these challenges was discussed in section 1.3.3. Unlike traditional regression approaches, MLLC analysis is fully able to address these challenges, which are discussed here with specific reference to the example dataset.

Structure and non-homogeneity

Data complexity is incorporated through a multilevel structure. Within the example dataset, different groups of patients attend different diagnostic centres (i.e. NHS Trusts), dependent on factors such as their area of residence. Patients are thus clustered within Trusts.

The ability to assign latent classes to subgroups of observations allows for non-homogeneity at both levels of a MLLC model. Assumptions of normality and independence, as required for MLM, are not necessary when discrete latent classes are incorporated in place of continuous latent variables. In the example dataset, neither patients nor Trusts are likely to be homogeneous, for reasons explored in sections 1.2.2 (considering variability in patient and provider characteristics), and 1.2.3 (considering differential selection).

Observed and unobserved variation

While traditional regression approaches cannot accommodate uncertainty in model covariates, latent variable techniques allow for covariates that may be measured with error, or that have missing values, to be modelled as class predictors within the class membership part of the model. They may therefore be removed from the regression part of the model, thus improving precision and minimising bias due to uncertainty or to measurement error.

Within the example dataset, stage at diagnosis suffers from missing values and, although these values are categorised, variation remains due to imprecise classification. Variability in the quality of pathology can lead to patients being classified incorrectly (Quirke and Morris, 2007); classification is thus prone to error. There is also potential bias in the grading of stage as the quality of pathology can sometimes lead to patients being ‘understaged’ (i.e. incorrectly assigned an earlier stage at diagnosis due to unidentified

lymph node metastases) (Morris et al., 2007a). For example, for the tumour to be classified at stage C, lymph nodes must be involved. The number of lymph nodes retrieved, however, is highly variable and if few nodes are available, this limits the likelihood of identifying node involvement, so the tumour may instead be classified at stage B. This has an impact on the treatment received, as patients diagnosed with a stage B tumour may not receive beneficial chemotherapy (Morris et al., 2007b). The recording of stage has also changed somewhat over time. If a tumour is initially graded at stage C, but clinical evidence of metastatic disease is then found, the policy in the NYCRIS region at time of data extraction was to 'up-stage' the tumour to stage D. This may not have occurred in previous years, leading to potential bias. Stage should therefore be included as a class predictor, when modelling for causal inference, in order to minimise bias due to measurement error. Use of a latent variable modelling approach may help to incorporate the additional uncertainty of having a missing stage category.

Considering unobserved variation, the latent constructs of a MLLC model implicitly accommodate unmeasured differences across observations, thereby minimising bias due to unmeasured covariates. The example dataset contains only a small selection of possible variables that may be associated with survival from colorectal cancer. As such, its use within traditional modelling techniques may introduce bias if, for example, matching was undertaken on limited covariates. The MLLC approach does not have this disadvantage.

Complex observed relationships

Variables that may either moderate or mediate the main exposure-outcome relationship may be modelled as class predictors, thus removing them from the regression part of the model and minimising exacerbated bias due to measurement error (i.e. interaction terms are not required for effect modifiers) or due to the reversal paradox. Traditional regression approaches cannot accommodate these complex observed relationships without the risk of invoking bias.

Previous studies investigating the association between survival from colorectal cancer and known potential risk factors, such as age, sex and

SES, have typically considered stage of disease at diagnosis (where available) as a potential confounder, for example Morris et al. (2011) and Downing et al. (2013). However, a higher level of socioeconomic deprivation may result in patients presenting with a more advanced stage at diagnosis (Jones et al., 2008; McPhail et al., 2015), which may also be associated with survival (Morris et al., 2011; Downing et al., 2013). SES therefore causally precedes stage at diagnosis and consequently stage does not qualify as a genuine confounder if causal inference modelling of the SES-survival relationship is required; it is a mediator. As such, if modelling for causal inference is required, stage should be removed from the regression part of the MLLC model.

Covariate relationships are explored in further detail in sections 3.2.1 and 3.2.2.

2.4.3 Broad modelling strategies

Two broad modelling strategies are sought. They are the basis for the construction of detailed modelling configurations that are unique for each research question, yet standard in approach.

(i) Grouping together providers in terms of *similar* patient characteristics

This yields provider-level latent classes that are homogeneous with respect to patient outcome and its relationship with model covariates. The focus is on patients and each provider-level class may contain differing proportions of patient classes; heterogeneity is thus accounted for at the provider level. This strategy allows for the exposure-outcome relationship to be determined at the patient level and is therefore utilised to answer research question (1).

(ii) Grouping together providers in terms of *different* patient characteristics

This yields provider-level latent classes that are heterogeneous with respect to patient characteristics. The focus is on providers and each provider-level class will contain the same proportions of patient classes, i.e. the provider classes are effectively patient casemix ‘adjusted’. These classes must differ with respect to non-patient-level characteristics, however, in order to be

separate classes; differences will be due to provider-level patient outcome differences, which in turn are due to underlying organisational factors. This strategy allows the researcher to assess difference in performance across providers, based on underlying provider-level factors rather than by patient casemix, and is therefore utilised to answer research questions (2) and (3).

2.4.4 Detailed parameterisations

Patient-class intercepts, covariates, class sizes and error variances can be either provider-class dependent or independent, as discussed in section 2.2.4.3. An overview of the specific parameterisations necessary to answer the research questions is given here, with full consideration within the following chapters.

Intercepts. For all research questions, class-independent intercepts are adopted to enable identical contrasts to be made amongst patient classes within provider classes, in a relative sense, i.e. the patient classes with 'best' and 'worst' mortality differ in relative terms identically for each provider class. If class-dependent intercepts were adopted instead, contrasts in survival amongst patient classes in one provider could, relatively speaking, mean different things according to which provider class is considered. In both instances, provider classes may differ in their overall outcome. It is helpful for illustration and ease of interpretation, though not essential, to adopt class-independent model intercepts. In other circumstances (especially for different datasets), class-dependent intercepts may be more appropriate.

Covariate effects. Initially, class-dependent covariate effects are adopted, to allow for random effects, although this may be switched to class independent for parsimony (if there is little evidence that a parameter value varies across the classes), or to avoid causal circularity between a covariate, mediator and outcome. A combination of configurations is possible; parameter estimates may be constrained to take one value over a number of classes and another value over the remaining classes. Although technically possible, a priori knowledge of how the data are generated is essential before utilising such complex model structures.

Class sizes. Modelling strategy (i) requires that patient-class sizes are provider-class dependent, so that the provider classes are grouped by their similarities with respect to patient outcome and its relationship with model covariates. Each provider class may therefore be made up of differing proportions of patient classes. Modelling strategy (ii) requires that patient-class sizes are provider-class independent; thus generating heterogeneous latent classes at the provider level and so accounting for differential selection.

Error variance. This is not applicable to research questions (1) or (2) as a binary outcome only is considered. For research question (3), class-independent error variances are adopted, based on the choices made during the simulation approach (as detailed in section 5.2). Patient classes are thus constrained to be homoscedastic, i.e. the variance of the outcome is fixed across the patient classes.

Class-dependent error variances may be more appropriate for other datasets, for example when utilising an observational dataset where the outcome may be expected to vary differently for different patient subgroups.

Chapter 3

Research Question (1); Focus on Patients

3.1 Introduction

Chapters 1 and 2 explored the overarching latent variable approach to modelling complex observational health data, making contrasts with traditional techniques with respect both to the comprehensive framework and to the generic data challenges. Distinctive aspects of the latent variable methodology were thoroughly examined, including: the use of discrete latent classes at all levels of a hierarchy (e.g. to account for heterogeneity), class predictors (e.g. to minimise bias due to effect modifiers or mediators), inactive covariates (e.g. to aid interpretability without affecting the primary relationship), and class-dependent or class-independent features (e.g. to precisely define model parameter requirements).

Three research questions were presented, reflecting queries commonly made within observational health data, with each question concerning a different aspect of the patient journey within the healthcare system. MLLC modelling was identified to provide a suitable approach to answer all of the research questions; rationale (considering the overarching framework and generic data challenges), broad modelling strategies and detailed parameterisations were presented in Chapter 2, and these specifications will be explored further in Chapter 3, with relevance to research question (1):

(1) What is the relationship between a health exposure and outcome, and what other factors affect this relationship?

The example dataset was also introduced, and explored in detail in Chapter 2, including specific data challenges (as examples of the generic data challenges discussed in section 1.2.2) that must be addressed within the modelling approaches to the research questions: data are hierarchical, with non-homogeneous groups at patient and Trust levels, there is variation due

to measurement error and unmeasured variables, and covariates have a complex observed relationship structure.

The latent variable approach to this research question is to account for heterogeneity at the provider level in order to make causal inference at the patient level. Careful assessment of the relationship between model covariates is essential to ensure appropriate adjustment for confounders. For example, as identified in section 2.4.2, stage at diagnosis is thought to be a mediator of the primary exposure-outcome relationship, and is measured with error.

Certain simplifications are implemented, as discussed in section 1.2.5; a binary outcome variable is utilised (i.e. whether or not the patient survived at three years following diagnosis) instead of a continuous survival measure, and the cross-classified effect of the small-area level is ignored.

Section 3.2 summarises the data and methods relevant to this research question, including construction of a DAG, consideration of related literature, the modelling approach, parameterisation, and optimum model construction. MLM is identified as the traditional comparison; assessment of MLM assumptions can also be made.

Section 3.3 presents all results, starting with the MLM analysis, through MLLC model construction to interpretation of the results for the latent classes at both patient and Trust levels, making appropriate contrasts with MLM.

Section 3.4 provides a discussion of the methods and results.

This chapter contains work based on two publications (Harrison et al., 2012; Harrison et al., 2013).

3.2 Data and methods

3.2.1 Example Dataset

The example dataset described in Chapter 2 is utilised, containing data on 24,640 patients diagnosed with colorectal cancer between 1998 and 2004; 12,708 patients (51.6%) died within three years of diagnosis. A literature search to identify factors that have been shown to impact on survival within this disease area is described in section 3.2.2. Interest lies in the association between SES (measured in these data using the TDI) and three-year mortality; the research question specific to these data is thus:

(1) What is the relationship between SES and three-year mortality from colorectal cancer, and what other factors affect this relationship?

As identified in section 2.4.2, these data are hierarchical with patients at the lower level and NHS Trusts at the upper level, and neither patient nor Trust groups are likely to be homogeneous. Stage at diagnosis suffers from missing data, imprecise classification, and may also lie on the causal path between SES and survival.

A DAG, as introduced in section 1.2.3, is essential to assess the key covariate relationships. There are, however, many ways to construct a DAG, based on differing theorised relationships between covariates; thus, a range of alternatively specified DAGs are presented in figures 3.1, 3.2 and 3.3.

Figure 3.1 illustrates a simplified approach. Stage at diagnosis is considered to lie on the causal path between SES and survival (represented in these data as three-year mortality); stage is therefore operating as a mediator of the exposure-outcome relationship. If included as a covariate in the regression part of a model, bias may be introduced due to the reversal paradox, as discussed in section 1.3.3. Age at diagnosis and sex are shown as competing exposures, with stage at diagnosis also potentially mediating their relationships with survival.

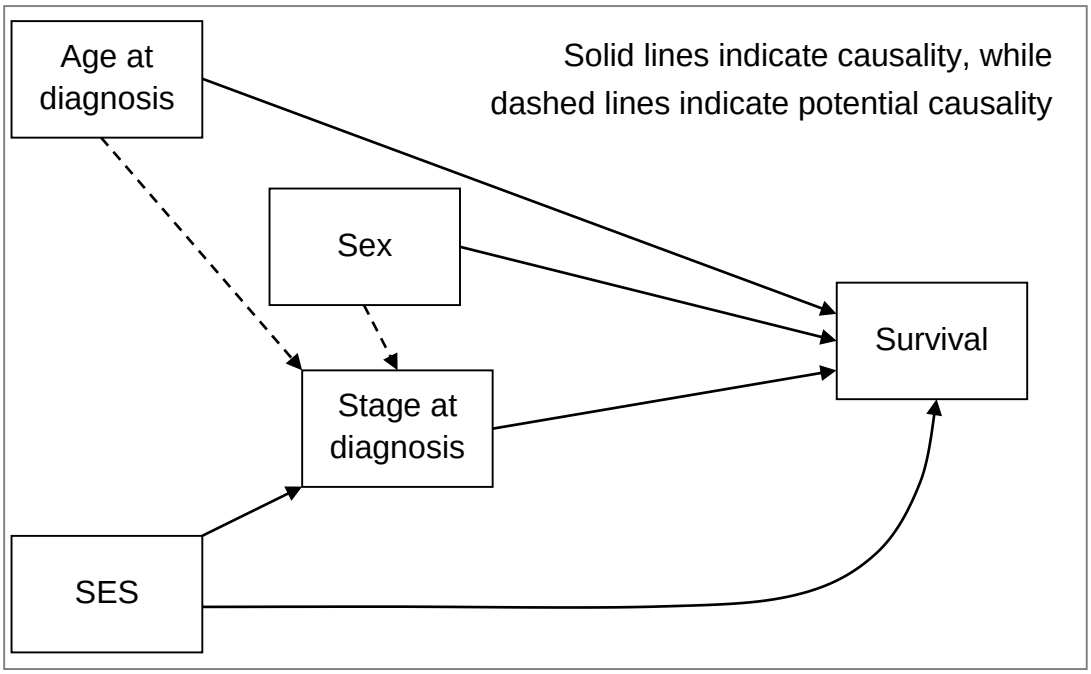


Figure 3.1 DAG (1) showing the inferred causal relationships amongst key variables at the population level

Figure 3.2 also includes whether or not the patient receives treatment (curative only, in these data), which is dependent upon stage at diagnosis, as discussed in section 2.4.2. Treatment may then also affect survival, as will be described in section 3.2.2.

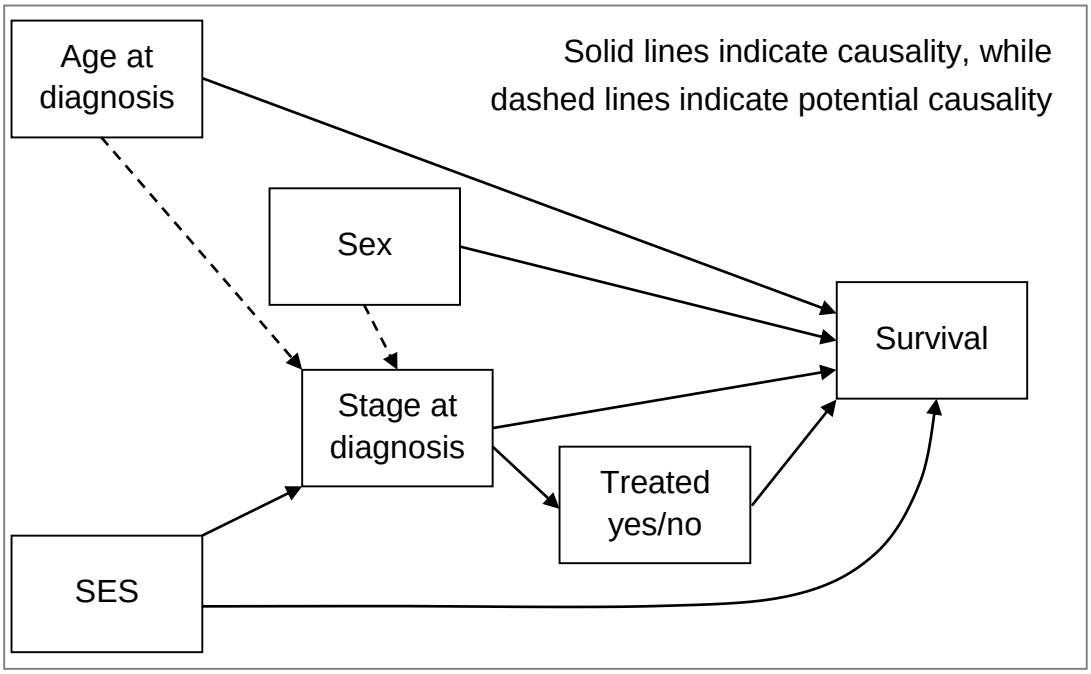


Figure 3.2 DAG (2) showing the inferred causal relationships amongst key variables at the population level

Figure 3.3 additionally shows a causal relationship between age at diagnosis and SES. Age at diagnosis is a complex measurement, as it comprises risks due both to the time period within which the patient was born and to the patient's age at which the tumour was diagnosed; differences in age at diagnosis have been seen to impact upon socioeconomic inequalities in colorectal cancer survival (Nur et al., 2015).

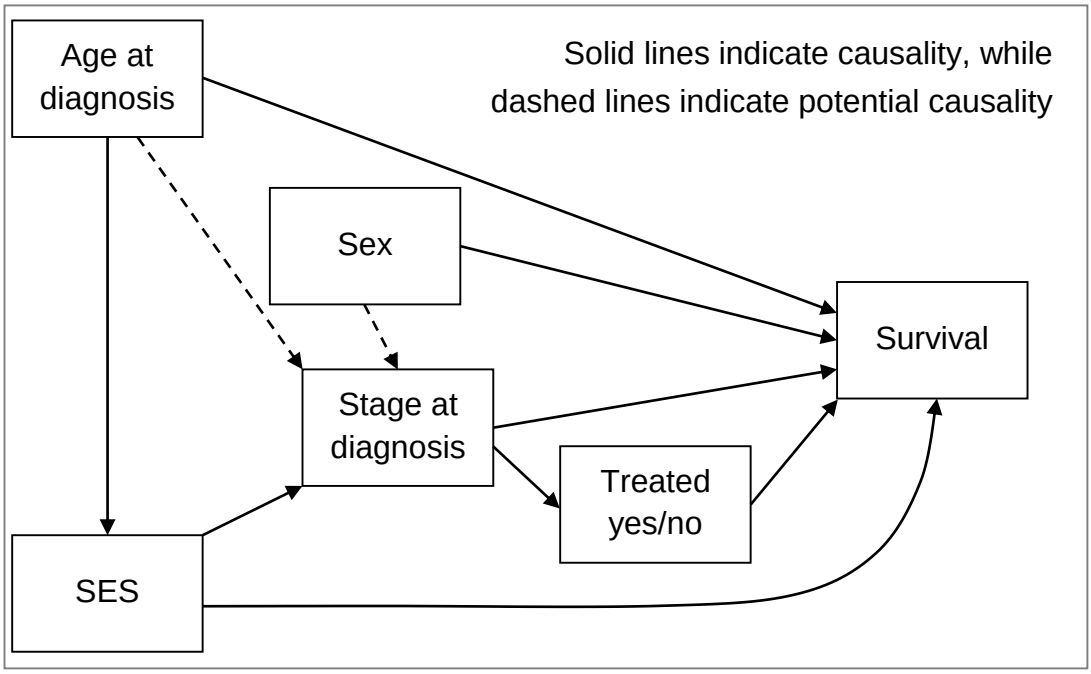


Figure 3.3 DAG (3) showing the inferred causal relationships amongst key variables at the population level

While any one of many DAGs, including those illustrated, may be appropriate for these data, the DAG shown in figure 3.1 is chosen for the purposes of this analysis, to simplify the inferred covariate relationships and thus demonstrate the utility of the latent variable techniques.

3.2.2 Literature review

A literature review is performed to assess risk factors associated with survival (or mortality) from colorectal cancer, with specific focus on the relationship between socioeconomic deprivation (quantified in these data as SES), stage and survival. Much research has been performed within this field and this search is not designed to cover it all, but will instead highlight the key findings and the methodological approach commonly taken when

answering this sort of research question. There are two parts to the search: (i) determination of the risk factors associated with survival from colorectal cancer; and (ii) exploration of the specific association between deprivation and stage at diagnosis (regardless of health outcome or other risk factors). The results are combined and summarised.

Medline is selected as the most appropriate database, and a combination of keyword searching and medical subject headings (MeSH) are used. The full literature search strategy can be seen in Appendix B. Consideration of colorectal cancer, survival and risk factors yields 17,853 results while focus on socioeconomic deprivation and stage at diagnosis yields 343; these are combined to yield 18,018 articles. A restriction to UK articles was thus applied, and will ensure generalisability to the example dataset. There is no expectation that risk factors will differ across countries, although there may be differences in how individuals within the countries are affected by these factors (Ait Ouakrim et al., 2015). With further limitations to include only articles with abstracts, in English, concerning humans and published within the last ten years, 263 results were found. Excluding duplications, 247 articles remain for consideration. Abstracts were initially reviewed for relevance, and additional articles were sourced from citations.

Screening for colorectal cancer was introduced into the United Kingdom in 2006, with nationwide availability by 2010, and all individuals aged over sixty are eligible (over fifty in Scotland). Uptake is around 54%, on average (von Wagner et al., 2011); and lower uptake is associated with younger age, male sex and a higher level of deprivation (Mansouri et al., 2013). There is also a low uptake in ethnically diverse areas (von Wagner et al., 2011), and for obese individuals (Beeken et al., 2014). Since its introduction, the screening programme has led to earlier stage diagnoses (Morris et al., 2012; Logan et al., 2012; Rees and Bevan, 2013), and patients diagnosed with tumours detected by screening have been seen to have better overall survival rates (Morris et al., 2012; Gill et al., 2014) compared with patients with non-screen-detected tumours, and after adjustment for stage at diagnosis. Further, screen-detected tumours are more likely to be treated curatively (Morris et al., 2012). So far, screening is estimated to have reduced mortality due to colorectal cancer by 18% (McClements et al., 2012).

Diagnosis or treatment centre may have an impact on survival, due to the volume of procedures performed, although study findings vary. Overall survival rates have been seen to improve when considering larger-volume hospitals (Borowski et al., 2010). Some studies observe improved postoperative mortality for high-volume surgeons in elective surgery (Borowski et al., 2007; Borowski et al., 2010), while others find no association (Burns et al., 2013), nor is there an association seen for emergency surgery (Borowski et al., 2007; Faiz et al., 2010).

Specialist surgical treatment improves survival from elective surgery both post-operatively (Brewster et al., 2011; Oliphant et al., 2013a; Oliphant et al., 2014b) and within five years (Oliphant et al., 2013a; Oliphant et al., 2014b), with those treated by a specialist more likely to undergo surgery in a high-volume hospital (Oliphant et al., 2014b).

Medication may also have an impact, with use of statins (Cardwell et al., 2014) or aspirin (Walker et al., 2012; McCowan et al., 2013) seen to reduce mortality. Vitamin D use is inconclusive (Zgaga et al., 2014; Jeffreys et al., 2015).

Regarding socio-demographic factors, there is a clear relationship between older age and higher rates of mortality, with older patients more likely to die within thirty days of diagnosis (Brewster et al., 2011; Downing et al., 2013; McPhail et al., 2013) (especially if they do not undergo an operation (Sheridan et al., 2014)), within thirty days of operation (Widdison et al., 2011; Faiz et al., 2011; Morris et al., 2011; Ahmed et al., 2014), and longer term (Faiz et al., 2011; Ahmed et al., 2014). Elderly patients are also more likely to present as an emergency (McPhail et al., 2013; Downing et al., 2013; Oliphant et al., 2014a), which is itself a risk factor for death in the early post-operative period (Brewster et al., 2011; Morris et al., 2011; Oliphant et al., 2014a; Askari et al., 2015), longer term (Oliphant et al., 2014a), and particularly so for an elderly population (Ihedioha et al., 2013). Patients presenting as emergencies are also more likely to receive non-specialist surgery (Oliphant et al., 2014a). Rates of emergency surgery are also decreasing in the screening age group (Hwang et al., 2014).

There are gender disparities in survival, with improved survival for younger women compared with younger men (Koo et al., 2008; Hendifar et al., 2009), and greater post-surgery mortality for males (McArdle et al., 2003). There are suggestions that tumour progression may be slowed by exposure to oestrogen (Arem et al., 2015), although consideration is also given to the discrepancy between male and female participation in screening (Mansouri et al., 2013).

Ethnic minorities have longer diagnostic and referral intervals (Martins et al., 2013), although referral route itself has not been shown to have an impact on survival (Zafar et al., 2012; Schneider et al., 2013). While surgical delays may affect survival (Nachiappan et al., 2015), the effect is not linear (Redaniel et al., 2014). South Asians, however, are seen to have an overall reduced mortality compared with all other ethnicities (Maringe et al., 2015).

The relationship between socioeconomic deprivation and survival varies. Some studies report increasing deprivation as a significant predictor of mortality, after adjusting for other factors, post-operatively (Morris et al., 2011), in the first year following diagnosis (Downing et al., 2013), and longer term (Lejeune et al., 2010); others find this effect only in univariable analyses (Smith et al., 2006; Bharathan et al., 2011; Brewster et al., 2011; Oliphant et al., 2013b; Oliphant et al., 2014a), while others find no effect at all (Nur et al., 2008; McMillan and McArdle, 2009; Nicholson et al., 2011). Patients living in more deprived areas are more likely to present for emergency surgery rather than elective (Bharathan et al., 2011; Oliphant et al., 2013b; McPhail et al., 2013) (although they are also more likely to have a specialist surgeon (Oliphant et al., 2013b)), to receive palliative treatment rather than curative (Bharathan et al., 2011; Oliphant et al., 2013b; Paterson et al., 2014), and to present with a more advanced stage at diagnosis (Jones et al., 2008; McPhail et al., 2015). Higher levels of socioeconomic deprivation are also associated with more adverse comorbidities (Bharathan et al., 2011; Oliphant et al., 2013b) and longer lengths of stay in hospital (Smith et al., 2006).

Socioeconomic deprivation may reflect how patients vary in terms of lifestyle factors such as diet and smoking (Davy, 2007; Macdonald et al., 2007).

Smoking is associated with increased mortality from colorectal cancer for males (Morrison et al., 2011), and there is some evidence that a healthy diet can improve survival (Norat et al., 2015). Factors such as body mass index (BMI), diabetes, blood pressure and physical activity are shown to have no effect on survival in a large-scale cohort study on males (Morrison et al., 2011), nor is an impact seen for BMI in a large-scale cohort study on females (Reeves et al., 2007). Diabetes, however, is associated with increased all-cause mortality after five years following a diagnosis of colon cancer (Walker et al., 2013).

Worse survival outcomes are seen for a more advanced stage of disease at diagnosis, with associations for early deaths following diagnosis (Brewster et al., 2011; Downing et al., 2013; McPhail et al., 2013; McPhail et al., 2015), particularly within an older population (Sheridan et al., 2014), for longer-term mortality (Nur et al., 2008; McMillan and McArdle, 2009), and for post-operative mortality (Morris et al., 2011; Nicholson et al., 2011; Ihedioha et al., 2013). Some studies do not report an effect of stage, however, after adjusting for other risk factors (Smith et al., 2006; Bharathan et al., 2011; Brewster et al., 2011; Oliphant et al., 2013b; Oliphant et al., 2014a). Late stage disease has also been linked to a higher likelihood of emergency surgery (McPhail et al., 2013; Askari et al., 2015). Differences in stage at diagnosis partly explain international differences in survival rates (Maringe et al., 2013).

Of the studies that account specifically for both socioeconomic deprivation and stage, none make explicit mention of, or accommodation for, potential mediation. Most include both deprivation and stage (among other variables) within a multivariable regression model (Smith et al., 2006; Nur et al., 2008; Lejeune et al., 2010; Morris et al., 2011; Bharathan et al., 2011; Brewster et al., 2011; Downing et al., 2013; Oliphant et al., 2013b; Oliphant et al., 2014a), while others exclude deprivation from multivariable analysis due to statistical non-significance within univariable analysis (McMillan and McArdle, 2009; Nicholson et al., 2011). Smith et al. (2006) recognise that the effect of deprivation is mediated on inclusion of tumour grade into the model, however no modifications are made to the analysis. Brewster et al. (2011) reflect that further research is required to determine any mediating effects

between deprivation and early death, but do not consider mediation within the analysis.

Only three of these studies used multilevel analysis (Smith et al., 2006; Morris et al., 2011; Downing et al., 2013), considering postcode districts, NHS Trusts, or cancer registries as the upper level for analysis. The remainder do not account for the potential multilevel structure of the data.

3.2.3 MLLC approach to the data

As discussed in section 2.4.1, MLLC modelling is the preferred analytical method to answer the research questions. This approach lies within the overarching latent variable framework, which allows modelling both for causal inference and for prediction at all levels of a hierarchy. For research question (1), causal inference is required at the patient level to determine the relationship between SES and survival, while variation at the Trust level is essentially 'nuisance', i.e. heterogeneity must be accounted for, but no inference is required. Broad modelling strategy (i), introduced in section 2.4.3, is therefore utilised. Thus, both patient and Trust classes are grouped together in terms of similar patient characteristics, and latent classes at both levels are therefore homogeneous with respect to the relationship between patient outcome and model covariates. In this manner, the exposure-outcome relationship may be determined within the patient-level classes, while heterogeneity is accommodated at the provider level.

The variables available for analysis within the example dataset are previously summarised in table 2.3, while the theorised covariate relationships are shown in the DAG in figure 3.1. Table 3.1 reiterates the available variables and specifies which are included, and how they are modelled, within the MLLC analytical approach to this research question.

As shown in table 3.1, the relationship between SES and survival is investigated within the regression part of the model, with adjustment for sex and age at diagnosis (centred around the study mean of 71.5 years to improve model precision). An age-squared term is also included as age is found to have a non-linear relationship with survival; the inclusion of age-squared allows for an adjustment to the linear effect of age and hence

additional accommodation of the non-linear relationship of age as a competing exposure.

Table 3.1 Variables included in analysis for research question (1)

Variables available for analysis	Variables included in analysis	Modelling approach
Deprivation	Deprivation	Regression
Sex	Sex	Regression
Age at diagnosis	Age at diagnosis	Regression
	Age-squared	Regression
Stage at diagnosis	Stage at diagnosis	Class predictor
ICD-10	ICD-10	Inactive covariate
Laterality	Laterality	Inactive covariate
Treated	Treated	Inactive covariate

Deprivation is a measure of SES, measured in these data using TDI.

Stage at diagnosis (coded A to D for increasing severity and missing values coded X), as a mediator of the primary relationship, is instead included as a class predictor. This removes stage from the regression part of the model and hence minimises bias due to the reversal paradox. As stage is also measured with error, its inclusion as a class predictor, rather than as a standard covariate, also avoids any exacerbated bias due to product interaction terms. The latent constructs may also incorporate the additional uncertainty due to missing data within the stage variable, although as discussed in section 1.2.5, methods to address missing data should generally be utilised (although not the focus of this thesis). The resultant latent classes may thus be identified by categories of stage, for example ‘early’ or ‘late’ stage disease at diagnosis.

The ICD-10 diagnosis code for the tumour, its laterality (position in the body), and whether or not curative treatment is received are included as inactive covariates, in order to examine their correlation with stage of disease, but to remove them from the SES-survival relationship.

3.2.4 Parameterisation

Detailed parameterisations, introduced in section 2.4.4, are summarised here with respect to research question (1).

Intercepts. Patient-class intercepts are designated class independent with respect to Trust classes; identical contrasts can therefore be made amongst patient classes regardless of which Trust class is considered. Therefore, the relative difference between patient classes with, for example, the 'best' and 'worst' mortality, remains constant across Trust classes.

Covariate effects. For SES, patient-class effects are designated class independent with respect to Trust classes, to avoid the causal circularity that may be introduced by means of the relationship between SES, stage at diagnosis, and three-year mortality (see section 2.2.4.1). Thus, SES has the same parameter value for each Trust class, and hence the SES-survival relationship is constrained to be the same within each patient class.

For all other covariates, this constraint is relaxed initially and thus patient-class effects are designated class dependent. If, however, parameter values are not seen to vary across the patient classes, this may be switched to class independent for parsimony.

Class sizes. Patient-class sizes are designated class dependent with respect to Trust classes, as required for modelling strategy (i), thus accommodating heterogeneity at the Trust level. Trust classes may therefore comprise differing proportions of each patient class.

Error variance. This is not applicable for a binary outcome.

3.2.5 Optimum model

Optimum model construction follows the process suggested in section 2.2.6, whereby a continuous latent variable is initially adopted at the Trust level while the preferred number of latent classes are identified at the patient level. Log-likelihood statistics (LL, BIC and AIC) and CE are assessed for guidance, and the optimum number of patient classes is chosen with consideration of both parsimony and interpretability. The continuous latent

variable at the upper level is then switched to categorical to identify the preferred number of latent classes at the Trust level. The same criteria are again considered in selection of the optimum number of Trust classes.

Model-evaluation statistics determined during model construction are presented in section 3.3.3.

3.2.6 Bootstrapping

Bootstrapping is a sampling technique, where random samples are drawn from a population and similarly modelled to assess the variability around an estimate. A useful introduction is provided by Efron and Tibshirani (1993).

200 bootstrapped datasets are generated, with replacement and with samples selected from within each Trust. Each is similarly analysed using the chosen MLLC model in order to generate 95% confidence intervals (CIs) for the model summary statistics and the model class profiles at both the patient and the Trust levels.

Model summary statistics (size and mortality statistics) are calculated for each of the bootstrapped datasets, and CIs are generated using percentile confidence intervals (2.5% to 97.5%).

For model class profiles, means or proportions (as appropriate) are calculated for each of the model covariates (SES, age at diagnosis and sex) and each of the class predictors (stage at diagnosis, whether treated, tumour site and laterality), based on their probabilistic assignment to each class. CIs are then calculated in the same manner as for the model summary statistics.

CIs for the model covariates in the regression part of the model at the patient level are determined directly from MLLC analysis of the example dataset.

3.2.7 Traditional comparison

MLLC modelling is compared with a traditional MLM approach, introduced in section 1.3.2, where hierarchical data are modelled with patients at the lower level and Trusts at the upper level. Within MLM, a continuous latent variable is incorporated at the Trust level, and parametric assumptions are made: the variation surrounding both intercepts and slopes is assumed to be normally

distributed, and independent of the variation in the individual measurements. As discussed in section 1.3.3, this may not be tenable in observational health data as neither patients nor Trusts are randomly assigned. These upper-level assumptions may also be evaluated, as model results will indicate whether or not a continuous latent variable at the Trust level is sufficient to model appropriately these data.

A single patient class is used within MLM therefore heterogeneity at the patient level cannot be incorporated; the same model is therefore applied to all patients and to all Trusts. Again, model results will indicate whether or not this is sufficient for these data.

SES, age at diagnosis, age-squared and sex are included as covariates in the regression model. Stage at diagnosis cannot be included, for reasons explored in section 3.2.3, and MLM does not have the capacity to model covariates as class predictors. The direct comparison between methods therefore is between use of MLLC modelling, including stage at diagnosis as a class predictor, versus use of MLM excluding stage entirely.

MLM analysis is performed on the original example dataset only, i.e. no bootstrapped datasets are reanalysed. Interest lies in the comparison between estimates of the effect of model covariates, and CIs are generated for the covariate estimates directly from analysis. Bootstrapping would provide CIs for the model statistics only in the MLM analysis, and so is not performed.

3.3 Results

3.3.1 Outline

Results are illustrated for both the MLM and MLLC analysis approaches. Table 3.2 summarises the variables contained within each model. Within the MLM, all variables are included as covariates within the regression model. For the MLLC model, variables may be included either as covariates with the regression model, as class predictors, or as inactive covariates, as discussed in section 3.2.3.

Table 3.2 Comparison of variables included in MLM and MLLC model

Variables included in analysis		
MLM	**MLLC**	**MLLC modelling approach**
Deprivation	Deprivation	Regression
Sex	Sex	Regression
Age at diagnosis	Age at diagnosis	Regression
Age-squared	Age-squared	Regression
-	Stage at diagnosis	Class predictor
-	ICD-10	Inactive covariate
-	Laterality	Inactive covariate
-	Treated	Inactive covariate

Deprivation is a measure of SES, measured in these data using TDI.

3.3.2 MLM analysis

Table 3.3 shows the results of the traditional MLM analysis, with a single patient class and a continuous latent variable at the Trust level. incorporating SES (measured in these data using the TDI), sex, age at diagnosis and age-squared as covariates in a multilevel regression model. Analysis is performed on the example dataset only, hence no CIs are available for model statistics.

Table 3.3 Results from MLM analysis; odds of death within three years

Model Statistics	**Mortality**
Overall	51.6%
Reference Group	49.3%
Model Covariates	**OR of death within three years (95% CI)**
Deprivation (per SD more)	1.18 (1.15, 1.21)
Female	0.87 (0.83, 0.92)
Age (per 5 years older)	1.31 (1.30, 1.33)
Age squared (per 5 years older)	1.006 (1.005, 1.007)

OR – Odds Ratio, CI – Confidence Interval, SD – Standard Deviation; LL = -16,081; Deprivation (measured using TDI) is inversely related to social status.

Overall 12,708 patients (51.6%) died within three years. The reference group comprises males of mean age (71.5 years), diagnosed with stage A colorectal cancer and attributed a Townsend deprivation score of zero.

Substantial and statistically significant associations are found between increasing deprivation and increased odds of death (OR=1.18, 95% CI 1.15 to 1.21 per SD increase in Townsend deprivation score); between female gender and decreased odds of death (OR=0.87, 95% CI 0.83 to 0.92); between increasing age and increased odds of death (OR=1.31, 95% CI 1.30 to 1.33 per 5-year increase in age); and between increasing age-squared and increased odds of death (OR=1.006, 95% CI 1.005 to 1.007). All covariates included in the analysis are identified as competing exposures as per the DAG described in figure 3.1 of section 3.2.1.

3.3.3 Building the MLLC model

As discussed in section 3.2.5, a continuous latent variable is initially adopted at the Trust level in order to determine the optimum number of patient-level classes. Table 3.4 summarises the model-evaluation criteria for the MLLC models with a continuous latent variable at the Trust level.

Table 3.4 Model-evaluation criteria for the patient classes in the MLLC models with a continuous Trust-level latent variable

Patient Classes	LL	BIC	AIC	No. of Parameters	Patient CE
1 class	-16,081	32,213	32,173	5	0.0%
2 classes	-12,122	24,396	24,275	15	8.8%
3 classes	-11,985	24,222	24,019	25	23.2%
4 classes	-11,975	24,305	24,021	35	33.2%
5 classes	-11,966	24,386	21,021	45	32.1%

LL – Log Likelihood, BIC – Bayesian Information Criterion, AIC – Akaike Information Criterion, CE – Classification Error.

Figure 3.4 displays the change in -2LL as the number of patient classes are increased.

Figure 3.4 -2LL plot to determine the optimum number of patient classes in the MLLC modelling approach

This approach suggests that three patient classes are optimum by both the BIC and AIC, while the LL shows model fit improving as the number of patient classes increase, as would be expected with increased model complexity and no penalty to invoke parsimony. After marked improvement in the LL from one to two patient classes, there is little further improvement for increased numbers of patient classes. Considering all other model-evaluation criteria and model interpretation (to distinguish patient effects), three patient classes are selected.

CE at the patient level is 23.2% for the three patient classes, which suggests that just under a quarter of observations are split across the classes, when considering probabilistic assignment, rather than these patients being assigned predominantly to a single patient class.

The continuous Trust-level latent variable is then switched to categorical in order to determine the optimum number of Trust classes; three classes remain fixed at the patient level. Table 3.5 summarises the model-evaluation criteria for the MLLC models with a categorical latent variable at the Trust level, and three latent classes at the patient level.

Table 3.5 Model-evaluation criteria for the Trust classes in the MLLC models with a categorical Trust-level latent variable; three patient-level latent classes

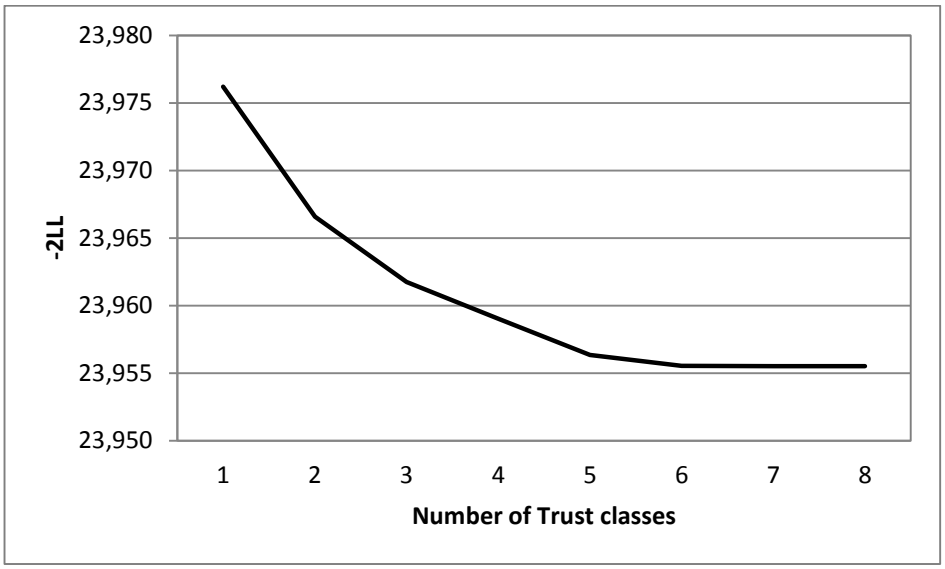
Trust Classes	LL	BIC	AIC	No. of Parameters	Patient CE	Trust CE
1 class	-11,988	24,209	24,022	23	22.7%	0.0%
2 classes	-11,983	24,240	24,021	27	23.2%	10.6%
3 classes	-11,981	24,275	24,024	31	23.1%	10.1%
4 classes	-11,980	24,313	24,029	35	23.9%	12.9%
5 classes	-11,978	24,351	24,034	39	23.2%	17.8%
6 classes	-11,978	24,390	24,042	43	24.1%	21.4%
7 classes	-11,978	24,431	24,050	47	24.1%	30.5%
8 classes	-11,978	24,471	24,058	51	24.1%	36.5%

LL – Log Likelihood, BIC – Bayesian Information Criterion, AIC – Akaike Information Criterion, CE – Classification Error.

This approach suggests that one Trust class is optimum by the BIC, while two Trust classes are just optimum by the AIC. The LL shows improved model fit as the number of Trust classes are increased, as anticipated, although this is a more gradual improvement than that seen for the patient classes. More than one Trust class is required at the Trust level to explain Trust differences therefore further assessment is made of the LL.

Figure 3.5 displays the change in -2LL as the number of Trust classes are increased. The -2LL value continues to improve up to that for five Trust classes.

Figure 3.5 -2LL plot to determine the optimum number of Trust classes in the MLLC modelling approach



The traditional MLM approach with a continuous latent variable at the Trust level, extended to include three patient classes, shows a LL of -11,985, as seen in table 3.4. Results in table 3.5 show that this figure is surpassed by using two Trust classes. The -2LL plot, however, shows a gradual improvement in model fit with increasing numbers of Trust classes, although there is a suggestion that after five Trust classes, the improvement in model fit is minimal. Again considering parsimonious model-evaluation criteria and model interpretability (to model Trust variability and to improve patient-class estimates), the model with five Trust classes is chosen.

CE at the patient level is unchanged from that seen with a continuous latent variable at the Trust level (23.2%), while CE at the Trust level is 17.8%. There is little concern regarding the value of the Trust CE, as upper-level classes are constructed primarily to account for heterogeneity at the Trust level, and thus improve estimates at the patient level.

3.3.4 Patient classes

Table 3.6 summarises the model summary statistics for the patient classes for the chosen three-patient, five-Trust-class MLLC model, where patients are apportioned to one of three groups, labelled post-hoc as ‘good prognosis’, ‘reasonable prognosis’, or ‘poor prognosis’.

Table 3.6 Model summary statistics for the patient classes in the three-patient, five-Trust-class MLLC model

Model Summary Statistics	Good Prognosis	Reasonable Prognosis	Poor Prognosis
	% patients (bootstrapped 95% CI)		
Class size	38.2 (30.0, 48.9)	27.6 (20.8, 38.2)	34.2 (23.7, 37.0)
Overall mortality	9.4 (2.2, 17.4)	58.3 (49.3, 72.9)	93.2 (92.0, 99.6)
Reference group mortality	8.0 (0.1, 16.5)	57.8 (36.7, 78.6)	94.1 (90.8, 100.0)

CI – Confidence Interval; the reference group comprises males, aged 71.5 years, classified as Stage A at diagnosis and attributed a Townsend deprivation score of zero; CIs from bootstrapping calculated using percentiles.

The good prognosis class contains 38.2% of cases of which 9.4% died within three years, compared with the reasonable prognosis class with 27.6% of cases of which 58.3% died within three years, and the poor prognosis class with 34.2% of cases of which 93.2% died within three years.

Tables 3.7 and 3.8 are to be interpreted together. Table 3.7 summarises the model covariate results for the patient classes in the same model, while table 3.8 summarises the mean (for Townsend deprivation score and age) or proportional values (for female gender) by patient class to aid the interpretation of covariate relationships with three-year mortality.

The effect of SES is constrained to take the same value across all patient classes, as discussed in section 3.2.4, in order to avoid the causal circularity between SES, stage at diagnosis, and three-year mortality. SES is therefore clearly associated with increased odds of death (Townsend deprivation score OR=1.33, 95% CI 1.26 to 1.41) for all patient classes. Mean deprivation scores differ somewhat across the classes, with negative values indicating greater affluence while positive values indicate greater deprivation. Patients in the poor prognosis class generally live in more deprived areas (mean 0.09, 95% CI 0.00 to 0.16), compared with patients in the good prognosis class, who generally live in more affluent areas (mean -0.17, 95% CI -0.23 to -0.10).

The impact of sex differs substantially across the classes. In the good prognosis class, females have significantly decreased odds of death compared with males (OR=0.59, 95% CI 0.40 to 0.87), while in the reasonable and poor prognosis classes the association is less clear (reasonable prognosis OR=0.88, 95% CI 0.64 to 1.21; poor prognosis OR=1.05, 95% CI 0.83 to 1.32). The proportions of females differ somewhat across the classes, with fewer females in the poor prognosis class (42.7%, 95% CI 41.1% to 44.1%) compared with the good and reasonable prognosis classes (good prognosis 44.0%, 95% CI 43.1% to 44.9%; reasonable prognosis 45.9%, 95% CI 44.4% to 47.8%), indicating that the majority of females have a decreased risk of death compared with males.

Table 3.7 Model covariate results for the patient classes in the three-patient, five-Trust-class MLLC model

Model Covariates	Good Prognosis	Reasonable Prognosis	Poor Prognosis	Wald p-value
	OR of death within three years (95% CI)			
Deprivation (per SD more)	1.33 (1.26, 1.41)	1.33 (1.26, 1.41)	1.33 (1.26, 1.41)	N/A
Female	0.59 (0.40, 0.87)	0.88 (0.64, 1.21)	1.05 (0.83, 1.32)	0.031
Age (per 5 years older)	1.46 (1.33, 1.60)	2.13 (1.69, 2.67)	1.46 (1.32, 1.62)	0.018
Age squared (per 5 years older)	1.011 (1.007,1.015)	1.009 (1.003,1.015)	1.009 (1.005,1.012)	0.710

OR – Odds Ratio, CI – Confidence Interval, SD – Standard Deviation; the Wald p-value indicates levels of statistical significance for differences in effect across the patient classes; CIs directly from analysis. Deprivation (measured using TDI) is inversely related to social status.

Table 3.8 Model class profiles for the model covariates by patient class in the three-patient, five-Trust-class MLLC model

Model Class profiles	Good Prognosis	Reasonable Prognosis	Poor Prognosis
	mean (bootstrapped 95% CI)		
Deprivation	-0.17 (-0.23, -0.10)	-0.03 (-0.11, 0.12)	0.09 (0.00, 0.16)
Age (years)	70.9 (70.7, 71.2)	72.6 (71.9, 73.5)	71.4 (70.7, 71.8)
	% patients (bootstrapped 95% CI)		
Female	44.0 (43.1, 44.9)	45.9 (44.4, 47.8)	42.7 (41.1, 44.1)

CI – Confidence Interval; CIs from bootstrapping calculated using percentiles. Deprivation (measured using TDI) is inversely related to social status.

Across all classes, older age is substantially and significantly associated with increased odds of death (good prognosis OR=1.46, 95% CI 1.33 to 1.60; reasonable prognosis OR=2.13, 95% CI 1.69 to 2.67; poor prognosis OR=1.46, 95% CI 1.32 to 1.62 per 5-year increase in age). Also across all classes, age-squared is substantially associated with increased odds of death (good prognosis OR=1.011, 95% CI 1.007 to 1.015; reasonable prognosis OR=1.009, 95% CI 1.003 to 1.015; poor prognosis OR=1.009, 95% CI 1.005 to 1.012 per 5-year increase in age). The mean age (in years) also differs across the classes (good prognosis 70.9, 95% CI 70.7 to 71.2; reasonable prognosis 72.6, 95% CI 71.9 to 73.5; poor prognosis 71.4, 95% CI 70.7 to 71.8), indicating that patients in the reasonable prognosis class are, on average, older than the patients in either of the other two classes.

Table 3.9 summarises the model class profiles for the patient classes in the same model.

The profile of stage differs across the patient classes. The good prognosis class corresponds to early-stage diagnosis with 70.8% (95% CI 66.0% to 75.1%) of the stage A and B patients compared with 36.3% (95% CI 8.2% to 44.0%) in the reasonable prognosis class and 6.0% (95% CI 4.2% to 13.1%) in the poor prognosis class. The poor prognosis class corresponds to late-stage diagnosis with 65.4% (95% CI 55.9% to 81.9%) of the stage D patients compared with 0.5% (95% CI 0.1% to 17.0%) in the reasonable prognosis class and 0.7% (95% CI 0.0% to 2.2%) in the good prognosis class. The reasonable prognosis class contains a large proportion of patients with missing values for stage (30.5%, 95% CI 20.6% to 47.1%).

A higher proportion of patients are treated in the good prognosis class (98.8%, 95% CI 97.6% to 99.4%) compared to either the reasonable prognosis class (81.4%, 68.2% to 86.7%) or the poor prognosis class (68.3%, 65.9% to 72.6%), which may be partly due to their stage at diagnosis, as early-stage patients are more likely to receive curative treatment (National Institute for Clinical Excellence, 2004).

Table 3.9 Model class profiles for the patient classes in the three-patient, five-Trust-class MLLC model

Model Class Profiles	Good Prognosis	Reasonable Prognosis	Poor Prognosis
	% patients (bootstrapped 95% CI)		
Stage A	23.2 (21.2, 25.1)	9.9 (0.2, 12.9)	0.0 (0.0, 2.1)
Stage B	47.6 (44.8, 50.0)	26.4 (8.0, 31.1)	6.0 (4.2, 11.0)
Stage C	26.5 (23.8, 28.4)	32.6 (26.9, 37.4)	17.2 (8.2, 19.7)
Stage D	0.7 (0.0, 2.2)	0.5 (0.1, 17.0)	65.4 (55.9, 81.9)
Missing stage	1.9 (0.0, 4.1)	30.5 (20.6, 47.1)	11.4 (0.2, 15.3)
Patients receiving treatment	98.8 (97.6, 99.4)	81.4 (68.2, 86.7)	68.3 (65.9, 72.6)
ICD-10 C18 (colon)	58.5 (57.5, 59.6)	56.0 (54.7, 58.0)	61.7 (60.6, 63.9)
ICD-10 C19 (rectosigmoid junction)	10.8 (10.2, 11.6)	9.7 (9.2, 10.5)	10.8 (9.9, 11.6)
ICD-10 C20 (rectum)	30.7 (29.7, 31.5)	34.3 (32.2, 35.7)	27.5 (25.3, 28.5)
Tumour on left side	68.7 (68.0, 69.5)	68.2 (65.2, 69.0)	61.2 (59.4, 62.4)
Tumour on right side	28.0 (27.1, 28.7)	25.2 (23.7, 26.7)	28.2 (27.0, 29.6)
Tumour across both sides	3.3 (2.9, 3.7)	6.6 (5.6, 9.8)	10.6 (9.5, 11.6)

CI – Confidence Interval; CIs from bootstrapping calculated using percentiles. Stage is modelled as a class predictor; patients receiving treatment, ICD-10 diagnosis code and laterality are modelled as inactive covariates.

There is some indication that the poor prognosis class contains a higher proportion of patients diagnosed with cancer of the colon (61.7%, 95% CI 60.6% to 63.9%), compared with the good and reasonable prognosis classes (good prognosis 58.5%, 95% CI 57.5% to 59.6%; reasonable prognosis 56.0%, 95% CI 54.7% to 58.0%), and a lower proportion of patients diagnosed with cancer of the rectum (27.5%, 95% CI 25.3% to 28.5%; versus good prognosis 30.7%, 95% CI 29.7% to 31.5%; and reasonable prognosis 34.3%, 95% CI 32.2% to 35.7%). This is also reflected in the results for laterality, as, while colon tumours may present on either side of the body, rectal tumours occur solely on the left side of the body.

There is no clinical rationale why the poor prognosis class should contain a higher proportion of colon tumours compared with rectal tumours, or those split across both sides of the body, compared with left side only; rather the difference is likely to be by stage (Lee et al., 2013).

The results seen for model covariates, in table 3.7, do not differ markedly from those obtained when different numbers of Trust classes are considered. Table 3.10 summarises the model covariate results for the three patient classes when considering between two and six Trust classes, with results grouped by model covariate for ease of interpretation.

SES remains clearly associated with increased odds of death across all classes and in all models (Townsend deprivation score OR ranges from 1.33 (95% CI 1.26 to 1.41, for five Trust classes) to 1.39 (95% CI 1.03 to 1.89, for six Trust classes)).

Females maintain decreased odds of death in the good prognosis class (OR ranges from 0.53 (95% CI 0.18 to 1.57, for six Trust classes) to 0.60 (95% CI 0.41 to 0.86, for three Trust classes)), although this only reaches statistical significance when considering three or five Trust classes. The association remains less clear, though consistent, in the reasonable and poor prognosis classes.

Older age remains substantially and significantly associated with increased odds of death across all classes, which is again consistent across all models. Age-squared also remains substantially associated with increased odds of death across all classes and all models.

Table 3.10 Model covariate results for the patient classes in the three-patient, two- to six-Trust-class MLLC models

Model Covariate	No. Trust Classes	Good Prognosis	Reasonable Prognosis	Poor Prognosis	Wald p-value
		OR of death within three years (95% CI)			
Deprivation (per SD more)	2	1.34 (1.16, 1.55)	1.34 (1.16, 1.55)	1.34 (1.16, 1.55)	N/A
	3	1.34 (1.27, 1.41)	1.34 (1.27, 1.41)	1.34 (1.27, 1.41)	N/A
	4	1.37 (1.12, 1.68)	1.37 (1.12, 1.68)	1.37 (1.12, 1.68)	N/A
	5	1.33 (1.26, 1.41)	1.33 (1.26, 1.41)	1.33 (1.26, 1.41)	N/A
	6	1.39 (1.03, 1.89)	1.39 (1.03, 1.89)	1.39 (1.03, 1.89)	N/A
Female	2	0.58 (0.31, 1.10)	0.97 (0.73, 1.28)	1.06 (0.71, 1.58)	0.250
	3	0.60 (0.41, 0.86)	0.88 (0.66, 1.18)	1.05 (0.83, 1.34)	0.046
	4	0.54 (0.24, 1.23)	0.94 (0.70, 1.25)	1.04 (0.53, 2.04)	0.190
	5	0.59 (0.40, 0.87)	0.88 (0.64, 1.21)	1.05 (0.83, 1.32)	0.031
	6	0.53 (0.18, 1.57)	0.92 (0.69, 1.24)	1.02 (0.35, 2.96)	0.220

OR – Odds Ratio, CI – Confidence Interval, SD – Standard Deviation; the Wald p-value indicates levels of statistical significance for differences in effect across the patient classes; CIs directly from analysis. Deprivation (measured using TDI) is inversely related to social status.

Table 3.10 continued Model covariate results for the patient classes in the three-patient, two- to six-Trust-class MLLC models

Model Covariate	No. Trust Classes	Good Prognosis	Reasonable Prognosis	Poor Prognosis	Wald p-value
		OR of death within three years (95% CI)			
Age (per 5 years older)	2	1.49 (1.39, 1.60)	2.20 (1.82, 2.66)	1.35 (1.15, 1.59)	<0.001
	3	1.47 (1.34, 1.61)	2.17 (1.72, 2.74)	1.45 (1.31, 1.61)	0.008
	4	1.48 (1.28, 1.72)	2.15 (1.74, 2.66)	1.39 (1.11, 1.74)	0.003
	5	1.46 (1.33, 1.60)	2.13 (1.69, 2.67)	1.46 (1.32, 1.62)	0.018
	6	1.49 (1.15, 1.94)	2.21 (1.45, 3.37)	1.41 (1.11, 1.79)	0.008
Age squared (per 5 years older)	2	1.010 (1.001, 1.019)	1.011 (1.007, 1.015)	1.008 (1.002, 1.013)	0.730
	3	1.011 (1.007, 1.015)	1.010 (1.004, 1.015)	1.009 (1.005, 1.012)	0.710
	4	1.011 (1.002, 1.020)	1.010 (1.005, 1.015)	1.007 (0.999, 1.015)	0.870
	5	1.011 (1.007,1.015)	1.009 (1.003,1.015)	1.009 (1.005,1.012)	0.710
	6	1.012 (1.001, 1.023)	1.011 (1.001, 1.021)	1.007 (0.997, 1.017)	0.910

OR – Odds Ratio, CI – Confidence Interval, SD – Standard Deviation; the Wald p-value indicates levels of statistical significance for differences in effect across the patient classes; CIs directly from analysis.

3.3.5 Patient-class comparison with MLM

Interest lies in the comparison between the patient classes identified in the MLLC model versus the MLM, with comparison between stage included as a class predictor in the MLLC model or excluded entirely in the MLM. Table 3.11 compares the MLM and MLLC patient-class results from tables 3.3, 3.6 and 3.7 side by side.

A single patient class is identified in the MLM with 51.6% of patients dying within three years of diagnosis. In contrast, the MLLC model identifies three patient classes that are distinct with respect to prognosis, representing variability in the patient groups. Proportions of patients dying within three years ranges from 9.4% (95% CI 2.2% to 17.4%) in the good prognosis class to 93.2% (95% CI 92.0% to 99.6%) in the poor prognosis class.

Although SES is constrained to be the same across the patient classes in the MLLC model, the effect size is greater than that seen in the MLM (MLLC OR=1.33, 95% CI 1.26 to 1.41; MLM OR=1.18, 95% CI 1.15 to 1.21; both per standard deviation increase).

For females, the MLM shows reduced odds of death compared with males (OR=0.87, 95% CI 0.83 to 0.92). In the MLLC model, this relationship is seen in the good prognosis class only (OR=0.59, 95% CI 0.40 to 0.87), suggesting an improved effect for females, compared with males, in early-stage diagnoses only.

Increased age at diagnosis, and higher values of age-squared, remain associated with increased odds of death both in the MLM and across all classes of the MLLC model. Effect sizes are greater in the MLLC model, with the OR for age ranging from 1.46 in both the good and poor prognosis classes (good prognosis 95% CI 1.33 to 1.60; poor prognosis 95% CI 1.32 to 1.62) to 2.13 in the reasonable prognosis class (95% CI 1.69 to 2.67), compared with 1.31 (95% CI 1.30 to 1.33) in the MLM (both per five years older).

Table 3.11 Comparison of results from MLM and MLLC analyses; odds of death within three years

Model Summary Statistics	MLM	MLLC			
		Good Prognosis / Early Stage Diagnosis	Reasonable Prognosis / Mid Stage Diagnosis	Poor Prognosis / Late Stage Diagnosis	Wald p-value
	% patients (bootstrapped 95% CI for MLLC model)				
Overall mortality	51.6	9.4 (2.2, 17.4)	58.3 (49.3, 72.9)	93.2 (92.0, 99.6)	N/A
Reference group mortality	49.3	8.0 (0.1, 16.5)	57.8 (36.7, 78.6)	94.1 (90.8, 100.0)	N/A
Model Covariates	OR of death within three years (95% CI)				
Deprivation (per SD more)	1.18 (1.15, 1.21)	1.33 (1.26, 1.41)	1.33 (1.26, 1.41)	1.33 (1.26, 1.41)	N/A
Female	0.87 (0.83, 0.92)	0.59 (0.40, 0.87)	0.88 (0.64, 1.21)	1.05 (0.83, 1.32)	0.031
Age (per 5 years older)	1.31 (1.30, 1.33)	1.46 (1.33, 1.60)	2.13 (1.69, 2.67)	1.46 (1.32, 1.62)	0.018
Age squared (per 5 years older)	1.006 (1.005, 1.007)	1.011 (1.007,1.015)	1.009 (1.003,1.015)	1.009 (1.005,1.012)	0.710

The reference group comprises males, aged 71.5 years, classified as Stage A at diagnosis and attributed a Townsend deprivation score of zero; OR – Odds Ratio, CI – Confidence Interval, SD – Standard Deviation; the Wald p-value indicates levels of statistical significance for differences in effect across the MLLC patient classes; CIs directly from analysis unless otherwise stated; LL (MLM) = -16,081, LL (MLLC) = -11,985. Deprivation (measured using TDI) is inversely related to social status.

3.3.6 Trust classes

Table 3.12 summarises the model summary statistics for the Trust classes from the chosen three-patient, five-Trust-class MLLC model, where Trusts are apportioned to one of five groups.

Trust classes are defined in order to account for heterogeneity at the Trust level, however interest lies in their comparison with the traditional MLM. In table 3.3, the MLM showed 12,708 patients (51.6%) dying within three years of diagnosis. In contrast, the MLLC model distinguishes five Trust classes with mean outcome (mortality) varying from 49.6% (95% CI 46.9% to 50.8%) to 54.5% (95% CI 52.6% to 59.6%) of patients dying within three years; Trust classes are hence labelled post-hoc from ‘best’ to ‘worst’ prognosis. Although ordered and labelled by prognosis, this is not to imply that Trusts within some classes perform better or worse than Trusts within other classes, as Trust classes contain differing profiles of patient characteristics, which will be described.

Class sizes range from 10.4% of patients (95% CI 5.2% to 47.0%; worst prognosis class) to 37.3% of patients (95% CI 3.7% to 51.6%; class 4). Confidence intervals are wide, indicating that the ordering of the best to worst prognosis classes varies considerably across bootstrapped datasets. In the example dataset, by modal assignment, the best and worst prognosis classes contain only two Trusts each, class 2 contains three Trusts, class three contains five Trusts, and class four contains seven Trusts.

Table 3.13 summarises the model class profiles for the Trust classes from the same model. Point values of SES differ somewhat across the Trust classes, with Trusts in the best prognosis class receiving patients on average from more affluent areas (mean Townsend score -0.39, 95% CI -1.04 to 1.53), while Trusts in class 2 receive patients on average from more deprived areas (mean Townsend score 0.38 (95% CI -1.18 to 1.33). Trusts in the worst prognosis class, however, receive the most affluent patients on average (mean Townsend score -0.45, 95% CI -1.22 to 0.45). Confidence intervals are very wide, however, indicating much variability in the values of SES across the bootstrapped datasets.

Table 3.12 Model summary statistics for the Trust classes in the three-patient, five-Trust-class MLLC model

Model Summary Statistics	**Best prognosis**	**Trust Class 2**	**Trust Class 3**	**Trust Class 4**	**Worst prognosis**
	% patients (bootstrapped 95% CI)				
Class size	11.1 (6.0, 38.0)	14.3 (6.0, 51.4)	26.9 (6.0, 54.3)	37.3 (3.7, 51.6)	10.4 (5.2, 47.0)
Mortality	49.6 (46.9, 50.8)	50.7 (48.5, 52.1)	50.9 (49.6, 53.5)	52.1 (50.8, 55.9)	54.5 (52.6, 59.6)

CI – Confidence Interval; CIs from bootstrapping calculated using percentiles.

Table 3.13 Model class profiles for the Trust classes in the three-patient, five-Trust-class MLLC model

Model Class Profiles	**Best prognosis**	**Trust Class 2**	**Trust Class 3**	**Trust Class 4**	**Worst prognosis**
	mean (bootstrapped 95% CI)				
Mean deprivation	-0.39 (-1.04, 1.53)	0.38 (-1.18, 1.33)	-0.05 (-1.11, 1.12)	0.05 (-1.49, 0.86)	-0.45 (-1.22, 0.45)
Mean age (years)	71.2 (70.7, 72.6)	71.6 (70.8, 72.5)	71.8 (71.0, 72.3)	71.5 (71.0, 72.5)	71.4 (71.0, 73.1)
	% patients (bootstrapped 95% CI)				
Female	43.4 (41.1, 47.6)	44.1 (42.1, 47.0)	44.3 (41.8, 47.2)	44.0 (41.6, 47.2)	44.6 (42.2, 47.8)

CI – Confidence Interval; CIs from bootstrapping calculated using percentiles. Deprivation (measured using TDI) is inversely related to social status.

Table 3.13 continued Model class profiles for the Trust classes in the three-patient, five-Trust-class MLLC model

Model Class Profiles	Best prognosis	Trust Class 2	Trust Class 3	Trust Class 4	Worst prognosis
	% patients (bootstrapped 95% CI)				
Stage A	12.2 (9.9, 13.7)	11.6 (9.5, 14.4)	11.6 (9.0, 13.6)	11.6 (9.4, 13.6)	10.8 (9.4, 13.3)
Stage B	26.9 (24.6, 29.7)	26.9 (24.7, 29.5)	27.1 (24.7, 29.4)	28.0 (24.9, 30.6)	28.6 (25.3, 34.2)
Stage C	25.3 (21.3, 28.2)	24.0 (22.4, 29.2)	26.6 (22.5, 29.8)	24.7 (21.5, 29.5)	22.9 (18.6, 26.7)
Stage D	22.7 (20.8, 25.2)	23.7 (20.4, 24.6)	23.0 (20.8, 24.7)	22.5 (20.3, 24.3)	22.0 (18.2, 24.2)
Missing stage	12.9 (9.9, 15.3)	13.8 (10.3, 14.8)	11.7 (10.3, 16.4)	13.2 (10.2, 16.0)	15.7 (11.5, 18.7)
Patients receiving treatment	84.5 (81.4, 87.7)	82.7 (82.1, 86.8)	84.7 (80.7, 87.5)	83.0 (81.2, 87.0)	81.7 (78.8, 84.6)
ICD-10 C18 (colon)	59.1 (54.9, 63.8)	57.3 (55.5, 61.5)	58.9 (55.5, 62.5)	58.5 (56.6, 63.1)	61.8 (57.2, 64.7)
ICD-10 C19 (rectosigmoid junction)	11.1 (8.8, 12.6)	10.9 (7.5, 12.6)	10.2 (5.5, 12.2)	10.9 (5.3, 12.3)	8.9 (7.5, 13.0)
ICD-10 C20 (rectum)	29.8 (25.4, 34.6)	31.8 (27.6, 33.7)	30.9 (27.9, 34.6)	30.7 (27.9, 35.8)	29.2 (24.3, 32.8)
Tumour on left side	64.8 (63.3, 70.0)	67.2 (64.1, 69.2)	67.1 (63.4, 69.5)	65.7 (62.3, 68.9)	63.9 (61.7, 67.8)
Tumour on right side	27.3 (24.7, 30.3)	26.7 (24.7, 29.7)	27.7 (24.5, 29.3)	27.7 (25.1, 29.7)	25.3 (23.5, 29.6)
Tumour across both sides	7.9 (4.2, 9.8)	6.1 (4.2, 9.4)	5.2 (4.4, 10.2)	6.6 (4.2, 11.1)	10.8 (5.4, 12.0)

CI – Confidence Interval; CIs from bootstrapping calculated using percentiles. Stage is modelled as a class predictor; patients receiving treatment, ICD-10 diagnosis code and laterality are modelled as inactive covariates.

Both the mean age of patients and the proportion of females remain consistent across the classes. Mean age ranges from 71.2 years (95% CI 70.7 to 72.6; best prognosis class) to 71.8 years (95% CI 71.0 to 72.3; class 3), while the proportion of females ranges from 43.4% (95% CI 41.1% to 47.6%; best prognosis class) to 44.6% (95% CI 42.2% to 47.8%; worst prognosis class).

Proportions of patients within each of the stage categories also remain consistent across the classes, although there may be an indication that the worst prognosis class contains slightly more patients with missing values for stage (15.7%, 95% CI 11.5% to 18.7%) compared with the other classes. The worst prognosis class also contains the fewest patients receiving curative treatment (81.7%, 95% CI 78.8% to 84.6%). Although not significant differences, taken together, there may be an indication that the two Trusts in this class are not treating as many early-stage patients as other Trusts.

Consistency is also predominantly seen across the classes for both the type and position of tumour, although there may be an indication that the worst prognosis class has the highest proportion of colon tumours (61.8%, 95% CI 57.2% to 64.7%) and the lowest proportion of tumours of the rectosigmoid junction (8.9%, 95% CI 7.5% to 13.0%). With colon tumours presenting across both sides of the body, and tumours of the rectosigmoid junction presenting entirely on the left side of the body, this may partly explain why the worst prognosis class also has the lowest proportion of tumours on the right side of the body (25.3%, 95% CI 23.5% to 29.6%) and the highest proportion of tumours split across both sides of the body (10.8%, 95% CI 5.4% to 12.0%).

3.4 Discussion

The MLLC model provided a better interpretation of the data compared with the MLM analysis. The MLM found sizeable and significant associations between increasing values of the Townsend deprivation score and increased odds of death, between being female and decreased odds of death, and between older age and increased odds of death. The MLLC analysis categorised patients into three latent classes (labelled as good, reasonable and poor prognosis), and in all classes, the overall impact of both SES (measured in these data using the TDI) and age was found to agree with the MLM. For sex, females had decreased odds of death in the good prognosis class, but less association in the reasonable and poor prognosis classes, thus recognising that the relationship between sex and mortality differs somewhat by prognosis. These differences indicate that a single patient class, as in the MLM, is not sufficient to model these data, due to heterogeneity at the patient level.

As stage at diagnosis was identified as a mediator of the relationship between SES and three-year mortality, its inclusion in a traditional regression model would introduce bias due to the reversal paradox; hence, it was excluded from the MLM. It was, however, included as a class predictor in the MLLC model. This therefore established the contrast between the two models as a comparison with and without the inclusion of stage. Good, reasonable and poor prognosis classes corresponded to early-, mid- and late-stage diagnosis respectively, with the majority of the advanced stage patients in the poor prognosis class. Females are seen to have decreased odds of death compared with males for early-stage disease.

The effect of SES was constrained to be the same across all patient classes, in order to avoid the causal circularity between SES, stage and survival. This, however, may not avoid some degree of residual bias due to the reversal paradox, as the exposure-outcome relationship is unlikely to be independent of within-class intercepts, which effectively are ‘adjusted’ by the consideration of stage as a class predictor. Stage could therefore be removed entirely from the MLLC model, and other studies have shown that latent classes remain similarly defined whether or not stage is included

(Downing et al., 2010). The removal of stage represents a cultural shift in approach, however, as it is commonly included in traditional analyses. Utilisation of a novel technique in addition to the removal of a covariate previously considered to be fundamental when modelling survival relationships may take time to be adopted.

Across all classes, the odds of death for both the Townsend deprivation score and age at diagnosis were greater in the MLLC model results, compared with the MLM analysis. This is perhaps due in part to the appropriate inclusion of stage in the model, and in an improved model fit (LL = -16,081 for the MLM versus -11,978 for the three-patient, five-Trust-class MLLC model), leading to improved estimates.

Although models with differing numbers of Trust classes could have been chosen, differences in output were minimal and the same patterns of association were seen for all model covariates. The five Trust classes identified outlying Trusts, indicating that the traditional MLM with a single, continuous latent variable at the Trust level, is not sufficient to model these data. Patient casemix differences can be seen across the Trust classes (e.g. in the different mean values of the Townsend deprivation score) and no adjustment has been made for these differences, so there can be no inference from this analysis as to the performance of the NHS Trusts included in the Trust classes. An alternative approach of grouping Trusts according to differences in characteristics is discussed in Chapter 4, where differences in survival at the Trust level may be as a result of underlying differences in Trust performance, rather than by patient casemix.

Chapter 4

Research Question (2); Casemix Adjustment

4.1 Introduction

Chapter 4 examines the second application of the latent variable methodological approach introduced in Chapter 1 and further defined in Chapter 2. The first application was seen in Chapter 3, where MLLC modelling was utilised to investigate the relationship between model covariates at the patient level, while accounting for heterogeneity at the provider level. Discrete latent classes, class predictors and unique class features were used to exactly specify model configurations, while accounting for data challenges specific to the example dataset.

Chapter 4 explores research question (2), again using the example dataset (although with minor differences in the number of deaths compared with the data used in Chapter 3):

(2) How does the performance of a healthcare provider vary after accommodating patient differences?

This is an important question within healthcare delivery, where provider performance may be assessed and compared in order to identify best practice and advocate changes in under-performing institutions. Some providers may perform better or worse than others in terms of average survival rates, for example, but these differences may reflect the characteristics of their patients rather than underlying differences in their effectiveness. Section 1.2.3 introduced the concept of patient ‘casemix’ leading to differential selection based on patient heterogeneity.

Results in Chapter 3 included a discussion of differences in prognosis across Trust-level latent classes (see section 3.3.6), and it was clear that

Trust performance could not be directly compared using patient-outcome differences, due to differences in patient-profile characteristics.

In this chapter, model parameterisation thus differs substantially from that seen in Chapter 3, with differential selection (due to patient heterogeneity and casemix differences) separated from the potential causal structure of factors influencing provider performance. Initially, for simplification, no provider-level covariates are considered. This allows for a focus on the accommodation of differential selection at the patient level, and enables comparison with traditional methodologies. Chapter 5 extends the approach to explore the principle of evaluating causal factors operating to influence provider performance.

Section 4.2 summarises the data and methods relevant to this research question, including the modelling approach, patient-level covariate configuration to account for differential selection, detailed parameterisation, optimum model construction and calculation of Trust performance rankings. Calculation of the SMR is identified as the traditional comparison.

Section 4.3 contains all results, including model construction, a summary of the composition of both patient and Trust classes, and the comparison of performance ranking between the MLLC approach and calculation of SMRs.

Section 4.4 provides a discussion of the methods and results.

This chapter contains work based on two publications (Gilthorpe et al., 2011; Harrison et al., 2012).

4.2 Data and methods

4.2.1 MLLC approach to the data

The example dataset described in Chapter 2 is again utilised, containing data on 24,640 patients diagnosed with colorectal cancer between 1998 and 2004; 12,856 patients (52.2%) died within three years of diagnosis, which, due to coding differences, slightly differs from the number of deaths seen in Chapter 3, as discussed in section 2.3.4. The same covariates are available for inclusion as summarised in section 2.3.4, although not all are utilised, as explained in section 4.2.2.

A MLLC approach is preferred to answer the research questions, for reasons discussed in section 2.4.1. As identified in section 2.4.2, these data have a two-level hierarchical structure with NHS Trust at the upper level, used here as an example of an area-level healthcare provider. Other datasets may contain different organisational groupings, such as clinical commissioning groups (CCGs), for example. There is likely to be heterogeneity at both levels; patient groups may differ in their characteristics, leading to differential selection, and thus Trusts may differ in their patient casemix. MLLC modelling can accommodate the data structure, while maintaining an overarching framework that can separate modelling for causal inference and for differential selection across different levels of a hierarchy.

Broad modelling strategy (ii), introduced in section 2.4.3, is utilised. While patient classes are constructed based on similarities in patient characteristics, Trust classes are instead determined based on differences in patient characteristics, and the same proportion of each patient class is allocated to each Trust class. Trust classes are therefore generated that are identical with respect to patient characteristics, i.e. they are patient casemix ‘adjusted’. Trust-class outcomes (for example mean three-year mortality), may differ, but these differences will then be due to underlying differences in Trust performance, due to unmodelled factors (potential covariates) operating at the Trust level, rather than patient casemix.

4.2.2 Casemix adjustment

As discussed in section 3.2.1, stage at diagnosis remains an imprecise measure, potentially exacerbating bias due to measurement error, particularly if interaction terms are considered, and stage contains missing data (though categorised for analysis). The DAGs constructed in figures 3.1, 3.2 and 3.3 demonstrate differing complex theorised relationships between covariates at the patient level, with the DAG shown in figure 3.1 also chosen for the purposes of analysis within this chapter. Although potential causality has been identified between SES, stage and survival, there are no concerns regarding bias due to the reversal paradox in this analysis, since there is no attempt to seek causal inference or to make any confounder adjustment at the patient level. Casemix adjustment can be viewed as purely predictive modelling, to maximally explain the outcome with respect to the model covariates, with no regard for their causal relationship.

The variables available for analysis within the example dataset are previously summarised in table 2.3. Table 4.1 reiterates the available variables and specifies which are included, and how they are modelled, within the MLLC analytical approach to this research question.

Table 4.1 Variables included in analysis for research question (2)

Variables available for analysis	Variables included in analysis	Modelling approach
Deprivation	Deprivation	Regression
Sex	Sex	Regression
Age at diagnosis	Age at diagnosis	Regression
	Age-squared	Regression
Stage at diagnosis	Stage at diagnosis	Regression
ICD-10	-	-
Laterality	-	-
Treated	-	-

Deprivation is a measure of SES, measured in these data using TDI.

Thus, optimum outcome prediction is sought by modelling patient characteristics in order to accommodate casemix differences. Consequently,

SES (measured in these data using the TDI), sex and age at diagnosis (centred around the study mean of 71.5 years) are included in the regression part of the model at the patient level, along with stage at diagnosis (coded A to D for increasing severity and missing values coded X), because stage plays a crucial role in affecting survival outcomes. Uncertainty due to measurement error, and to unmeasured covariates, is incorporated through the latent constructs.

An age-squared term is again included, to model appropriately the non-linear relationship between age and survival. Patient-level covariates are otherwise simplified for inclusion into the casemix-adjusted model, however, as interest lies in Trust-level comparisons rather than patient-level relationships. Generalised additive models (GAMs) (West, 2012) are used to visually identify threshold values for both SES and age, beyond which values become uncommon and thus relationships may become atypical. These tails of the distributions are then 'trimmed'; for age, rare values less than -10 (equivalent to 61.5 years of age) were assigned to equal -10, while for SES, rare values greater than 5 were assigned to equal 5.

No class predictors or inactive covariates are included, as this modelling configuration is designed to account for patient-level variation in the differentiation of Trust-level outcomes, rather than to investigate patient-class differences. Therefore, variables previously included as inactive for analysis in Chapter 3, i.e. tumour site (using ICD-10 diagnosis code), laterality and whether or not the patient received curative treatment, are excluded from analysis.

4.2.3 Parameterisation

Detailed parameterisations, introduced in section 2.4.4, are summarised here with respect to research question (2).

Intercepts. Class-independent intercepts are set for the patient classes, in relation to Trust classes, as also utilised in Chapter 3. Identical contrasts can thus be made amongst patient classes, within all Trust classes. Detailed interpretation of patient classes is not intended, however, as focus is on the Trust classes and their implications on Trust-level outcome differences.

Covariate effects. All patient-class covariate effects are initially designated Trust-class dependent, thus allowing parameter values to vary across the Trust classes and hence, for the relationship between each exposure and outcome to differ across the patient classes. Prediction modelling, rather than modelling for causal inference, is required at the patient level to account for patient casemix, therefore there are no concerns regarding any causal circularity between SES, stage and survival. It is not necessary, therefore, to constrain the effect of SES across the classes.

Nevertheless, any of the covariate effects may be constrained to be Trust-class independent for parsimony, if there is evidence that a relationship does not vary across the patient classes.

Class sizes. In contrast to Chapter 3, patient-class sizes are designated class independent with respect to Trust classes, as required for modelling strategy (ii). This ensures that each Trust class contains the same proportion of each patient class; thus Trust classes each contain the same patient casemix.

Error variance. This is not applicable for a binary outcome.

4.2.4 Optimum model

Optimum model construction again follows the process suggested in section 2.2.6. Initially, a continuous latent variable is adopted at the Trust level while the number of patient classes are sequentially increased from one to identify the optimum number of patient-level classes based on interpretability, but with parsimonious assessment from model-evaluation criteria and CE. The continuous latent variable is then switched to categorical to identify the optimum Trust-level structure. Log-likelihood statistics, model parsimony and CE are again explored, although also with a mind on utility, since a minimum of two Trust classes is both necessary and sufficient to exhibit discretised Trust-class differences in patient outcomes. Indeed, it may be desirable to consider more than two Trust classes to obtain optimal utility from this approach, even if model likelihood statistics are not improved by an increased number of Trust classes.

A description of model construction for the MLLC approach, using the example dataset, is presented in section 4.3.2.

4.2.5 Bootstrapping

200 bootstrapped datasets are generated, following the same process as described in section 3.2.6, both for dataset generation and calculation of 95% confidence intervals (CI) for the model summary statistics and model class profiles, using percentiles (2.5% to 97.5%), with model class profile figures based on probabilistic assignment to classes. The chosen MLLC model, as constructed in section 4.3.2 is utilised.

There are no model class profiles at the patient level, as all covariates are included in the regression part of the model; CIs for the model covariates are determined directly from MLLC analysis of the example dataset.

The primary utility of the bootstrapped datasets within this chapter, however, is in the comparison of Trust performance rankings, as described in sections 4.2.6 and 4.2.7.

4.2.6 Trust performance rankings

In a MLLC analysis, Trust classes will exhibit a graduated patient outcome (i.e. three-year mortality), which is used to generate ranks of Trust performance. Trusts are ordered based on their probabilistic assignment to the best survival Trust class, thus generating a performance ranking for each Trust that is comparable across Trusts. Across the nineteen Trusts, a rank of one indicates that a Trust has a high probabilistic assignment to the best survival Trust class while a rank of nineteen indicates that a Trust has a low probabilistic assignment to the best survival Trust class.

In order to ascertain the variability of the Trust performance rankings, MLLC analysis is replicated for each of the 200 bootstrapped datasets, using the chosen model as selected in section 4.3.2. Trusts are ranked from one to nineteen within each dataset; thus a median rank can be calculated together with a credible interval (CI; 2.5% to 97.5%) (Marshall and Spiegelhalter, 1998) for each Trust, over all datasets.

4.2.7 Traditional comparison

Trust performance ranking using MLLC modelling is compared with calculation of the SMR, an indirect standardisation approach introduced in section 1.3.4. As an indirect adjustment, a standard population distribution is not required; instead, a comparison of rates is utilised. For the SMR, comparison is therefore made between the number of observed and expected deaths within each Trust (scaled by Trust size), with the observed data used to calculate the figures for both the observed and expected deaths.

Logistic regression is first performed across the entire dataset, using the same exposure and outcome variables as for the MLLC analysis. The probability of death within three years can then be determined for each patient, based on (i.e. standardised by) observed values of age, sex, SES and stage. The number of expected deaths per Trust is calculated as the sum of these probabilities across all patients within a Trust. The number of observed deaths within a Trust is straightforward, and available explicitly within the example dataset.

Once numbers of observed and expected deaths are determined for each Trust, the SMR can be calculated using the equation:

$$SMR = {no.observed\ deaths}/{no.expected\ deaths}$$

An SMR value equal to one indicates that the numbers of expected and observed deaths are the same, while a figure greater than one indicates a higher number of observed deaths than expected and a figure less than one indicates a lower number of observed deaths than expected. The difference from a SMR value of one is calculated for each Trust, with negative values indicating better outcomes and positive values indicating worse outcomes, compared with expected figures.

This SMR difference is scaled by the Trust population size (by dividing by the square root of the Trust size), to calculate a scaled value that can be used to make direct comparisons across Trusts. Trusts are ranked from one to nineteen in increasing order of this scaled difference. Thus, a rank of one

is given to the ‘best’ survival Trust, while a rank of nineteen is given to the ‘worst’.

As for the MLLC modelling, the same 200 bootstrapped datasets are similarly analysed by calculation of the scaled SMR difference, and each Trust within each dataset is ranked from one to nineteen as described above. Again, the median rank and CI (2.5% to 97.5%) is calculated for each Trust, over all datasets.

Each Trust therefore has a median rank and CI calculated by each approach, which can thus be contrasted.

4.3 Results

4.3.1 Outline

Results are illustrated for the MLLC analysis approach, with comparison made in section 4.3.5 between Trust performance ranks generated using this approach and by calculation of the SMR. Table 4.2 summarises the variables contained within each model. For both approaches, all variables are included as covariates within the regression model, as discussed in sections 4.2.2 and 4.2.7.

Table 4.2 Comparison of variables included in MLLC model and calculation of SMR

Variables included in analysis	
MLLC	**SMR**
Deprivation	Deprivation
Sex	Sex
Age at diagnosis	Age at diagnosis
Age-squared	Age-squared
Stage at diagnosis	Stage at diagnosis

Deprivation is a measure of SES, measured in these data using TDI.

4.3.2 Building the MLLC model

As described in section 4.2.4, a continuous latent variable is initially adopted at the Trust level in order to ascertain the optimum number of latent classes at the patient level. Table 4.3 summarises the model-evaluation criteria for the MLLC patient classes in this situation.

One patient class is seen to be optimum according to the BIC, the statistic that favours maximum parsimony, whilst selection of four patient classes minimises the value of the AIC, also geared to favour parsimony, although less so. As seen in Chapter 3, and as expected due to the lack of accommodation for parsimony, the LL shows continual improvement in model fit as the number of patient classes are increased, although this increase slows beyond two patient classes.

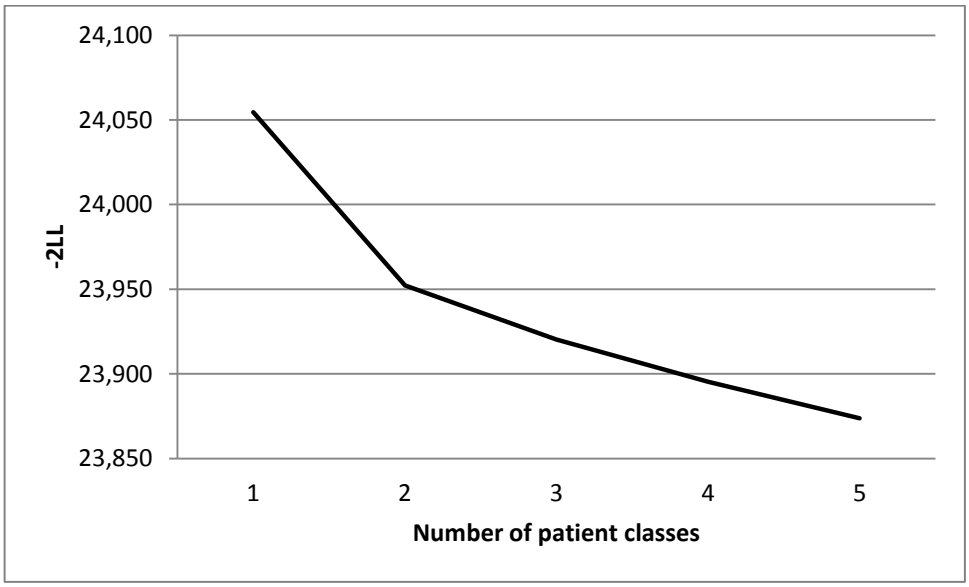
Table 4.3 Model-evaluation criteria for the patient classes in the MLLC models with a continuous Trust-level latent variable

Patient Classes	LL	BIC	AIC	No. of Parameters	Patient CE
1 class	-12,027	24,156	24,075	10	0.0%
2 classes	-11,976	24,165	23,994	21	34.6%
3 classes	-11,960	24,244	23,984	32	48.1%
4 classes	-11,948	24,330	23,981	43	42.2%
5 classes	-11,937	24,420	23.982	54	54.2%

LL – Log Likelihood, BIC – Bayesian Information Criterion, AIC – Akaike Information Criterion, CE – Classification Error.

Figure 4.1 displays the change in -2LL for increasing numbers of patient classes.

Figure 4.1 -2LL plot to determine the optimum number of patient classes in the MLLC modelling approach



Patient CE also increases with the number of patient classes, with 54.2% for five patient classes indicating that the majority of patients are split probabilistically across these classes, rather than being assigned mostly to a single class; the patient classes thus become generally more ‘virtual’ as the number of classes are increased. CE is not a concern, however, when modelling for prediction.

Considering the model-evaluation criteria for guidance, and recognising that more than one patient class is preferred to fully model patient variability, two patient classes are chosen.

Table 4.4 shows the model-evaluation criteria for the Trust classes, when the continuous latent variable at the Trust level is switched to categorical. Two patient classes remain fixed.

Table 4.4 Model-evaluation criteria for the Trust classes in the MLLC models with a categorical Trust-level latent variable; two patient-level latent classes

Trust Classes	LL	BIC	AIC	No. of Parameters	Patient CE	Trust CE
1 class	-11,979	24,150	23,996	19	35.0%	0.0%
2 classes	-11,977	24,166	23,995	21	35.4%	18.6%
3 classes	-11,976	24,184	23,998	23	35.6%	22.1%
4 classes	-11,976	24,204	24,002	25	35.6%	24.6%
5 classes	-11,976	24,225	24,006	27	35.6%	25.4%

LL – Log Likelihood, BIC – Bayesian Information Criterion, AIC – Akaike Information Criterion, CE – Classification Error.

Figure 4.2 displays the change in -2LL for increasing numbers of Trust classes.

Figure 4.2 -2LL plot to determine the optimum number of Trust classes in the MLLC modelling approach

The BIC shows one Trust class to be optimum, whilst the AIC just prefers two. The LL shows some improvement as the number of Trust classes are increased, up to three Trust classes, although differences are small and care must be taken not to over-interpret within this narrow range of figures. Again, the model-evaluation criteria are considered for guidance only. As discussed in section 4.2.4, at least two Trust classes are required in order to distinguish Trust-class differences, therefore two Trust classes are chosen on this occasion. The impact of a greater number of Trust classes is explored through simulations in Chapter 5.

The chosen model therefore contains two patient classes and two Trust classes. Patient CE is 35.4% and Trust CE is 18.6%, indicating that the patient classes are more ‘virtual’ than the Trust classes, which is acceptable for predictive modelling at the patient level.

4.3.3 Patient classes

Table 4.5 summarises the patient classes in the two-patient, two-Trust-class MLLC model selected in section 4.3.2.

Patients are assigned to two latent classes of similar size, one labelled ‘best’ prognosis (45.7% of cases (95% CI 18.1% to 82.2%), of which 39.3% (95% CI 33.3% to 48.2%) died within three years), and one labelled ‘worst’ prognosis (54.3% of cases (95% CI 17.8% to 81.9%), of which 63.0% (95% CI 54.9% to 84.5%) died within three years). The reference group comprises males of mean age (71.5 years), classified as stage A colorectal cancer at diagnosis, and attributed a Townsend deprivation score of zero.

While prognosis classes are categorised by overall mortality, and there are two distinct classes (considering the range of the CIs), both class size and reference group mortality are more variable. This is due to the variability in values within each class, across the bootstrapped datasets, leading to wide CIs.

Table 4.5 Results for the patient classes in the two-patient, two-Trust-class MLLC model; odds of death within three years

Model Summary Statistics	**Best prognosis**	**Worst prognosis**
	% of patients (bootstrapped 95% CI)	
Class size	45.7 (18.1-82.2)	54.3 (17.8-81.9)
Overall mortality	39.3 (33.3-48.2)	63.0 (54.9-84.5)
Reference group mortality	7.0 (0.0-86.2)	23.2 (1.3-69.4)
Model Covariates	**OR of death within three years (95% CI)**	
Deprivation (per SD more)	1.03 (0.81-1.31)	1.32 (1.21-1.43)
Female	0.58 (0.38-0.88)	0.94 (0.78-1.14)
Age (per 5 years older)	2.53 (1.31-4.90)	1.51 (1.42-1.60)
Age squared (per 5 years older)	0.984 (0.960-1.008)	1.005 (0.997-1.012)
Stage = B	0.55 (0.21-1.43)	2.40 (1.63-3.54)
Stage = C	1.74 (0.75-4.06)	7.72 (4.61-12.94)
Stage = D	Infinite†	20.19 (8.88-45.89)
Stage = X	33.41 (7.93-140.68)	6.30 (1.89-20.97)

OR – Odds Ratio, CI – Confidence Interval; CIs directly from analysis unless otherwise stated; †The odds ratio cannot be estimated as there were zero patients who survived 3 years in this subcategory. Deprivation (measured using TDI) is inversely related to social status.

Although patient classes are determined using predictive modelling, i.e. to account for differential selection, interest remains regarding any differences in covariate effects across the classes. Interpretation is cautious, as, for example, stage at diagnosis has not been modelled appropriately for causal inference. Thus, it is the differences across the classes that are of interest, rather than the overall magnitude of the relationships.

A significant association is seen between increasing deprivation and increased odds of death in the worst prognosis class (Townsend deprivation score OR=1.32, 95% CI 1.21 to 1.43), compared with little association in the best prognosis class (OR=1.03, 95% CI 0.81 to 1.31). In contrast, a significant association is seen between female gender and decreasing odds of death in the best prognosis class (OR=0.58, 95% CI 0.38 to 0.88), compared with little association in the worst prognosis class (OR=0.94, 95% CI 0.78 to 1.14). Substantial and significant associations are seen between older age and increased odds of death in both classes (best prognosis OR=2.53, 95% CI 1.31 to 4.90; worst prognosis OR=1.51, 95% CI 1.42 to 1.60). Little association is seen for age-squared in either class (best prognosis OR=0.984, 95% CI 0.960 to 1.008; worst prognosis OR=1.005, 95% CI 0.997 to 1.012), perhaps due to the ‘trimming’ of the tail of the age distribution, as described in section 4.2.2.

Model covariate relationships therefore generally agree with those seen in Chapter 3, where both increasing deprivation and older age were associated with increased odds of death across three patient classes. Further, females were shown to have decreased odds of death compared with males in the good prognosis class only, as also seen here.

The effect of stage at diagnosis also differs across the patient classes; stage A (earliest stage) is designated as the comparison group. In the worst prognosis class, all other stage categories are associated with increased odds of death, and the odds increase as severity increases (stage B OR=2.40, 95% CI 1.63 to 3.54; stage C OR=7.72, 95% CI 4.61 to 12.94; stage D OR=20.19, 95% CI 8.88 to 45.89). Odds of death are also increased for missing values of stage, compared with stage A at diagnosis (OR=6.30, 95% CI 1.89 to 20.97). In the best prognosis class, the association is not as

clear, although there remains a graduation in point values of odds of death with increasing severity from early- to late-stage diagnosis. There is little association seen, however, either between stage B or C at diagnosis (stage B OR=0.55, 95% CI 0.21 to 1.43; stage C OR=1.74, 95% CI 0.75 to 4.06), compared with stage A. The association for missing values of stage remains evident (OR=33.41, 95% CI 7.93 to 140.68), but the association between stage D at diagnosis and three-year mortality cannot be estimated, as all patients in this category died.

There is a noticeable pattern in deaths by stage and prognosis categories, as summarised in table 4.6, based on modal class assignment (i.e. by allocation of patients to classes according to their largest class probability).

All patients in the worst prognosis class diagnosed at either stage B or C died within three years; in the best prognosis class, all patients diagnosed at stage A, B or C survived, while all patients diagnosed at stage D died. This difference is anticipated, as stage at diagnosis is an important predictor of survival. Thus, while all of the early- and mid-stage patients survived at three years in the best prognosis class, most died within three years in the worst prognosis class.

Table 4.6 Deaths by stage and patient class, for the two-patient, two-Trust-class MLLC model

Stage at Diagnosis	**Modal Class; No. (%) of patients died within three years**					
	Best prognosis			**Worst prognosis**		
	Survived	**Died**	**Total**	**Survived**	**Died**	**Total**
A	1,210 (100.0%)	0 (0.0%)	1,210	1,099 (66.6%)	550 (33.3%)	1,649
B	4,829 (100.0%)	0 (0.0%)	4,829	0 (0.0%)	1,955 (100.0%)	1,955
C	3,437 (100.0%)	0 (0.0%)	3,437	0 (0.0%)	2,736 (100.0%)	2,736
D	0 (0.0%)	1,962 (100.0%)	1,962	437 (12.0%)	3,202 (88.0%)	3,639
Missing (X)	359 (79.8%)	91 (20.2%)	450	413 (14.9%)	2,360 (85.1%)	2,773
Total	**9,835 (82.7%)**	**2,053 (17.3%)**	**11,888**	**1,949 (15.3%)**	**10,803 (84.7%)**	**12,752**

4.3.4 Trust classes

Table 4.7 summarises the Trust classes in the two-patient, two-Trust-class MLLC model selected in section 4.3.2.

Trusts are also assigned to two latent classes of similar size, one labelled ‘best’ prognosis (53.1% of cases (95% CI 25.7% to 92.5%), of which 51.3% (95% CI 49.6% to 52.3%) died within three years), and one labelled ‘worst’ prognosis (46.9% of cases (95% CI 7.5% to 74.3%), of which 53.2% (95% CI 52.6% to 60.0%) died within three years).

Classes are ordered and labelled by prognosis, and although there is only a small difference in overall mortality, classes are again distinct, considering the range of the CIs. Class size ranges, however, are again wide, due to variability in class size across the bootstrapped datasets.

Model class profiles are balanced across the Trust classes, as would be expected for a casemix-adjusted model. A direct comparison between the Trust-class results in table 4.7 and those in tables 3.12 and 3.13, generated by means of the three-patient, five-Trust-class MLLC model, would not be appropriate due to the substantial differences in model parameterisations. While the Trust classes defined in analysis in Chapter 3 deliberately contain different proportions of patient classes, leading to differences in patient composition across the classes, those in table 4.7 are adjusted for differential selection and hence each represents the entire spectrum.

Table 4.7 Results for the Trust classes in the two-patient, two-Trust-class MLLC model; odds of death within three years

Model Summary Statistics	**Best prognosis**	**Worst prognosis**
	% patients (bootstrapped 95% CI)	
Class Size	53.1 (25.7-92.5)	46.9 (7.5-74.3)
Mortality	51.3 (49.6-52.3)	53.2 (52.6-60.0)
Model Class Profiles	**Mean (bootstrapped 95% CI)**	
Mean deprivation	-0.13 (-0.48 to 0.00)	-0.25 (-1.05 to 0.02)
Mean age (years)	73.1 (72.9-73.5)	73.0 (72.8-74.3)
	% patients (bootstrapped 95% CI)	
Female	44.2 (43.3-45.2)	44.0 (42.8-47.1)
Stage A	11.9 (10.9-12.4)	11.3 (10.1-12.6)
Stage B	27.2 (26.4-28.6)	28.0 (26.7-33.3)
Stage C	25.2 (24.2-26.4)	24.9 (19.3-26.3)
Stage D	23.1 (22.2-23.8)	22.3 (19.0-23.5)
Missing stage	12.7 (11.7-13.7)	13.5 (11.9-17.3)

CI – Confidence Interval; CIs from bootstrapping calculated using percentiles. Deprivation (measured using TDI) is inversely related to social status.

4.3.5 Performance ranking comparison

Results comparing Trust performance rankings between the MLLC approach and the calculation of SMRs are summarised in table 4.8; a low ranking value indicates a better survival rate than expected. Trusts are ordered by their median probability of belonging to the best survival MLLC Trust class, by methods described in section 4.2.6.

Table 4.8 Trust ranks from the MLLC model and the calculation of Trust SMRs

Trust	Median probability of belonging to best survival Trust class	Median Rank (95% CI)	
		MLLC	SMR
1	1.000	1 (1-9.5)	6 (2-11)
2	0.999	3 (1-11)	4 (1-10.5)
3	0.997	4 (1-11)	3 (1-10.5)
4	0.996	4 (1-15)	8 (3-14.5)
5	0.993	5 (1-12.5)	5 (1-13)
6	0.956	8 (2-16)	9 (2-17)
7	0.912	9 (3-17)	5 (1-17)
8	0.908	9 (2-17)	6 (1-18)
9	0.897	9 (3-18)	5 (1-18)
10	0.816	10 (3-17)	8 (1-18)
11	0.575	11 (3.5-18)	11 (3-17)
12	0.476	13 (5.5-18)	12.5 (3-18)
13	0.372	12 (4-18.5)	11.5 (5.5-17)
14	0.359	12 (3-19)	12 (7-17)
15	0.152	14 (5.5-19)	15 (4.5-18)
16	0.070	14 (4-19)	13 (7-18)
17	0.070	15 (7.5-19)	16 (7.5-18)
18	0.003	18 (7-19)	15 (10-18)
19	0.002	18 (13.5-19)	19 (18-19)

CI – Credible interval (2.5% to 97.5%); point values from interpolation.

Figure 4.3 then provides a graphical representation of these results, in order of increasing median probability of belonging to the best survival Trust class by the MLLC analytical approach.

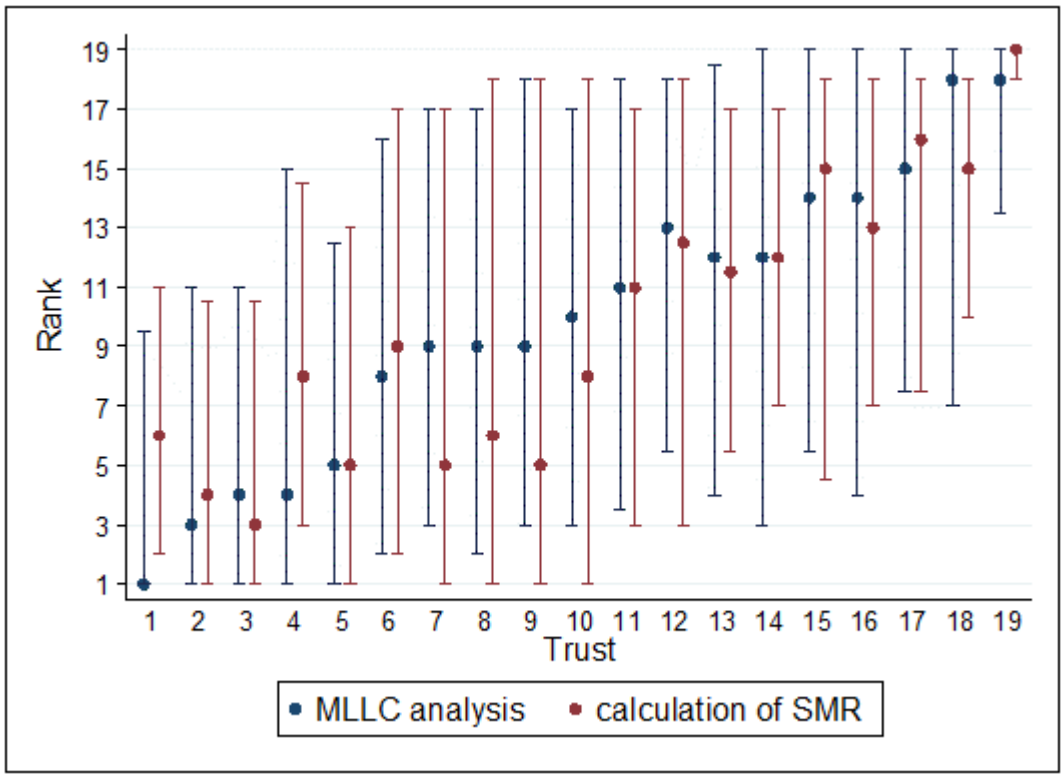


Figure 4.3 Trust median ranks and 95% CIs, ordered by the MLLC analysis

Differences in the median rank of Trust performance between the MLLC approach and the calculation of Trust SMRs are within their estimated CIs, which are very wide, indicating a large amount of heterogeneity remaining by both methods. The MLLC approach is thus comparable with the traditional approach.

4.4 Discussion

The MLLC modelling approach within this chapter incorporates available patient-level covariates and models patient-class uncertainty associated with unavailable patient-level covariates, to ensure that the resulting Trust-class differences in patient outcome are effectively adjusted for casemix. Model parameterisation has separated differential selection at the patient level (due to patient heterogeneity and casemix differences), from the potential causal structure of factors influencing Trust performance. Therefore, while the variation in Trust outcomes seen in Chapter 3 was attributable to explicit variation due to differential selection, in this chapter, differences are instead attributable to residual variation due to the influence of latent factors operating at the Trust level. This variation depends upon unmeasured Trust-level characteristics, for example, differences in healthcare delivery processes. This analytical strategy has considerable prognostic utility to inform health service providers of disparities within patient care.

Similar results are seen in Trust rankings across the two approaches, with estimates well within the CIs. The same general Trust-rank progression can be seen, with both methods broadly identifying the ‘best’ and ‘worst’ NHS Trusts based on patient outcome. Results are not identical, however, and those from the MLLC model should be preferred, as the method demonstrates a more sophisticated approach to the research question by accounting for uncertainty due to patient heterogeneity and measurement error. Use of the SMR does not fully accommodate patient casemix or imprecise measurements. The assumption of patient casemix only to the point of entry into the healthcare system is naïve, however, as heterogeneity may remain. Section 1.2.1 highlighted the complex relationships between the patient, treatment, and healthcare provider, thus the inclusion of treatment characteristics may further accommodate heterogeneity and narrow the CIs. This information is not available within the example dataset, however.

The probabilities of Trust-class membership in table 4.8 are marked, with most Trusts belonging entirely or predominantly to one Trust class, by the MLLC approach. This is unsurprising, as there is only a modest difference

between the two classes in median survival, and probabilistic assignment differentiates between the two, providing a class-weighted combined survival rate. It is not feasible, however, for a Trust to be assigned a class-weighted survival rate below that of the worst survival class, or above that of the best survival class. This is an implicit constraint on the estimated weighted survival for Trusts allocated entirely to one of the two Trust classes. To alleviate this, more Trust classes could be sought, increasing the number until no Trust has a probabilistic assignment of exactly one, for classes at the extremities of the range of Trust outcome means. Further inclusion of Trust classes is considered through simulations in Chapter 5, but as applied here, the estimated ranks are robust.

This chapter demonstrates an interim solution, extending the latent variable approach only as far as is feasible to still be able to make comparisons to a traditional approach. This comparison therefore shows proof of principle for the novel techniques. While the traditional methodologies cannot develop further, however, MLLC analysis offers improvement and extension to include both patient pathway adjustment (i.e. treatment differences, where available) and provider-level process variables. Further, the causal inference that may be investigated at the Trust level is now free from the patient-level differential selection issues that may conflate predictive modelling with causal inference modelling.

Chapter 5 explores the principle of evaluating causal factors operating to influence Trust performance, while accommodating patient casemix.

Chapter 5

Research Question (3); Provider-level Covariates

5.1 Introduction

Chapters 1 and 2 introduced and further defined the latent variable methodologies, putting MLLC approaches into context within an overarching causal framework. Model aspects and features were fully described, demonstrating how to precisely configure modelling approaches. Three research questions were posed, together with examination of the example dataset and a broad MLLC approach to each question.

Analysis within Chapter 3 applied this approach to the example dataset to investigate covariate effects at the patient level, while accounting for heterogeneity at the provider level. Analysis within Chapter 4 applied the same approach, with different parameterisation, to investigate performance comparison at the provider level, while accounting for differential selection (i.e. patient casemix) at the patient level. This analysis was necessarily simplified by not considering provider-level covariates.

Chapter 5 thus extends the investigation commenced in Chapter 4, to incorporate covariates at the provider level within a latent variable framework that also accounts for differential selection. Incorporation of organisational level features (such as surgeon speciality or available beds) can lead to improved comparisons and hence a more appropriate assessment of differences in levels of patient care. There are, however, no provider-level covariates in the example dataset. Data are therefore simulated, with both the data structure and the distribution of patient-level covariates based on values sourced from the example dataset.

The research question appropriate to this chapter is therefore:

(3) Can causal provider-level covariate effects be identified, after accommodating patient differences?

Use of simulated data allows for proof of principle to be demonstrated by the recovery of provider-level covariate effects as designed into the data simulations. Both continuous and binary outcome variables are investigated, with binary outcomes analogous to those explored in Chapters 3 and 4, i.e. three-year mortality. The continuous outcomes, as generated, do not represent a true survival outcome but may be considered to represent other outcomes, such as cost of care, and as such they are an important consideration for this and future research. There is no traditional comparison to the MLLC approach to this research question, as traditional techniques commonly adjust only for patient characteristics up to the point of entry into the healthcare system.

Both binary and continuous covariate effects are considered, although for simplification, they are analysed separately. A further simplification is to simulate a homogeneous patient group, such that focus may be placed on the accommodation of covariates at the provider level.

Section 5.2 describes the simulation approach, including consideration of the data structure, calculation of patient outcomes and of Trust-level coefficient effects. Simulated data combinations are described and the sensitivity of the approach is assessed.

Section 5.3 explores the modelling approach, considering the MLLC modelling strategy, adjustment for patient casemix, model parameterisation and construction, and the process of recovering the Trust-level coefficient values.

Section 5.4 contains all model results for both continuous and binary outcomes, with full interpretation, and further assessment as appropriate for each outcome and its findings.

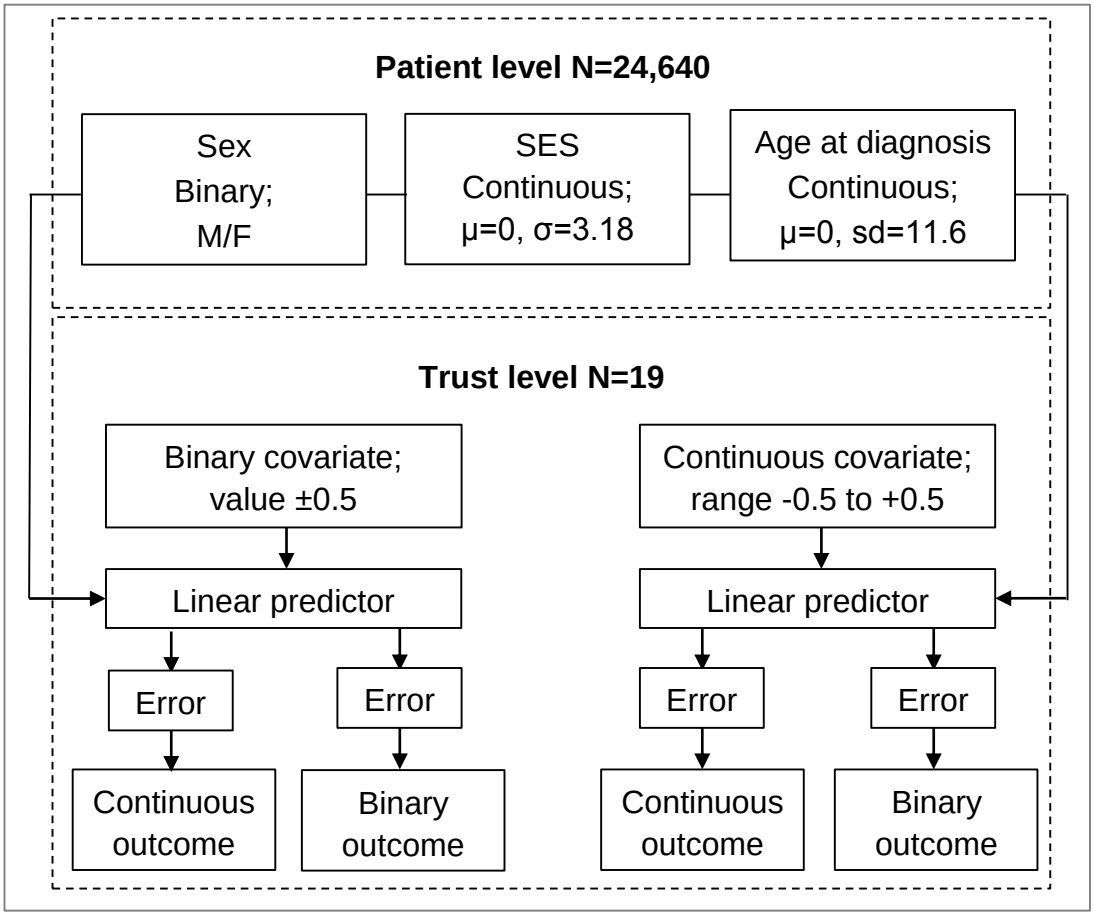
Section 5.5 provides a discussion of the methods and results.

The Stata code used for data simulation can be seen in Appendix C.

5.2 Simulation approach

5.2.1 Data structure

Figure 5.1 graphically displays the overarching simulation approach to the patient and Trust levels; 24,640 patients and nineteen Trusts are utilised, to correspond to the example dataset. Trust size is allowed to vary, thus reflecting differing Trust sizes by geographical area.



μ – mean, σ – standard deviation, N – total number of unique observations at patient or Trust level

Figure 5.1 Overarching simulation approach to the patient and Trust levels

The patient-level covariates sex, SES and age at diagnosis are simulated first using a trivariate covariance matrix, and values are drawn from a random normal distribution. Sex is defined as either male or female and there are approximately equal numbers of each. SES and age are centred on zero, with standard deviations as per descriptive statistics of the example

dataset (see table 2.3 in Chapter 2: Townsend deprivation score (used to measure SES) SD = 3.18, age SD = 11.6). Each patient is then assigned to one of nineteen Trusts, and randomisation ensures that no pattern is generated in the distribution of either Trusts or Trust-level covariates.

The binary Trust-level covariate equals approximately ±0.5, and a random normal distribution (with a small standard deviation of 0.01) is used to introduce some variability around these figures. Values are randomly allocated across Trusts, thus allowing for variability in the number of Trusts assigned each value within each simulated dataset. While this may widen the uncertainty when estimating the covariate effect due to boundary values that may not easily be modelled, it is a strength of the simulation to allow for a majority one way or the other as this reflects real-world Trust-level effects.

The continuous Trust-level covariate ranges from -0.5 to +0.5 across the nineteen Trusts, using a random uniform distribution to allocate values to Trusts. Values are generated without replacement, thus allowing for duplication, again to reflect real-world possibilities.

5.2.2 Patient outcomes

Patient outcomes are based on a linear predictor, as shown in figure 5.1, and calculated using the equation:

$$Linear\ predictor = \beta_{0i} + (\beta_{1i} \times sex) + (\beta_{2i} \times SEB) + (\beta_{3i} \times age) + (\beta_T \times Trust\ level\ covariate)$$

where β_{0i} is a constant term at the patient level i, β_{1i}, β_{2i} and β_{3i} are the effects of the patient-level covariates sex, SES and age respectively, and β_T is the coefficient effect of the Trust-level covariate (binary or continuous) as set during simulation. Values of β_{0i}, β_{1i}, β_{2i} and β_{3i} are log odds values taken from the MLM analysis of the example dataset in Chapter 3 (OR can be seen in table 3.3), with $\beta_{0i} = -0.0265$, $\beta_{1i} = -0.1368$, $\beta_{2i} = 0.0527$ and $\beta_{3i} = 0.0547$, and a range of β_T values (defined in section 5.2.3) are utilised. The linear predictor therefore includes the same effect of the patient covariates, but a different Trust-level effect dependent on the simulated values of the Trust-level covariate and coefficient effect.

Two outcomes are then generated for each Trust-level covariate. The continuous outcome is equal to the value of the linear predictor plus a normally distributed error term; a range of error variances are utilised, as described in section 5.2.4. The binary outcome is drawn from a random binomial distribution based on the inverse logit of the linear predictor, with a fixed error variance of $\pi^2/3$ at the patient level (Snijders and Bosker, 1999). This fixed binomial error term has implications on the effect of the variance structure at other levels, which is discussed further in section 5.5.

5.2.3 Trust-level coefficient effects

A range of values are utilised for the coefficient effect at the Trust level (β_T), to show consistency of recovery from the simulated value and to allow for graphical representation of the relationship between simulated and recovered values.

As an informed basis for analysis, patient-level coefficient values are selected from previous analyses; the MLM analysis of the example dataset performed in Chapter 3 is again utilised, with applicable effects recorded in section 5.2.2. For the binary Trust-level covariate, the absolute effect of sex is used (0.137) while for the continuous Trust-level covariate, the effect of deprivation is used (0.053). Five coefficient values are investigated for each Trust-level covariate: the effect of the chosen coefficient, one fifth the effect, five times the effect, and two additional values within this range. Table 5.1 summarises the values used for each Trust-level covariate.

Table 5.1 Trust-level coefficient values for the binary and continuous Trust-level covariates

β_T **Effect**	β_T **Coefficient Values**	
	Binary Trust-Level Covariate	**Continuous Trust-Level Covariate**
One fifth effect	0.027	0.011
Effect of sex or deprivation	0.137	0.053
Additional value	0.250	0.120
Additional value	0.500	0.200
Five times effect	0.684	0.264

5.2.4 Error variance

Normally distributed error terms are included in the calculation of each continuous outcome, as shown in figure 5.1. As the variance of these terms is unknown, a range of values are utilised; the error variance is calculated as 33%, 50% or 67% of the median variance of the outcome, when calculated without error. Error terms are therefore generated as normally distributed random numbers with mean equal to zero and variance equal to each error variance described above, before being added to the linear predictor to generate each continuous outcome. Table 5.2 summarises the values of the error variance used for each Trust-level covariate.

Table 5.2 Values of the error variance for the binary and continuous Trust-level covariates

Error variance (%)	Values of the Error Variance	
	Binary Trust-Level Covariate	**Continuous Trust-Level Covariate**
33%	0.150	0.144
50%	0.225	0.218
67%	0.300	0.293

5.2.5 Simulated data combinations

In addition to the five Trust-level coefficient values (β_T) described in section 5.2.3 and the three error variances described in section 5.2.4, three simulation seeds are also used to generate unique sets of 100 simulated datasets. Thus for the continuous outcome, forty-five sets are simulated, while for the binary outcome, fifteen sets are simulated (as there are no associated error variances). Table 5.3 summarises the combinations used; the same number of sets are required for each of the binary and continuous Trust-level covariates.

Table 5.3 Summary of combinations used in data simulation for both continuous and binary outcomes

Simulation seed	**Error variance**	**β_T coefficient**				
		One fifth effect	Effect of sex or deprivation	Additional value	Additional value	Five times effect
Continuous outcome						
Seed 1	33%	15 sets of 100 datasets				
Seed 2						
Seed 3						
Seed 1	50%	15 sets of 100 datasets				
Seed 2						
Seed 3						
Seed 1	67%	15 sets of 100 datasets				
Seed 2						
Seed 3						
Binary outcome						
Seed 1	N/A	15 sets of 100 datasets				
Seed 2						
Seed 3						

β_T coefficient effects differ by Trust-level covariate; sets are thus produced for each covariate.

5.2.6 Sensitivity of the simulation approach

Some of the choices described in the previous sections may have an impact on modelling outcomes. The sensitivity of these choices is therefore assessed.

In section 5.2.1, nineteen Trusts are simulated. This relatively small number may not give robust results for either fixed or random effects, therefore the implication of up to fifty Trusts may also be considered, as appropriate.

In section 5.2.1, Trust size is allowed to vary. Restricting Trust size to be the same throughout has no effect on the modelling outcomes.

In section 5.2.1, for the binary Trust-level covariate, ±0.5 is selected such that, if balanced across Trusts, values would average to zero. On balancing these values, there is little difference seen in the modelling outcomes. Examination showed that, when values were not balanced across Trusts, the smallest number of Trusts to be allocated to one of the binary categories is four. Additional uncertainty surrounding covariate effect estimates due to boundary values is thus minimised. Alternative values of 0/+1 or -1/0 are also considered, and the effect of eliminating the normally distributed variation is investigated, however neither option introduces much variability into the modelling outcomes obtained.

In section 5.2.1, for the continuous Trust-level covariate, the range from -0.5 to +0.5 is also selected such that, across all Trusts, values would average to zero. Although duplication of values is allowed, on investigation, all values were found to be unique within each simulated dataset.

In section 5.2.5, sets of 100 simulated datasets are described. Initial investigation increased the number to 1,000 (per combination of Trust-level coefficient value (β_T), error variance and simulation seed) but there is no measurable difference in the recovered values of β_T obtained. Therefore, to minimise computational requirements, 100 simulated datasets per combination are utilised.

5.3 Modelling approach

5.3.1 MLLC approach to the data

MLLC analysis is used to analyse the simulated data. While the simulation approach starts at the patient level and progresses upwards, modelling consideration starts at the Trust level with a latent construct, followed by casemix adjustment. There is therefore no direct overlap of simulation to analysis, thus allowing a more robust assessment of the analytical strategy.

These data have a hierarchical structure, with patients at the lower level and NHS Trusts at the upper level. Although patients are simulated as a single homogeneous group, accommodation is made for differential selection within the modelling strategy and parameterisation, and the use of a latent variable approach explicitly accommodates uncertainty within the latent structures at both levels. This approach may therefore also be utilised for heterogeneous patient groups, and thus, heterogeneous Trust groups (due to different patient casemix), as both are common within observational health datasets.

The modelling configuration is based on broad modelling strategy (ii), introduced in section 2.4.3, and demonstrated in Chapter 4, where patients are grouped into latent classes based on similarities in characteristics, while Trust classes are determined based on differences in patient characteristics. Trust classes therefore contain the same mixture of patient characteristics, i.e. they are balanced with respect to patient casemix. Differences in patient outcome are therefore due to effects operating at the Trust level. These effects are simulated within these data and interest thus lies in the comparison between simulated and recovered Trust-level coefficient values.

5.3.2 Casemix adjustment

Simulated values of sex, SES and age are included in the regression part of the model, at the patient level. No higher order terms, class predictors or inactive covariates are included. Bias is not a concern, since modelling at the patient level is required only to account for heterogeneity due to differential selection, and no causal inference will be made. Additional

covariates, such as stage at diagnosis, may therefore be included easily if available in observational data

MLLC modelling is utilised to partition modelling for prediction at the patient level from modelling for causal inference at the Trust level, although as binary and continuous Trust-level covariates are investigated separately, no assessment of potential causal relationships between covariates at the upper level are required for these data. Any combination of Trust-level effects could be incorporated however, with construction of a DAG to model relationships at the Trust level.

5.3.3 Parameterisation

Parameterisation, as introduced in section 2.4.4, is of a similar set-up to that described in section 4.2.3.

Intercepts. Class-independent intercepts are set such that identical contrasts may be made amongst patient classes, regardless of Trust class.

Covariate effects. Class-dependent covariate effects are set such that patient-class parameter values may vary across Trust classes, to allow covariate-outcome relationships to vary across the patient classes.

Class sizes. Class-independent patient-class sizes are required for modelling strategy (ii), to balance patient casemix across Trust classes.

Error variance. For the continuous outcome measure, class-independent error variances are adopted. This restricts error variances to be the same across the Trust classes, thus patient classes are set to be homoscedastic (i.e. the variance of the outcome remains the same within each patient class). This is appropriate, as no heteroscedasticity is built in to the simulated patient-level data.

5.3.4 Optimum model

A range of models are explored to allow for Trust-level variation. Two Trust classes are required as a minimum, in order to distinguish outcome differences, and this is increased as required to fully model variation at the Trust level. Models are not selected on the basis of model-evaluation

criteria, parsimony, CE or interpretability; rather, Trust classes are increased from two to a point of no further improvement in recovered Trust-level coefficient values.

Patient-level variation is incorporated by a single patient class, as patient-level data are simulated to be homogeneous. The parameterisation detailed in section 5.3.3 details how the patient classes are to be organised within the Trust classes and hence, with only one patient class, there is essentially no difference between Trust-class dependence or independence. Nevertheless, the intent is to model as described, which will appropriately accommodate any increase in the number of patient classes utilised in future modelling.

As Trust classes are increased, the time required to complete the modelling also increases. For each outcome and Trust-level covariate, fifteen sets of 100 datasets are simulated, as described in section 5.2.5, for each combination of error variance (continuous outcome only), Trust-level coefficient and simulation seed.

For models using two or three Trust classes, fifteen sets of 100 datasets can usually be analysed overnight. For four Trust classes, an additional half a day is often required, while for five Trust classes, a further overnight session should be allowed. Timings are estimated as they are affected by other computational issues such as disk space and system failures. Thus, to obtain the results seen in table 5.4, for example, where fifteen sets of 100 datasets are analysed using three different MLLC models (with two, three and four Trust-classes) across three error variances, a minimum of ten days may be required. If additional modelling using five Trust-classes is necessary, as seen in table 5.5, a further six days may be required.

There is no measureable difference in time required to perform the modelling for either nineteen or fifty Trusts. There is, however, a large increase in time required when considering MLLC models with ten Trust classes. To analyse just five sets of 100 datasets (considering only one simulation seed), using ten Trust classes, may take up to one week. Thus, two weeks are required to achieve the results seen in table 5.8.

These computational requirements necessarily limit the scope and range of the models considered within the time available, however sufficient models

are conducted and presented within this chapter to assess the utility of the methodological approach to recover simulated values of the Trust-level coefficient effects.

5.3.5 Trust-level coefficient recovery

Each simulated dataset is similarly modelled using the approach described. A weighted mean outcome for each Trust is calculated, based on the overall weighted mean outcome within each Trust class and the probabilistic assignment of each Trust to each Trust class. As highlighted in section 5.3.1, differences in mean outcome are due to simulated Trust-level covariate effects. Recovered values of the Trust-level coefficient (β_T) are therefore obtained by performing single level regression analysis to regress the Trust weighted mean outcome on the relevant binary or continuous Trust-level covariate. This process is repeated for each simulated dataset, with medians and credible intervals (CIs; 2.5% to 97.5%) calculated over each set of 100 datasets, for each combination of MLLC model, simulated Trust-level coefficient value, and error variance (where appropriate). Recovered values are averaged over the three simulation seeds.

5.4 Results

5.4.1 Continuous outcome

5.4.1.1 Binary Trust-level covariate

Table 5.4 shows the results of the analysis using a continuous outcome and a binary Trust-level covariate. Results are consistent across simulation seeds; models contained one patient class (1P) and up to four (4T) Trust classes.

For all combinations, the simulated values of the Trust-level coefficient (β_T) are found to be within the credible intervals for each recovered β_T value, and results are very consistent across the different models and error variances. In general, as the error variance increases, the credible intervals became gradually wider, as would be expected, but the difference is small. For example, for simulated β_T = 0.250, 1P-2T model, the 33% error returns a credible interval of 0.239 to 0.259, the 50% error returns 0.237 to 0.261 and the 67% error returns 0.235 to 0.263.

The median recovered β_T is almost identical to the simulated β_T for all simulated values except the lowest, regardless of error variance or MLLC model. There is some suggestion that, for the lowest simulated β_T = 0.027, the recovered β_T value reduces as the error variance is increased. At the 33% error variance, recovered β_T ranges from 0.017 to 0.018 across the MLLC models, at 50% it ranges from 0.014 to 0.015 and at 67%, this is 0.012 to 0.013. For this lowest simulated β_T, it was not possible to combine the results from all datasets in order to produce an estimate of the recovered β_T value. At the 33% error variance, across the MLLC models and simulation seeds, between 0 and 1 datasets are excluded from each set of 100. At the 50% error variance, this increased to between 0 and 5, and at the 67% error variance, to between 4 and 11. As would be expected, there are more datasets excluded for higher values of the error variance

Table 5.4 Simulated and recovered values of the Trust-level coefficient for the continuous outcome and binary Trust-level covariate; nineteen simulated Trusts

Error Variance	Simulated β_T Coefficient	MLLC Model	No. Datasets Excluded	Median Recovered β_T Coefficient (CI)
33% (0.150)	0.027	1P-2T	0-1	0.017 (0.005-0.030)
		1P-3T	0-1	0.018 (0.005-0.031)
		1P-4T	0-1	0.018 (0.005-0.031)
	0.137	1P-2T	0	0.137 (0.126-0.146)
		1P-3T	0	0.136 (0.126-0.146)
		1P-4T	0	0.136 (0.126-0.146)
	0.250	1P-2T	0	0.250 (0.239-0.259)
		1P-3T	0	0.250 (0.239-0.259)
		1P-4T	0	0.250 (0.239-0.259)
	0.500	1P-2T	0	0.499 (0.489-0.509)
		1P-3T	0	0.499 (0.489-0.509)
		1P-4T	0	0.499 (0.489-0.509)
	0.684	1P-2T	0	0.683 (0.672-0.693)
		1P-3T	0	0.683 (0.673-0.693)
		1P-4T	0	0.683 (0.673-0.693)
50% (0.225)	0.027	1P-2T	0-5	0.014 (0.002-0.029)
		1P-3T	0-5	0.015 (0.003-0.030)
		1P-4T	0-5	0.015 (0.003-0.030)
	0.137	1P-2T	0	0.136 (0.123-0.148)
		1P-3T	0	0.136 (0.123-0.148)
		1P-4T	0	0.136 (0.123-0.149)
	0.250	1P-2T	0	0.250 (0.237-0.261)
		1P-3T	0	0.250 (0.237-0.261)
		1P-4T	0	0.250 (0.237-0.261)
	0.500	1P-2T	0	0.499 (0.486-0.511)
		1P-3T	0	0.499 (0.486-0.511)
		1P-4T	0	0.499 (0.486-0.511)
	0.684	1P-2T	0	0.683 (0.670-0.695)
		1P-3T	0	0.683 (0.670-0.695)
		1P-4T	0	0.683 (0.670-0.695)

β_T – Trust-level coefficient value; median averaged over 3 simulation seeds; CI – Credible Interval.

Table 5.4 continued Simulated and recovered values of the Trust-level coefficient for the continuous outcome and binary Trust-level covariate; nineteen simulated Trusts

Error Variance	Simulated β_T Coefficient	MLLC Model	No. Datasets Excluded	Median Recovered β_T Coefficient (CI)
67% (0.300)	0.027	1P-2T	4-10	0.012 (0.002-0.028)
		1P-3T	4-10	0.013 (0.002-0.029)
		1P-4T	4-11	0.013 (0.002-0.029)
	0.137	1P-2T	0	0.136 (0.119-0.149)
		1P-3T	0	0.136 (0.118-0.150)
		1P-4T	0	0.136 (0.118-0.150)
	0.250	1P-2T	0	0.250 (0.235-0.263)
		1P-3T	0	0.249 (0.235-0.263)
		1P-4T	0	0.249 (0.235-0.263)
	0.500	1P-2T	0	0.499 (0.484-0.513)
		1P-3T	0	0.499 (0.484-0.512)
		1P-4T	0	0.499 (0.485-0.513)
	0.684	1P-2T	0	0.683 (0.668-0.697)
		1P-3T	0	0.683 (0.668-0.696)
		1P-4T	0	0.683 (0.668-0.697)

β_T – Trust-level coefficient value; median averaged over 3 simulation seeds; CI – Credible Interval.

The reason for these exclusions is that these datasets, when analysed, show exactly the same weighted mean outcome for each Trust class and the same probability of class membership for each of the nineteen Trusts. Hence, the weighted mean outcome by Trust is also identical for all Trusts and so the regression analysis cannot be performed, as the outcome does not vary. The β_T coefficient cannot therefore be recovered. There is consistency across the models i.e. no more or less datasets are excluded on average for the models using four Trust classes compared to those using two Trust classes. All simulated datasets are included in the results for all other values of β_T. It is hypothesised that, at very small values of β_T, the noise introduced when simulating the data dominates the value of the β_T coefficient and hence the modelling process is unable to divide the Trusts into identifiably different Trust classes. As the numbers of excluded datasets are relatively small, the simulated β_T = 0.027 value remains included in the results, although some bias may remain in the recovered β_T value. This may explain why these recovered β_T values are seen to reduce as the error variance is increased. Given this possibility of bias, it is reassuring that the simulated β_T = 0.027 value remains within the 95% credible intervals of the recovered values throughout.

Figure 5.2 shows the results from table 5.4 plotted by error variance, demonstrating that the line of equality (where recovered β_T equals simulated β_T) lies almost exactly on the data points and is well within the credible intervals. All MLLC models are included and no distinction is made between the number of Trust classes.

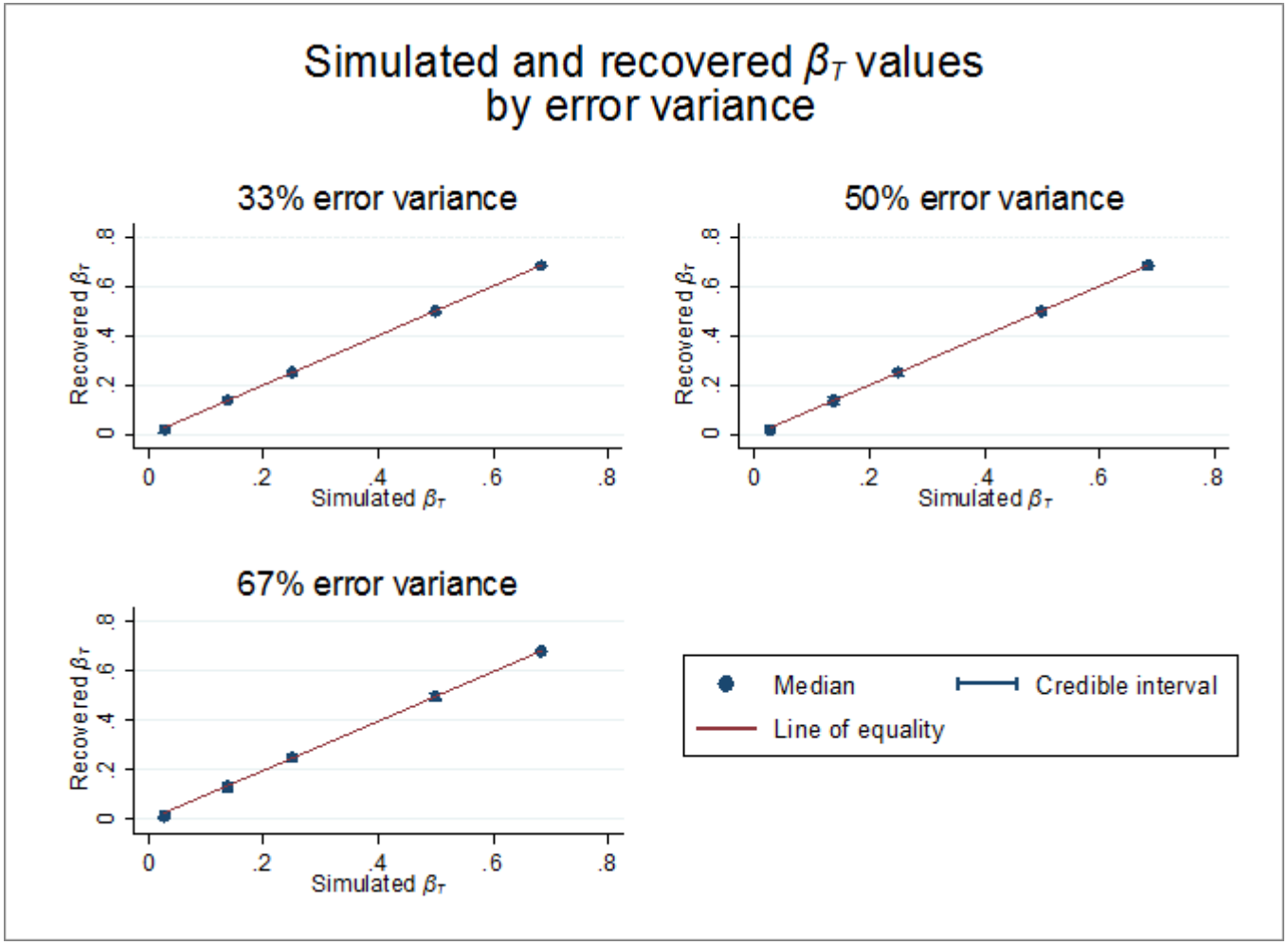


Figure 5.2 Plot showing β_T relationship for the continuous outcome and binary Trust-level covariate

5.4.1.2 Continuous Trust-level covariate

Table 5.5 shows the results of the analysis using a continuous outcome and a continuous Trust-level covariate, for nineteen simulated Trusts. Results are again consistent across simulation seeds. Again, just one patient class (1P) was used in the modelling, but up to five Trust classes (5T) are initially used to reflect the gradual improvement seen as the number of Trust classes are increased.

For all combinations, the median recovered values of the Trust-level coefficient (β_T) are lower than those simulated, although most simulated values are within the credible intervals for each recovered β_T value, for models with three Trust classes or more. As seen when modelling the binary Trust-level covariate, credible intervals widen as error variance increases, but for the continuous Trust-level covariate models, they also widen as the simulated β_T value is increased.

Estimates are better for smaller values of the error variance (e.g. for the simulated β_T = 0.120, 1P-2T model: 33% error variance returns 0.090, 50% error variance returns 0.087 and 67% error variance returns 0.085), and this pattern is seen for all simulated values of β_T except the lowest. This indicates that an increase in simulated error variance may be dominating the value of the β_T coefficient such that the modelling process is less able to separate the Trusts into distinct Trust classes. For the larger simulated β_T values, estimates also improve as the number of Trust classes are increased (e.g. for the simulated β_T = 0.120, 33% error variance: 1P-2T returns 0.090, 1P-3T returns 0.105, 1P-4T returns 0.107 and 1P-5T returns 0.109). This pattern is seen for all values of the error variance. This relationship is expected, as more Trust classes are required in order to fully distinguish differences between values of the continuous Trust-level covariate.

Table 5.5 Simulated and recovered values of the Trust-level coefficient for the continuous outcome and continuous Trust-level covariate; nineteen simulated Trusts

Error Variance	Simulated β_T Coefficient	MLLC Model	No. Datasets Excluded	Median Recovered β_T Coefficient (CI)
33% (0.144)	0.011	1P-2T	45-53	0.003 (-0.001 to 0.013)
		1P-3T	42-50	0.003 (-0.001 to 0.014)
		1P-4T	43-52	0.003 (-0.001 to 0.014)
		1P-5T	41-47	0.003 (-0.001 to 0.014)
	0.053	1P-2T	0	0.032 (0.011-0.051)
		1P-3T	0	0.036 (0.012-0.056)
		1P-4T	0	0.036 (0.012-0.056)
		1P-5T	0	0.036 (0.012-0.056)
	0.120	1P-2T	0	0.090 (0.063-0.113)
		1P-3T	0	0.105 (0.079-0.124)
		1P-4T	0	0.107 (0.084-0.127)
		1P-5T	0	0.109 (0.085-0.127)
	0.200	1P-2T	0	0.153 (0.113-0.184)
		1P-3T	0	0.182 (0.154-0.201)
		1P-4T	0	0.188 (0.165-0.207)
		1P-5T	0	0.191 (0.168-0.210)
	0.264	1P-2T	0	0.201 (0.153-0.241)
		1P-3T	0	0.240 (0.207-0.263)
		1P-4T	0	0.250 (0.223-0.269)
		1P-5T	0	0.254 (0.230-0.274)

β_T – Trust-level coefficient value; median averaged over 3 simulation seeds; CI – Credible Interval.

Table 5.5 continued Simulated and recovered values of the Trust-level coefficient for the continuous outcome and continuous Trust-level covariate; nineteen simulated Trusts

Error Variance	Simulated β_T Coefficient	MLLC Model	No. Datasets Excluded	Median Recovered β_T Coefficient (CI)
50% (0.218)	0.011	1P-2T	51-57	0.003 (-0.002 to 0.015)
		1P-3T	46-52	0.003 (-0.002 to 0.015)
		1P-4T	47-53	0.003 (-0.002 to 0.015)
		1P-5T	43-50	0.003 (-0.002 to 0.015)
	0.053	1P-2T	0-4	0.029 (0.008-0.052)
		1P-3T	0-4	0.031 (0.008-0.055)
		1P-4T	0-4	0.032 (0.008-0.055)
		1P-5T	0-3	0.031 (0.007-0.055)
	0.120	1P-2T	0	0.087 (0.058-0.115)
		1P-3T	0	0.101 (0.071-0.126)
		1P-4T	0	0.103 (0.073-0.126)
		1P-5T	0	0.104 (0.073-0.129)
	0.200	1P-2T	0	0.152 (0.111-0.185)
		1P-3T	0	0.180 (0.148-0.203)
		1P-4T	0	0.186 (0.159-0.209)
		1P-5T	0	0.188 (0.161-0.210)
	0.264	1P-2T	0	0.202 (0.149-0.242)
		1P-3T	0	0.240 (0.203-0.264)
		1P-4T	0	0.249 (0.218-0.271)
		1P-5T	0	0.252 (0.225-0.277)

β_T – Trust-level coefficient value; median averaged over 3 simulation seeds; CI – Credible Interval.

Table 5.5 continued Simulated and recovered values of the Trust-level coefficient for the continuous outcome and continuous Trust-level covariate; nineteen simulated Trusts

Error Variance	Simulated β_T Coefficient	MLLC Model	No. Datasets Excluded	Median Recovered β_T Coefficient (CI)
67% (0.293)	0.011	1P-2T	53-61	0.003 (-0.003 to 0.017)
		1P-3T	50-58	0.003 (-0.002 to 0.017)
		1P-4T	51-57	0.003 (-0.002 to 0.017)
		1P-5T	46-53	0.004 (-0.003 to 0.017)
	0.053	1P-2T	3-6	0.027 (0.003-0.052)
		1P-3T	3-6	0.028 (0.004-0.055)
		1P-4T	2-6	0.028 (0.004-0.054)
		1P-5T	2-6	0.028 (0.005-0.054)
	0.120	1P-2T	0	0.085 (0.054-0.115)
		1P-3T	0	0.098 (0.062-0.126)
		1P-4T	0	0.099 (0.063-0.127)
		1P-5T	0	0.100 (0.063-0.129)
	0.200	1P-2T	0	0.151 (0.109-0.186)
		1P-3T	0	0.178 (0.143-0.205)
		1P-4T	0	0.183 (0.149-0.209)
		1P-5T	0	0.186 (0.154-0.212)
	0.264	1P-2T	0	0.202 (0.147-0.243)
		1P-3T	0	0.238 (0.202-0.266)
		1P-4T	0	0.248 (0.213-0.274)
		1P-5T	0	0.251 (0.220-0.277)

β_T – Trust-level coefficient value; median averaged over 3 simulation seeds; CI – Credible Interval.

There is less of a pattern seen for the lowest simulated β_T values. In fact, for the lowest simulated β_T value of 0.011, neither the median returned value nor the credible interval show much difference at all across the error variances or models. Additionally, for the second lowest simulated β_T value of 0.053, while there is improvement for reduced amounts of error, and initial improvement between models containing two and three Trust classes, there is no additional improvement as the number of Trust classes are increased further.

As seen in analysis using the binary Trust-level covariate, not all datasets are able to be included in calculation of the median recovered β_T coefficient, for the same reasons. There are more datasets excluded here, however, for the same reasons as examined in section 5.4.1.1, and these figures can also be seen in table 5.5. For the lowest simulated β_T value of 0.011, between 41 and 61 datasets are excluded, with more exclusions seen at increased values of the error variance. For the second lowest simulated β_T value of 0.053, between 0 and 6 datasets are excluded, following the same pattern across values of the error variance. Exclusions again remain consistent across MLLC models. The lowest simulated β_T value of 0.011 is therefore excluded from further investigation into the relationship between simulated and recovered values, as too many datasets have been excluded to rely on the results seen. The second lowest value of 0.053 remains included, as numbers of exclusions are small, although some bias may remain. It is again reassuring that the simulated β_T = 0.053 value remains within credible intervals, for MLLC models with at least three Trust classes.

Figures 5.3, 5.4 and 5.5 show the results from table 5.5, excluding the lowest value of β_T = 0.011, plotted for 33%, 50% and 67% error variance respectively, and showing the gradually improving relationship between the simulated and recovered values of the Trust-level coefficient as the number of Trust classes are increased.

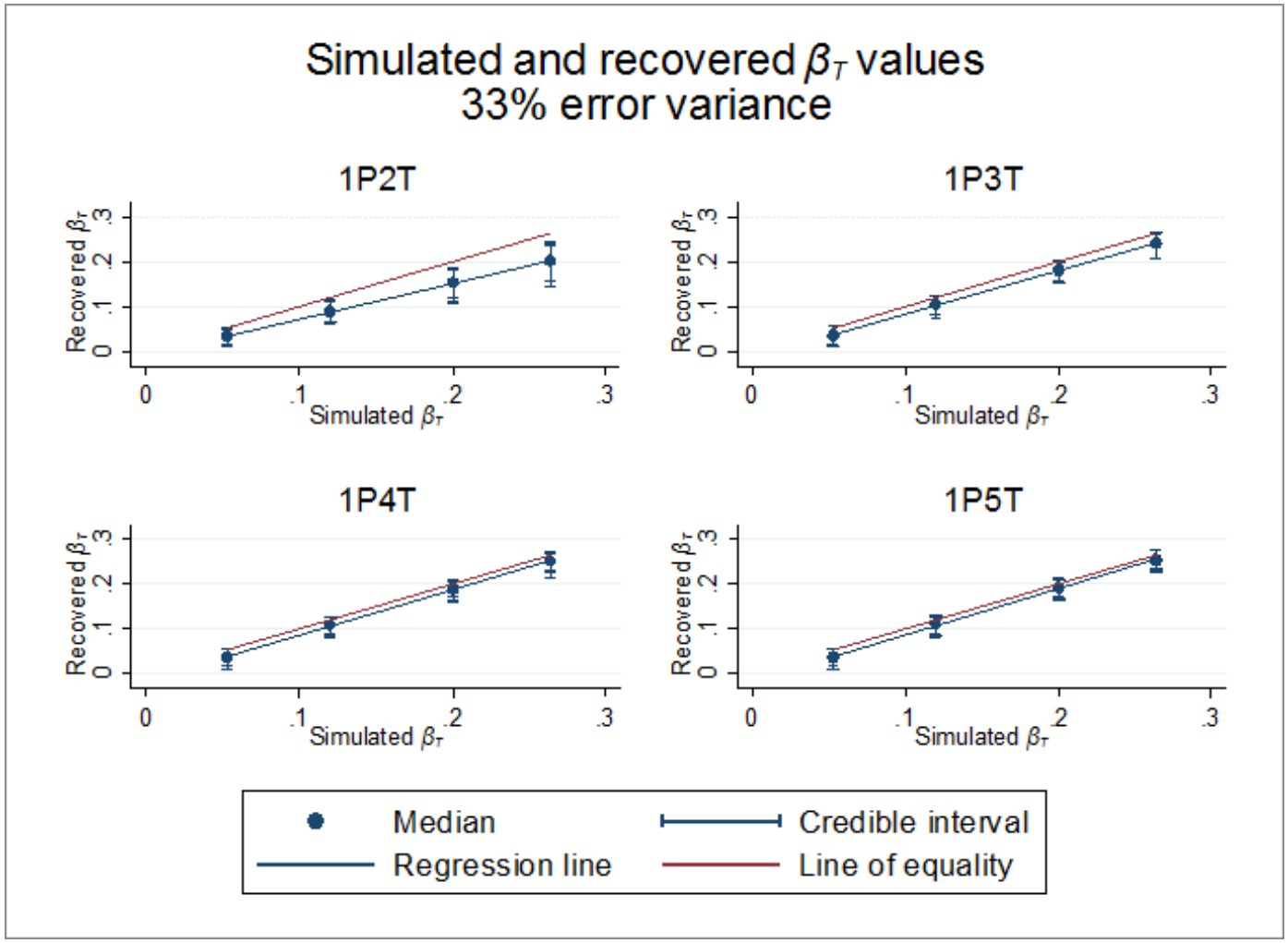


Figure 5.3 Plot showing β_T relationship for the continuous outcome and continuous Trust-level covariate; 33% error variance

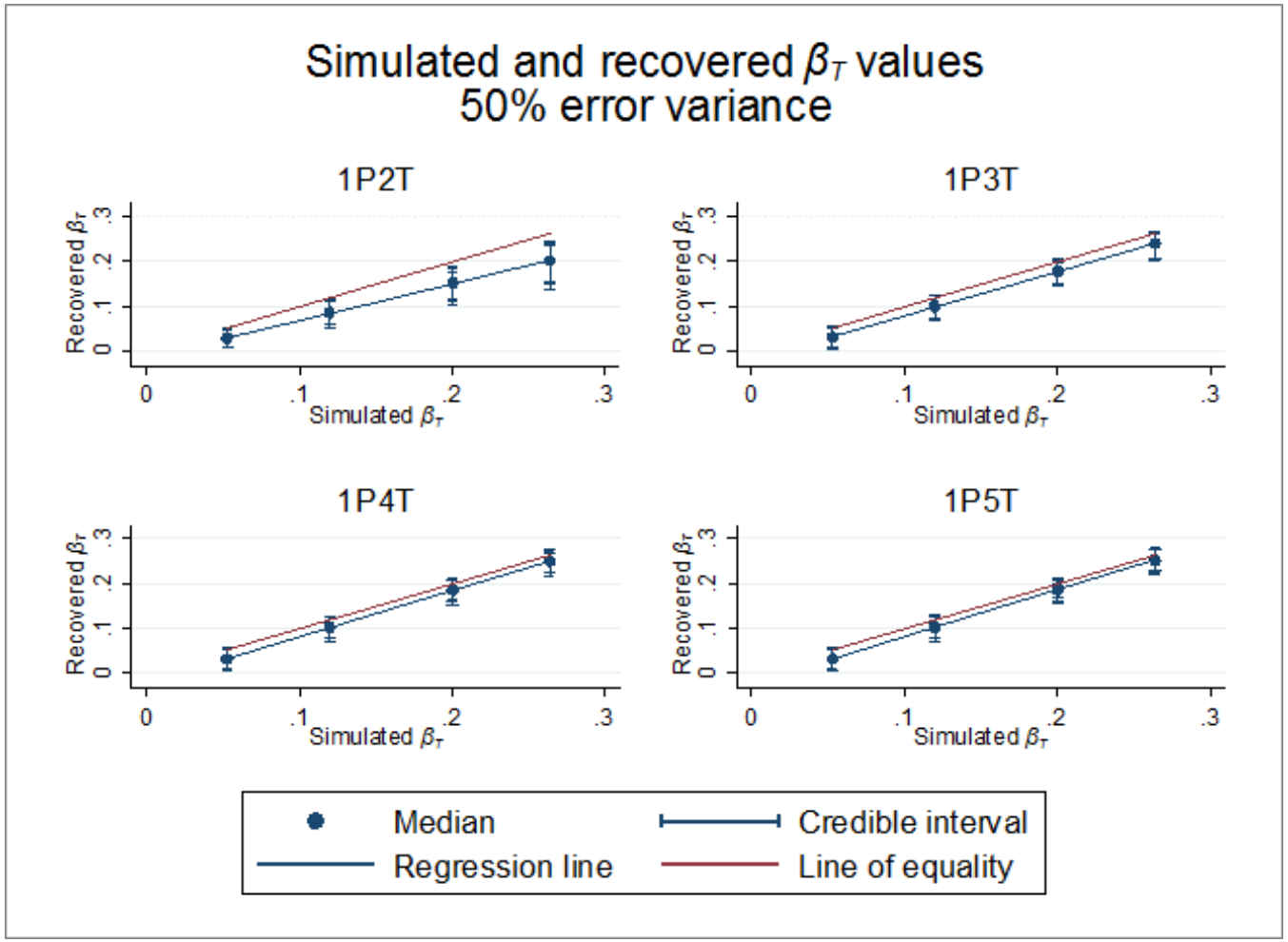


Figure 5.4 Plot showing β_T relationship for the continuous outcome and continuous Trust-level covariate; 50% error variance

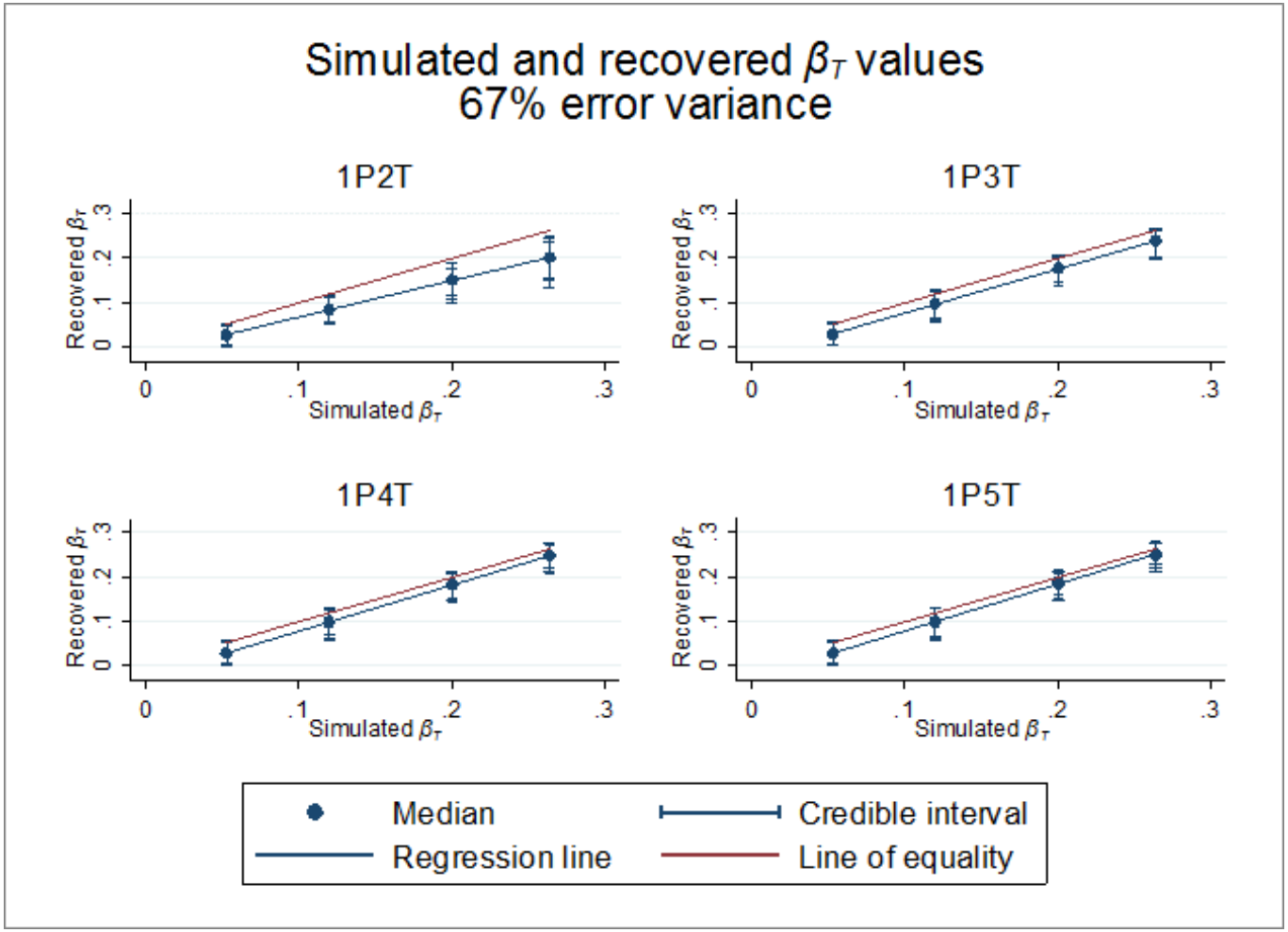


Figure 5.5 Plot showing β_T relationship for the continuous outcome and continuous Trust-level covariate; 67% error variance

5.4.1.3 Fifty Trusts

As discussed in section 5.2.6, interest lies in the implication of increasing the number of simulated Trusts from nineteen to fifty, to investigate the idea that use of a greater number of Trusts may lead to improved results. No other changes are made to the simulation process. Only 50% error variance is considered, as the intent here is merely to ascertain whether improved results are seen, rather than to replicate the entire set of results.

Table 5.6 shows the results of the analysis for the binary Trust-level covariate, while table 5.7 shows the results for the continuous Trust-level covariate. The same models are used for fifty Trusts, as were used for nineteen Trusts. For the binary Trust-level covariate, therefore, up to four (4T) Trust classes are considered, to compare directly to model results in table 5.4. For the continuous Trust-level covariate, up to five (5T) Trust classes are considered, to compare directly to model results in table 5.5.

For the binary Trust-level covariate, results are consistent across MLLC models, as also seen for nineteen Trusts. Recovered estimates, however, are reduced for fifty Trusts, for the lowest β_T values of 0.027, 0.137 and 0.250. Whilst, for nineteen Trusts, all simulated β_T values lie within the credible intervals of the recovered values, this is not the case for β_T = 0.027 when considering fifty Trusts. At this lowest value of β_T, credible intervals are much reduced. The rest of the recovered values lie within credible intervals, however.

For the continuous Trust-level covariate, recovered estimates are reduced for fifty Trusts, compared with those seen for nineteen Trusts, across all values of β_T. Recovered estimates increase as the number of Trust classes are increased, as also seen for nineteen Trusts, for all values of β_T, and credible intervals are generally narrower. Whilst, for nineteen Trusts, simulated β_T values lie within the credible intervals of the recovered values when considering at least three Trust classes, this is not the case for fifty Trusts, where most recovered values lie outside of the credible intervals for any number of Trust classes.

Table 5.6 Simulated and recovered values of the Trust-level coefficient for the continuous outcome and binary Trust-level covariate; fifty simulated Trusts

Error Variance	Simulated β_T Coefficient	MLLC Model	No. Datasets Excluded	Median Recovered β_T Coefficient (CI)
50% (0.225)	0.027	1P-2T	2-9	0.008 (0.001-0.019)
		1P-3T	2-9	0.008 (0.001-0.019)
		1P-4T	2-9	0.008 (0.001-0.019)
	0.137	1P-2T	0	0.133 (0.115-0.146)
		1P-3T	0	0.133 (0.118-0.147)
		1P-4T	0	0.133 (0.118-0.147)
	0.250	1P-2T	0	0.249 (0.235-0.261)
		1P-3T	0	0.249 (0.236-0.262)
		1P-4T	0	0.249 (0.233-0.262)
	0.500	1P-2T	0	0.499 (0.486-0.512)
		1P-3T	0	0.499 (0.483-0.512)
		1P-4T	0	0.499 (0.485-0.512)
	0.684	1P-2T	0	0.683 (0.667-0.696)
		1P-3T	0	0.683 (0.669-0.696)
		1P-4T	0	0.683 (0.670-0.696)

β_T – Trust-level coefficient value; median averaged over 3 simulation seeds; CI – Credible Interval; comparison to 19 Trusts in table 5.4.

Table 5.7 Simulated and recovered values of the Trust-level coefficient for the continuous outcome and continuous Trust-level covariate; fifty simulated Trusts

Error Variance	Simulated β_T	MLLC Model	No. Datasets Excluded	Median Recovered β_T Coefficient (CI)
50% (0.218)	0.011	1P-2T	51-60	0.001 (-0.003 to 0.010)
		1P-3T	47-59	0.002 (-0.003 to 0.010)
		1P-4T	46-58	0.002 (-0.003 to 0.011)
		1P-5T	43-59	0.002 (-0.003 to 0.013)
	0.053	1P-2T	4-6	0.018 (0.003-0.034)
		1P-3T	4	0.019 (0.003-0.037)
		1P-4T	4	0.019 (0.003-0.038)
		1P-5T	4	0.019 (0.003-0.037)
	0.120	1P-2T	0	0.076 (0.053-0.099)
		1P-3T	0	0.085 (0.056-0.107)
		1P-4T	0	0.086 (0.058-0.109)
		1P-5T	0	0.086 (0.058-0.109)
	0.200	1P-2T	0	0.144 (0.116-0.170)
		1P-3T	0	0.169 (0.141-0.190)
		1P-4T	0	0.174 (0.146-0.197)
		1P-5T	0	0.175 (0.146-0.199)
	0.264	1P-2T	0	0.195 (0.163-0.225)
		1P-3T	0	0.230 (0.202-0.251)
		1P-4T	0	0.239 (0.212-0.262)
		1P-5T	0	0.243 (0.214-0.266)

β_T – Trust-level coefficient value; median averaged over 3 simulation seeds; CI – Credible Interval; comparison to 19 Trusts in table 5.5.

5.4.1.4 Apparent β_T suppression

Analysis for nineteen Trusts using a continuous outcome and a binary Trust-level covariate demonstrates consistent results with median recovered β_T values seen to be almost identical to the simulated values for all values except the lowest. See section 5.4.1.1 and table 5.4 for these results. For a continuous outcome and a continuous Trust-level covariate, however, there is some attenuation of the effect seen. As indicated in section 5.4.1.2, and seen in table 5.5, median recovered β_T values are lower than those simulated throughout, although estimates are generally better both for smaller values of the error variance, and for a greater number of Trust classes.

When considering fifty Trusts, further attenuation is seen. For a continuous outcome and a binary Trust-level covariate, recovered estimates are reduced in comparison those seen for nineteen Trusts, for the lowest simulated β_T values, as indicated in section 5.4.1.3, and seen in table 5.6. For a continuous outcome and a continuous Trust-level covariate, recovered estimates are again reduced for all simulated values of β_T, when compared with those seen for nineteen Trusts, as also indicated in section 5.4.1.3, and seen in table 5.7. Similar to the results seen for nineteen Trusts, however, estimates are generally better for a greater number of Trust classes. Only a 50% error variance was used when considering fifty Trusts, so no further comment can be made on the effect of different sizes of error variance on the recovered estimates.

This apparent suppression of recovered β_T values is important when performing analyses using this methodological approach. Results showing the effect of Trust-level covariates should be interpreted cautiously, with consideration that effects seen may be lower than the ‘true’ effects, particularly for larger numbers of Trusts, greater values of the error variance, and smaller numbers of Trust classes.

5.4.1.5 Ten Trust classes

Tables 5.5 and 5.7 show that, for a continuous outcome and a continuous Trust-level covariate, there is a gradual improvement in estimates of the recovered value of β_T as the number of Trust classes are increased. Interest lies in whether this improvement continues beyond five Trust classes, for either nineteen or fifty Trusts.

Table 5.8 shows the results of the analysis using a one patient-class (1P), ten Trust-class (10T) MLLC model. Due to the computational resources required to run these larger models, only a 50% error variance is considered, and just one simulation seed is used to generate the simulated datasets.

Results do not generally show improvement compared to the one patient-class (1P) five Trust-class (5T) MLLC models for either nineteen or fifty Trusts. For nineteen simulated Trusts, all simulated values of β_T lie within the recovered credible intervals for both 5T and 10T models, whilst for fifty simulated Trusts, only simulated values of β_T = 0.011 or 0.264 lie within recovered credible intervals for either of the 5T or 10T models.

Table 5.8 Simulated and recovered values of the Trust-level coefficient for the continuous outcome and continuous Trust-level covariate; 1P-10T model

Error Variance	Simulated β_T	No. Trusts	No. Datasets Excluded	Median Recovered β_T Coefficient (CI)
50% (0.218)	0.011	19	48	0.004 (-0.001 to 0.016)
		50	54	0.001 (-0.004 to 0.011)
	0.053	19	4	0.032 (0.009 to 0.056)
		50	4	0.018 (0.003-0.035)
	0.120	19	0	0.104 (0.070-0.127)
		50	0	0.086 (0.051-0.109)
	0.200	19	0	0.190 (0.160-0.211)
		50	0	0.175 (0.144-0.199)
	0.264	19	0	0.256 (0.226-0.278)
		50	0	0.244 (0.212-0.266)

β_T – Trust-level coefficient value; simulation seed 1 only; CI – Credible Interval.

5.4.2 Binary outcome

Table 5.9 shows the results of the analysis using a binary outcome and a binary Trust-level covariate. Results are again consistent across simulation seeds; models contain one patient class (1P) and up to four (4T) Trust classes.

The simulated values of the Trust-level coefficient (β_T) are not found within the credible intervals for any of the recovered β_T values, with all recovered estimates lying considerably lower than the respective simulated values. Results are, however, very consistent across the different models, indicating that further modelling with more Trust classes is unlikely to improve the estimates. Again, not all datasets are able to be included in calculation of the average recovered β_T coefficient, for the same reasons as explained in section 5.4.1.1, with figures also shown in table 5.9. Exclusions again remain consistent across MLLC models. Due to the high number of exclusions, the lowest simulated value of β_T = 0.027 is therefore excluded from further investigation into the relationship between simulated and recovered values, while the second lowest value of β_T = 0.137 remains included, although some bias may be present.

Figure 5.6 shows the results from table 5.9, excluding the lowest value of β_T = 0.027. All MLLC models are included and no distinction is made between the number of Trust classes, as results are consistent. A linear regression line is added to aid interpretation and to show the pattern of the recovered β_T values compared with those simulated. It is clear that the line of equality (where recovered β_T equals simulated β_T) does not lie within the credible intervals of the recovered β_T values.

As yet, the rationale behind this observed relationship is unknown and further investigation is required. There may be a scalar effect operating to distort the relationship, which may be a function of the intraclass correlation coefficient (ICC), or the association may be more complex. Section 5.5 discusses the ICC, while section 6.5 considers suggestions for further study.

Table 5.9 Simulated and recovered values of the Trust-level coefficient for the binary outcome and binary Trust-level covariate; nineteen simulated Trusts

Simulated β_T Coefficient	MLLC Model	No. Datasets Excluded	Median Recovered β_T Coefficient (CI)
0.027	1P-2T	48-60	0.001 (-0.002 to 0.010)
	1P-3T	46-57	0.002 (-0.002 to 0.010)
	1P-4T	46-57	0.002 (-0.002 to 0.010)
0.137	1P-2T	1-2	0.018 (0.007 to 0.037)
	1P-3T	1-2	0.018 (0.007 to 0.037)
	1P-4T	1-2	0.018 (0.007 to 0.037)
0.250	1P-2T	0	0.051 (0.034 to 0.066)
	1P-3T	0	0.050 (0.034 to 0.066)
	1P-4T	0	0.050 (0.034 to 0.066)
0.500	1P-2T	0	0.112 (0.097 to 0.125)
	1P-3T	0	0.112 (0.097 to 0.126)
	1P-4T	0	0.113 (0.097 to 0.126)
0.684	1P-2T	0	0.154 (0.140 to 0.167)
	1P-3T	0	0.154 (0.140 to 0.168)
	1P-4T	0	0.154 (0.140 to 0.168)

β_T – Trust-level coefficient value; median averaged over 3 simulation seeds; CI – Credible Interval.

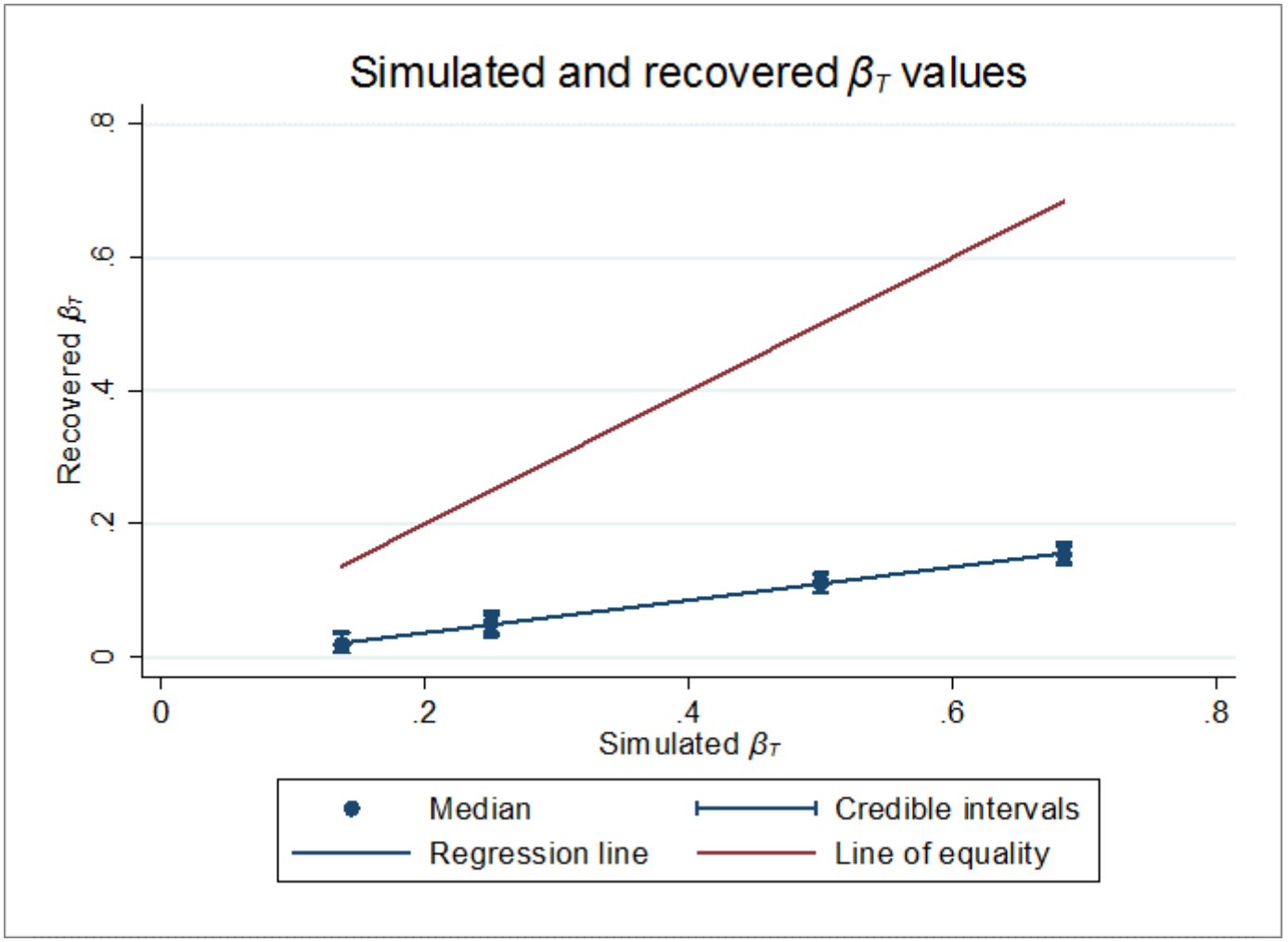


Figure 5.6 Plot showing β_T relationship for the continuous outcome and binary Trust-level covariate

5.5 Discussion

The MLLC modelling approach to the recovery of Trust-level coefficient values shows a successful recovery for both a binary and a continuous Trust-level covariate, when a continuous outcome is investigated. For the binary Trust-level covariate, recovered values are almost equal to the simulated values, for all except the lowest value, across MLLC models, simulated coefficients and error variances; all simulated values are within credible intervals of the recovered values. For the continuous Trust-level covariate, median estimates are lower than simulated values throughout, showing some attenuation of effect as described in section 5.4.1.4. Improvement is seen, however, as the number of Trust classes are increased; all simulated values are within credible intervals of the recovered values when the number of Trust classes are at least three. This difference is anticipated, as more Trust classes are required to fully distinguish differences between values of a continuous Trust-level covariate. Additional covariates between the binary and continuous could be explored, i.e. three or more categories would be simulated, and the same analysis performed.

Lower simulated values of the Trust-level covariate are not recovered as well as higher values. It is possible that the variation introduced during simulation dominates the coefficient value such that it is harder to identify within the modelling process. The additional variation in the continuous Trust-level covariate may also mean that a small simulated value may be even harder to identify, thus higher values of the Trust-level covariate are recovered more successfully.

The use of fifty Trusts does not improve estimates when compared to nineteen Trusts. Estimates, in fact, are lower, with some simulated values of the Trust-level coefficient lying outside recovered credible intervals. Whilst simulations with fifty Trusts can support more Trust classes compared with those with nineteen Trusts, there is no evidence that this is a solution. An increase to ten Trust classes does not improve estimates compared with MLLC models with up to five Trust classes, for simulated data with either nineteen or fifty Trusts.

When investigating a binary outcome, results are not as clear. Consideration is given to the possibility of a scalar effect operating to distort the relationship between simulated and recovered values of the Trust-level coefficient, which may be a function of the ICC, as discussed in section 5.4.2. In a MLLC model, the binomial error variance at the lower level is fixed at a value of $\pi^2/3$, as indicated in section 5.2.2, while the variance at the upper level is allowed to vary. The modelling process includes sex, SES and age at the lower level, thus explaining some of the variability at this level. Since the lower-level variance is fixed, however, this may impact on the upper-level variance, creating a scalar effect in the results. The effect of $\pi^2/3 = 3.290$ alone, however, does not appear to explain the relationship seen in figure 5.6. Further research is therefore required, which is discussed in section 6.5.

Nevertheless, the MLLC approach as demonstrated here has many advantages. Modelling for prediction and for causal inference are partitioned, thus performing adjustment for differential selection at the patient level, while allowing for a causal inference structure at the provider level. This approach is feasible only through use of latent variable methodologies; traditional approaches cannot replicate the modelling performed here. Although upper-level covariates are included individually in this analysis, there is much scope to extend by consideration of multiple covariates at the provider level. A multivariable DAG could be constructed at this level in order to adjust appropriately for multiple provider-level covariates within a single analytical framework, and is discussed further in section 6.5.

Chapter 6

Discussion

6.1 Introduction

This study set out to explore the utility of unexploited, novel statistical techniques in the analysis of complex observational health data. A more advanced modelling approach has been demonstrated that lies within an overarching causal framework, thus allowing for different modelling approaches to be partitioned across levels of a hierarchy, while accommodating uncertainty within the model covariates. Through the exploration of three research questions, aspects of the patient journey have been highlighted for detailed methodological consideration by latent variable techniques, and compared (where feasible) with traditional modelling approaches. The process was not, however, straightforward. While proof of principle has been demonstrated in most of the explored circumstances, care must be taken in the interpretation of results, due to inherent limitations within the available data.

Observational health data (as described in the Preface) are commonly obtained by routine data collection as patients progress through the healthcare system, with events recorded and linked to the patient. Thus, data are collected by distinct organisations, for differing purposes (Department of Health, 2011). Specifically for cancer registrations, data may be recorded such that incidence and mortality statistics can be calculated and monitored (The National Cancer Registration and Analysis Service, 2017), and to assess whether targets, such as the two week wait process (NICE, 2017b), are met within a population. As discussed in section 1.2.1, data have not been collected in response to a specific research question, which raises additional challenges to be addressed during analysis, due to the structure and complexity of the available data.

Within the field of cancer research, single level regression methods are most commonly used to analyse observational health data, as discussed in section 3.2.2. Multilevel analysis is not commonly used, as also discussed, despite the fact that there are clearly distinct organisations within which patients are diagnosed and treated. Stage at diagnosis is frequently included in analysis, although its relationship with other covariates may not be carefully considered or explored, as indicated in section 3.2.2. When accommodation for patient casemix is sought, approaches are commonly restricted to the casemix adjustment strategies discussed in section 1.3.4, for example, the use of direct or indirect standardisation procedures (Haynes et al., 2009; Fidler et al., 2015). MLLC analysis has not often been utilised in the context of observational health data, as indicated in section 1.4.5, with only one of the five multilevel latent variable approaches considering data similar to that described within this thesis.

It is possible that routinely collected data, such as that utilised here, may not be amenable to improved assessment or evaluation. Further, the requirements and rationale behind data collection, as described within this section, may not promote the production of datasets that benefit from sophisticated analytical developments to address service-relevant questions. Although this thesis demonstrates that latent variable methods can be applied, it should be recognised that observational health datasets may not be the most appropriate context within which to test or to develop these novel methods of data analysis.

Section 6.2 summarises the findings from analysis of the three research questions.

Section 6.3 considers the strengths and limitations of the analytical approach to the available data.

Section 6.4 discusses the implications of the study.

Section 6.5 considers recommendations for future research.

Section 6.6 offers conclusions to the thesis as a whole, taking into account strengths, limitations and future research.

6.2 Summary of findings

6.2.1 Research question (1)

(1) What is the relationship between a health exposure (SES) and outcome (three-year mortality), and what other factors affect this relationship?

For this research question, causal inference was sought at the patient level, whilst accommodating variation at the provider level. Use of MLLC analysis allowed for the appropriate modelling of patient-level covariates; stage at diagnosis was identified as a mediator of the SES-survival relationship, and is an imprecise measure, thus stage was removed from the regression part of the model and instead included as a class predictor. Bias due to both the reversal paradox and measurement error was therefore minimised. A range of DAGs demonstrated that alternative models could have been selected, which will be addressed in section 6.5.

MLLC modelling provided a better interpretation of the data, compared with the traditional MLM approach, and offered an enhanced interpretation based on three latent classes at the patient level. MLM found increasing Townsend deprivation score (as a measure of SES) and increasing age to increase the odds of death, while female gender decreased the odds. MLLC modelling found similar effects (although of a greater magnitude) for deprivation and age, while reduced odds of death for females were seen only in the good prognosis (early stage) class.

Five Trust-level latent classes were chosen, identifying outlying Trusts and thus indicating that a continuous latent variable at the Trust level (as required for MLM), was not sufficient to model these data. Differences in prognosis at the Trust level were due primarily to patient casemix differences.

Classification of patients and Trusts was not straightforward, and a different number of patient and Trust classes could have been chosen, potentially resulting in different results and interpretation. Sensitivity analyses are described in section 6.5.

6.2.2 Research question (2)

(2) How does the performance of a healthcare provider vary after accommodating patient differences?

For this research question, accommodation for differential selection was sought at the patient level such that differences in patient outcome at the provider level were then considered to be due to underlying organisational factors, rather than patient casemix. There was no concern regarding bias at the patient level, as causal inference was not required. Thus, all modelled covariates were included in the regression part of the model.

In the MLLC approach, two patient classes were chosen to model patient-level variability, and two Trust classes were identified, which showed a small but distinct difference in overall prognosis. A Trust performance ranking was allocated to each Trust based on its probability of membership of the best survival Trust class. In the traditional comparison, each Trust was assigned a rank based on its scaled difference from an SMR of one (where numbers of expected deaths equalled those of observed deaths).

The approaches were shown to be comparable, providing similar results with the same general Trust-rank progression. MLLC modelling is preferred, however, due to the use of a more sophisticated method that has accommodated both heterogeneity and measurement error. Confidence intervals were wide, and may be improved by the addition of treatment characteristics. This is discussed further in section 6.3.2.

Again, a different number of patient and Trust classes could have been chosen, potentially resulting in different results and interpretation. Sensitivity analyses are described in section 6.5.

6.2.3 Research question (3)

(3) Can causal provider-level covariate effects be identified, after accommodating patient differences?

This research question is an extension of research question (2), considering covariates at the provider level. Simulation was performed to assess proof of principle of the approach to accommodate patient casemix whilst performing causal modelling at the provider level. Assessment was made of the ability of the MLLC approach to recover simulated values of either a binary or continuous provider-level covariate, when combined with either a binary or continuous outcome measure. No traditional comparison was available.

A homogeneous patient group was simulated, hence MLLC models incorporated only one patient class. This is discussed further in section 6.5. A range of Trust classes were utilised, from two upwards until no further improvement was seen in recovered estimates.

Initial simulations generated nineteen Trusts, comparable to those within the example dataset. Consistent results were seen for a continuous outcome and a binary Trust-level covariate, with simulated values of the Trust-level coefficient within confidence intervals of the recovered values throughout. For the same outcome, with a continuous Trust-level covariate, estimates were slightly lower than simulated values throughout, showing some attenuation of effect, although they improved as the number of Trust classes were increased. Simulated values of the Trust-level coefficient were within confidence intervals of the recovered values, however, for three or more Trust classes. The use of fifty Trusts showed further attenuation of effect, which is addressed in section 6.5.

When considering a binary outcome and a binary Trust-level covariate, recovered values of the Trust-level coefficient were considerably lower than simulated values throughout. The rationale behind this observed association is unknown and thus requires further research, as discussed in section 6.5.

6.3 Strengths and limitations

6.3.1 Improvements over traditional techniques

In a traditional multilevel setting, where a continuous latent variable is adopted at the upper level, the implicit assumption is that provider-level outcomes have an underlying normal distribution (conditional on provider-level covariates) (see section 1.3.2). Healthcare providers are therefore effectively treated as a random sample of a larger (infinite) population. Providers are not, however, randomly placed geographically, nor are patients randomly assigned to providers (see section 1.2.3). The latent variable approach allows for parametric assumptions to be circumvented by the inclusion of discrete latent classes at the upper level, although there may remain a degree of geographical dependency that is not accounted for.

The use of latent classes at any level accommodates heterogeneity, with covariate relationships identified within each latent class, rather than over all observations, as seen in traditional regression approaches.

Within the latent variable approach, class membership models allow covariates to be removed from the regression part of the model, offering the capability to model appropriately moderators or mediators of an exposure-outcome relationship, and thus to minimise bias due to the reversal paradox. If imprecisely measured covariates are also included as class predictors, interactions are implicit, so exacerbated bias due to measurement error is also minimised. This approach was demonstrated for research question (1), and other scenarios may also benefit. Traditional approaches do not have this capability; all covariates are included within a single model, risking bias.

Modelling for differential selection can be accommodated at the patient level, as demonstrated for research question (2). As an advantage over traditional casemix adjustment techniques, latent variable approaches can also incorporate modelling for causal inference at the provider level, as demonstrated for research question (3). Traditional casemix adjustment strategies may increase bias and none can be extended to partition modelling approaches in the same manner (see section 1.3.4).

6.3.2 Limitations

Stage at diagnosis contained 13.1% missing data in the example dataset. To demonstrate proof of principle for the latent variable approach, this limitation was avoided by the inclusion of an additional category for stage, such that all observations could be included in analysis. The missing data category was therefore interpreted separately. Although considered sufficient for the focus of this thesis, generally, missing data techniques should be implemented. Techniques such as multiple imputation (MI) or inverse probability weighting (IPW) could be employed (Carpenter et al., 2006; Cattle et al., 2011; Seaman and White, 2013). Each method requires careful specification of the relevant imputation model for each variable with missing data, including investigation of parametric assumptions. Further, the congeniality principle (Meng, 1994) requires that the imputation models are compatible with the analytical model, which becomes more complex when there are interactions or non-linear relationships (Sterne et al., 2009; von Hippel, 2009; White et al., 2011). Ideally, latent variable modelling would lie within an integrated framework that includes approaches to the problems of missing data, but this is not yet resolved for large and complex datasets incorporating multilevel data. Methods are in development to (i) perform MI within a multilevel framework, and (ii) develop latent class approaches to imputation techniques that can be applied to Big Data. Missing data is hence an important challenge to be recognised and addressed in other work.

For research questions (1) and (2), survival was represented by a binary outcome of mortality status at three years. This was a necessary simplification in order to be comparable to existing research. Survival analysis in the standard methodological sense has not been explored within the proposed latent variable methodological framework. Results are therefore expressed as odds ratios of death within three years. It would be common to explore survival as a continuous measure, for example using Cox proportional hazards regression (Armitage et al., 2002), which is technically feasible within the proposed latent variable methodological framework, although more powerful software would be required, e.g. MPlus (Muthén and Muthén, 1998-2015), and this extension is left for future work.

SES (as a measure of socioeconomic deprivation) is included at the patient level, although it is derived at the small-area level (see section 1.2.2). Again, this limitation has not been addressed by the methodological approach. This can lead to the ecological fallacy, as discussed in section 1.2.2. Thus, strictly speaking, interpretation of the relationship between SES and survival should be made at the small-area level, while effects may vary for individuals within each small area. SES should ideally be considered as a separate level, effectively cross-classified with the Trust level, as discussed in section 1.2.5. More sophisticated and as yet unavailable alternative software would be required to accommodate a latent variable approach with a cross-classified level of analysis; MPlus is under development to achieve cross-classified Cox proportional hazards regression, thus eventually addressing the limitations of both a binary outcome and use of a small-area measure simultaneously. Much more complex modelling is possible with MLLC analysis, but all models must be thought through and the context in which they are interpreted carefully considered.

Patients were identified as attending the healthcare organisation at which they received their latest diagnosis (see section 2.3.3), so that all patients could be included in analysis regardless of whether or not they received treatment. There may be error introduced, however, due to patients attending multiple centres throughout their care. This error is mitigated by modelling with Trust of diagnosis at the upper level, rather than by hospital, as a higher proportion of patients who are treated remain within their Trust of diagnosis throughout their care (74.3%) compared with those who remain within their hospital of diagnosis throughout (59.2%). Nevertheless, some variability may remain, which could impact upon the CE of the chosen models. These inaccuracies could be further addressed by screening each patient journey to determine where the majority of interventions take place, or by using multilevel multiple membership models (Goldstein, 2011) for multiple treatment centres.

Accommodation for differential selection has thus far considered only patient characteristics. Ideally, patient pathways through the healthcare system (e.g. treatment effects) would also be included, as introduced in section 1.1, and considered with respect to research question (2) in section 4.4. While the

latent variable approach will extend to accommodate these, no information was available in the example dataset. Thus, their inclusion has not been addressed within the analysis, which may impact upon the width of the confidence intervals around the Trust ranks in figure 4.3, when comparing Trust performance for research question (2). The inclusion of treatment effects in the analysis may explain more of the variability in the outcome, potentially reducing the confidence intervals, if this information were available. Nevertheless, the latent variable approach may still offer an improved alternative over traditional techniques when evaluating treatment effects in an observational setting. As discussed in section 1.3.4, the propensity score is commonly used to accommodate patient differences and compare treatment subgroups, but this approach conflates modelling for prediction with that for causal inference.

Additionally, it may be challenging to make the analyses presented within this thesis accessible to potential users, such as health service policy makers or healthcare providers. The methodological approach applied when using the example dataset, i.e. for research questions (1) and (2), was unable to entirely address the difficulty in selection of the optimum number of both patient and Trust classes, as discussed in sections 6.2.1 and 6.2.2. Thus, the results and interpretations presented may be somewhat speculative. Discussion of future research in section 6.5 makes recommendations as to how best to achieve accessibility to users.

6.4 Implications of the study

This study has shown that latent variable modelling has utility in the analysis of a complex, hierarchical dataset, with improved accommodation of the data challenges in comparison to traditional techniques, and the ability to incorporate both differential selection and causal inference within the same modelling approach. It must be noted, however, that observational health datasets may not be the most appropriate context within which to explore these novel techniques, and that further research is necessary to minimise potentially speculative results and interpretation. Nevertheless, a generalised framework has been demonstrated, which offers the opportunity to utilise the methods in other contexts.

Potential area-level factors are numerous, and are not limited to known subgroups (e.g. communities, healthcare providers or schools); any environmental features utilised by groups of individuals may be modelled (e.g. access to fast food outlets, opportunities for physical activity or availability of green space). Possibilities to assess health outcomes are also numerous. For example, both the availability of fast food outlets, and opportunities for physical activity may impact on levels of obesity (Coombes et al., 2010; Williams et al., 2015). An integrated approach could therefore be pursued, considering the interplay between characteristics of both individuals and areas within an single analytical framework.

Further, preliminary investigation of observational data using latent variable techniques may inform prospective cluster-randomised trials targeted at improving public health outcomes. Trials can then focus on the modification of either individual or area characteristics identified by the MLLC approach as potential causes of differences in health outcomes. This pertains to existing approaches for quality improvement research, and is consistent with the principles of the Medical Research Council (MRC) framework for the development and evaluation of complex interventions (Craig et al., 2008). Latent variable techniques may also be utilised alongside the cluster-randomised trial approach (e.g. to assess compliance), to minimise the bias that may occur when small numbers of clusters are considered (Campbell et al., 2007) by accommodating uncertainty within the latent constructs.

6.5 Recommendations for future research

As discussed in sections 6.2.1 and 6.2.2, the selection of both patient and Trust classes for research questions (1) and (2) was not straightforward, which may lead to differing results and interpretation, and potentially speculative results, as discussed in section 6.3.2. In analysis for research question (1), the same patterns of association were seen for all model covariates when the number of patient classes was fixed at three, and different numbers of Trust classes were investigated (see table 3.10). Comprehensive sensitivity analyses should, however, be performed with respect to both patient and Trust classification. For both these research questions, alternative numbers of classes could have been selected, based on model-evaluation criteria. Further analyses should therefore be performed to investigate the effect of selecting the number of patient classes either side of the initial selection. Re-assessment of Trust classes would then follow for each new selection of patient classes, with comparisons made between the sets of results achieved. Thus, assessment could be made regarding the sensitivity of results and interpretation to the classification of both patients and Trusts. Further, for research question (2), more than two Trust classes may be required at the Trust level to show optimal utility from the latent variable approach, even if model likelihood statistics are not improved by an increased number of Trust classes.

In analysis for research question (1), stage at diagnosis was included as a class predictor in the MLLC model, while it was excluded entirely in the MLM analysis, as MLM does not have the capacity to model covariates as class predictors. MLM could, however, be stratified by stage, in order to make a more direct comparison between the two analytical methods. It must be noted, however, that this amounts to introducing stage at diagnosis explicitly as a covariate, which may introduce bias due to the reversal paradox where causal interpretation is sought. Equivalent results may be achieved by the exclusion of stage entirely in the MLLC analysis (Downing et al., 2010), although this approach may not be well received by healthcare organisations because, as considered in section 3.4, stage at diagnosis is commonly considered to be a key covariate.

Discrete latent variables have been utilised at the provider level throughout, to avoid the parametric assumptions required when using a continuous latent variable. It may be appropriate in some situations to consider both continuous and discrete latent variables within a single model approach. Studies have shown that use of a continuous latent variable, in place of discrete latent classes, may provide a better fit to the data (Downing et al., 2010; Henry and Muthén, 2010). Continuous and discrete latent variables, if combined, may prove more parsimonious, with variation within each provider-level class captured by the continuous latent variable, potentially leading to fewer provider classes needed to describe overall provider-level variation. For research question (2), use of a continuous Trust-level latent variable alongside the discrete Trust-level latent variables may alleviate the probabilities of Trust-class membership in table 4.8 being so marked, although the estimation of Trust survival rates would then become more complex.

In analysis for research question (3), simulated data were modelled in order to assess the utility of the latent variable modelling approach to recover simulated values of covariates at the Trust level. Nineteen Trusts were simulated initially, with an extension to fifty Trusts also considered. Modelling for a continuous outcome and a continuous Trust-level covariate showed some suppression of recovered values, as described in section 5.4.1.4, with increased suppression for fifty Trusts compared with nineteen Trusts. The rationale behind this effect is as yet unclear and additional research to further investigate simulations that generate larger numbers of Trusts (i.e. upper level units) would be beneficial.

Results seen for research question (3) with respect to the binary outcome were inconclusive (see sections 5.4.2 and 5.5). Further investigation is required to ascertain the rationale behind the observed association before proof of principle can be said to be complete. There are a number of potential approaches that could be employed. Much larger simulated values of the Trust-level coefficient could be explored to investigate whether or not the relationship remains linear over a wider range. Mathematical formulas could be studied, to investigate the expected relationship between the explained variance at each level of the hierarchy. Further estimation

procedures could be performed under controlled situations, perhaps by initially simulating very simple models without the inclusion of Trust-level covariates. Alternative statistical software could be utilised to reproduce models and compare results, or the parameters of the Latent GOLD software could be varied. Which method may yield the most useful information is, as yet, unknown.

In simulations for research question (3), analysis incorporated a single latent class only at the patient level, as data were simulated for a homogeneous patient group. As described in section 5.3.4, this meant that the parameterisation used to accommodate patient casemix was essentially irrelevant, as there were no multiple patient cases to be organised across the Trust classes. Nevertheless, this step was important, as it allowed demonstration of recovery of the provider-level covariates. An extension to the analysis as performed could be to simulate, and thus model, a heterogeneous patient group to assess whether there is any impact on Trust-level covariate recovery.

Analysis for research question (3) incorporated the assessment of individual provider-level covariates. Simulation methods can be extended to incorporate multiple covariates at the provider level, as indicated in section 5.3.2, with causal inference modelling supported by the construction of a DAG at this level. Thus, many potential provider-level effects may be modelled together, including both binary (e.g. whether or not the surgeon is a specialist) and continuous (e.g. volume of procedures). Interest may lie in the extent to which additional complexities in casemix (e.g. patient pathway variables; discussed in section 6.3.2) and multiple provider-level covariates (both competing exposures and confounders) may dilute the precision of estimates sought for a main provider-level covariate under investigation.

The utility of DAGs can thus be explored further, and extended to other applications or different healthcare contexts. While identification of covariate relationships is relatively straightforward for small numbers of variables, construction of a DAG becomes more complex if many covariates are included, for example in Big Data. As such, a web application ‘DAGitty’ (Textor et al., 2011) has been developed, whereby causal diagrams can be

produced and analysed. DAGitty can identify sets of covariates that should be included in the model in order to minimise bias in the estimate of the main exposure-outcome effect, thus aiding identification of appropriate models when using larger datasets. Additionally, the R package ‘dagitty’ has been developed (Textor et al., 2016), whereby the functionally of the web application, and additional capabilities, can be accessed from within the R software.

Finally, as raised in section 6.3.2, there are challenges to be addressed in order to make the work within this thesis accessible to potential users within healthcare organisations. Although the methodology presented offers a novel, integrated approach to the analysis of complex observational health data, with improvements seen over traditional techniques, findings remain somewhat speculative due to the limitations previously discussed. A guide to ‘best practice’ in the use of MLLC methodological approaches could assist. This guidance document, publishable as a review article, would consider the use of multilevel latent variable approaches within health research to accommodate patient casemix, aid performance comparison, and thus identify potentially causal effects that may affect provider performance. Publishing of analyses presented so far has been piecemeal, with modelling approaches described independently with reference to data-specific research questions (Gilthorpe et al., 2011; Harrison et al., 2012; Harrison et al., 2013). It would aid uptake of the methods presented to make explicit the integrated analytical approach such that other researchers could utilise the same techniques, with adaptation or extension as applicable to their data. It remains, however, that observational health datasets may not be the most appropriate context within which to explore the methods.

6.6 Conclusions

In conclusion, a generalised multilevel latent framework has been developed whereby individual and area characteristics can be modelled appropriately to assess their contribution to a health outcome. The research detailed within this thesis demonstrates one area of application, in the field of cancer survival, showing speculative results and interpretation due to inherent limitations within observational health data and the appropriateness of the methods within this context. Nevertheless, the framework may offer an enhanced analytical approach with extension to accommodate missing data, and future research allows opportunities to further extend the capabilities of the approach in the area of causal inference modelling.

References

Abel, GA, Saunders, CL and Lyratzopoulos, G. 2014. Cancer patient experience, hospital performance and case mix: evidence from England. *Future Oncology.* **10**(9), pp.1589-1598. https://dx.doi.org/10.2217/fon.13.266

Ahmed, S, Howel, D, Debrah, S and Northern Region Colorectal Cancer Audit Group. 2014. The influence of age on the outcome of treatment of elderly patients with colorectal cancer. *Journal of Geriatric Oncology.* **5**(2), pp.133-140. http://dx.doi.org/10.1016/j.jgo.2013.12.005

Ait Ouakrim, D, Pizot, C, Boniol, M, Malvezzi, M, Boniol, M, Negri, E, Bota, M, Jenkins, MA, Bleiberg, H and Autier, P. 2015. Trends in colorectal cancer mortality in Europe: retrospective analysis of the WHO mortality database. *British Medical Journal.* **351**, h4970. http://dx.doi.org/10.1136/bmj.h4970

Akaike, H. 1974. A new look at the statistical identification model. *Institute of Electrical and Electronics Engineers Transactions on Automatic Control.* **19**(6), pp.716-723. http://dx.doi.org/10.1109/TAC.1974.1100705

Anderson, TW. 1954. On estimation of parameters in latent structure analysis. *Psychometrika.* **19**(1), pp.1-10. http://dx.doi.org/10.1007/BF02288989

Anderson, TW. 1959. Some scaling methods and estimation procedures in the latent class model. In: Grenander, U ed. *Probability and statistics.* New York: Wiley.

Arem, H, Park, Y, Felix, AS, Zervoudakis, A, Brinton, LA, Matthews, CE and Gunter, MJ. 2015. Reproductive and hormonal factors and mortality among women with colorectal cancer in the NIH-AARP Diet and Health Study. *British Journal of Cancer.* **113**(3), pp.562-568. http://dx.doi.org/10.1038/bjc.2015.224

Arminger, G and Küsters, U. 1989. Construction principles for latent trait models. *Sociological Methodology.* **19**, pp.369-393. http://dx.doi.org/10.2307/270958

Armitage, P, Berry, G and Matthews, JNS. 2002. *Statistical methods in medical research.* 4th ed. Massachusetts: Blackwell.

Askari, A, Malietzis, G, Nachiappan, S, Antoniou, A, Jenkins, J, Kennedy, R and Faiz, O. 2015. Defining characteristics of patients with colorectal cancer requiring emergency surgery. *International Journal of Colorectal Disease.* **30**(10), pp.1329-1336. http://dx.doi.org/10.1007/s00384-015-2313-8

Auer, MF, Hickendorff, M, Van Putten, CM, Beguin, AA and Heiser, WJ. 2016. Multilevel latent class analysis for large-scale educational assessment data: exploring the relation between the curriculum and students' mathematical strategies. *Applied Measurement in Education.* **29**(2), pp.144-159. http://dx.doi.org/10.1080/08957347.2016.1138959

Bartholomew, DJ. 1980. Factor analysis for categorical data. *Journal of the Royal Statistical Society (Series B).* **42**(3), pp.293-321.

Beeken, RJ, Wilson, R, McDonald, L and Wardle, J. 2014. Body mass index and cancer screening: findings from the English Longitudinal Study of Ageing. *Journal of Medical Screening.* **21**(2), pp.76-81. http://dx.doi.org/10.1177/0969141314531409

Bennink, M, Croon, MA, Keuning, J and Vermunt, JK. 2014. Measuring student ability, classifying schools, and detecting item bias at school level, based on student-level dichotomous items. *Journal of Educational and Behavioral Statistics.* **39**(3), pp.180-202. http://dx.doi.org/10.3102/1076998614529158

Bharathan, B, Welfare, M, Borowski, DW, Mills, SJ, Steen, IN, Kelly, SB and Northern Region Colorectal Cancer Audit Group. 2011. Impact of deprivation on short- and long-term outcomes after colorectal cancer surgery. *British Journal of Surgery.* **98**(6), pp.854-865. http://dx.doi.org/10.1002/bjs.7427

Borowski, DW, Bradburn, DM, Mills, SJ, Bharathan, B, Wilson, RG, Ratcliffe, AA, Kelly, SB and Northern Region Colorectal Cancer Audit Group. 2010. Volume-outcome analysis of colorectal cancer-related outcomes. *British Journal of Surgery.* **97**(9), pp.1416-1430. http://dx.doi.org/10.1002/bjs.7111

Borowski, DW, Kelly, SB, Bradburn, DM, Wilson, RG, Gunn, A, Ratcliffe, AA and Northern Region Colorectal Cancer Audit Group. 2007. Impact of

surgeon volume and specialization on short-term outcomes in colorectal cancer surgery. *British Journal of Surgery.* **94**(7), pp.880-889. http://dx.doi.org/10.1002/bjs.5721

Bostan, C, Oberhauser, C, Stucki, G, Bickenbach, J and Cieza, A. 2015. Which environmental factors are associated with lived health when controlling for biological health? - a multilevel analysis. *BioMed Central Public Health.* **15**(508). http://dx.doi.org/10.1186/s12889-015-1834-y

Boyd, D and Crawford, K. 2012. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication and Society.* **15**(5), pp.662-679. http://dx.doi.org/10.1080/1369118X.2012.678878

Breslow, NE and Day, NE. 1987. *Statistical methods in cancer research. Volume II - the design and analysis of cohort studies.* International Agency for Research on Cancer Scientific Publications. 82, pp.1-406. Available from: http://www.iarc.fr/en/publications/pdfs-online/stat/sp82/. Accessed: 22 Aug 2016.

Brewster, DH, Clark, DI, Stockton, DL, Munro, AJ and Steele, RJ. 2011. Characteristics of patients dying within 30 days of diagnosis of breast or colorectal cancer in Scotland, 2003-2007. *British Journal of Cancer.* **104**(1), pp.60-67. http://dx.doi.org/10.1038/sj.bjc.6606036

Burns, EM, Bottle, A, Almoudaris, AM, Mamidanna, R, Aylin, P, Darzi, A, Nicholls, RJ and Faiz, OD. 2013. Hierarchical multilevel analysis of increased caseload volume and postoperative outcome after elective colorectal surgery. *British Journal of Surgery.* **100**(11), pp.1531-1538. http://dx.doi.org/10.1002/bjs.9264

Burns, RA, Byles, J, Magliano, DJ, Mitchell, P and Anstey, KJ. 2015. The utility of estimating population-level trajectories of terminal wellbeing decline within a growth mixture modelling framework. *Social Psychiatry and Psychiatric Epidemiology.* **50**(3), pp.479-487. http://dx.doi.org/10.1007/s00127-014-0948-3

Campbell, MJ, Donner, A and Klar, N. 2007. Developments in cluster randomized trials and Statistics in Medicine. *Statistics in Medicine.* **26**(1), pp.2-19. https://dx.doi.org/10.1002/sim.2731

Cardwell, CR, Hicks, BM, Hughes, C and Murray, LJ. 2014. Statin use after colorectal cancer diagnosis and survival: a population-based cohort study. *Journal of Clinical Oncology.* **32**(28), pp.3177-3183. http://dx.doi.org/10.1200/JCO.2013.54.4569

Carpenter, JR, Kenward, MG and Vansteelandt, S. 2006. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society (Series A).* **169**(3), pp.571-584. http://dx.doi.org/10.1111/j.1467-985X.2006.00407.x

Carroll, RJ, Ruppert, D, Stefanski, LA and Crainiceanu, C. 2006. *Measurement error in nonlinear models: a modern perspective.* 2nd ed. London: Chapman and Hall.

Cattle, BA, Baxter, PD, Greenwood, DC, Gale, CP and West, RM. 2011. Multiple imputation for completion of a national clinical audit dataset. *Statistics in Medicine.* **30**(22), pp.2736-2753. http://dx.doi.org/10.1002/sim.4314

Cohen, ME, Dimick, JB, Bilimoria, KY, Ko, CY, Richards, K and Hall, BL. 2009. Risk adjustment in the American College of Surgeons National Surgical Quality Improvement Program: A comparison of logistic versus hierarchical modeling. *Journal of the American College of Surgeons.* **209**(6), pp.687-693. http://dx.doi.org/10.1016/j.jamcollsurg.2009.08.020

Coombes, E, Jones, AP and Hillsdon, M. 2010. The relationship of physical activity and overweight to objectively measured green space accessibility and use. *Social Science and Medicine.* **70**(6), pp.816-822. https://dx.doi.org/10.1016/j.socscimed.2009.11.020

Craig, P, Dieppe, P, Macintyre, S, Michie, S, Nazareth, I, Petticrew, M and Medical Research Council, G. 2008. Developing and evaluating complex interventions: the new Medical Research Council guidance. *British Medical Journal.* **337**(a1655). http://dx.doi.org/10.1136/bmj.a1655

D'Agostino, RB, Jr. 1998. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine.* **17**(19), pp.2265-2281.

da Costa, LP and Dias, JG. 2014. Perceptions of poverty attributions in Europe: a multilevel mixture model approach. *Quality and Quantity: International Journal of Methodology.* **48**(3), pp.1409-1419. http://dx.doi.org/10.1007/s11135-013-9843-3

da Costa, LP and Dias, JG. 2015. What do Europeans believe to be the causes of poverty? A multilevel analysis of heterogeneity within and between countries. *Social Indicators Research.* **122**(1), pp.1-20. http://dx.doi.org/10.1007/s11205-014-0672-0

Damman, OC, Stubbe, JH, Hendriks, M, Arah, OA, Spreeuwenberg, P, Delnoij, DM and Groenewegen, PP. 2009. Using multilevel modeling to assess case-mix adjusters in consumer experience surveys in health care. *Medical Care.* **47**(4), pp.496-503. http://dx.doi.org/10.1097/MLR.0b013e31818afa05

Davy, M. 2007. Socio-economic inequalities in smoking: an examination of generational trends in Great Britain. *Health Statistics Quarterly.* (34), pp.26-34.

Deeks, JJ, Dinnes, J, D'Amico, R, Sowden, AJ, Sakarovitch, C, Song, F, Petticrew, M and Altman, DG. 2003. Evaluating non-randomised intervention studies. *Health Technology Assessment.* **7**(27), pp.iii-x, 1-173. http://dx.doi.org/10.3310/hta7270

Department for Business, Innovation and Skills. 2013. *Eight great technologies: infographics.* [Online]. [Accessed 12 August 2016]. Available from: https://www.gov.uk/government/publications/eight-great-technologies-infographics

Department of Health. 2011. *Sources of routine mortality and morbidity data, including primary care data, and how they are collected and published at international, national, regional and district levels.* [Online]. [Accessed 07 February 2017]. Available from: https://www.healthknowledge.org.uk/public-

health-textbook/health-information/3b-sickness-health/collection-routine-ad-hoc-data

Derks, EM, Boks, MPM and Vermunt, JK. 2012. The identification of family subtype based on the assessment of subclinical levels of psychosis in relatives. *BioMed Central Psychiatry.* **12**(71). http://dx.doi.org/10.1186/1471-244X-12-71

Dixon, A, Appleby, J, Robertson, R, Burge, P, Devlin, N and Magee, H. 2010. *Patient choice. How patients choose and how providers respond.* Available from: http://www.kingsfund.org.uk/publications/patient-choice. Accessed: 30 Aug 2016.

Downing, A, Aravani, A, Macleod, U, Oliver, S, Finan, PJ, Thomas, JD, Quirke, P, Wilkinson, JR and Morris, EJ. 2013. Early mortality from colorectal cancer in England: a retrospective observational study of the factors associated with death in the first year after diagnosis. *British Journal of Cancer.* **108**(3), pp.681-685. http://dx.doi.org/10.1038/bjc.2012.585

Downing, A, Harrison, WJ, West, RM, Forman, D and Gilthorpe, MS. 2010. Latent class modelling of the association between socioeconomic background and breast cancer survival status at 5 years incorporating stage of disease. *Journal of Epidemiology and Community Health.* **64**(9), pp.772-776. http://dx.doi.org/10.1136/jech.2008.085852

Dukes, CE. 1949. The surgical pathology of rectal cancer. *Journal of Clinical Pathology.* **2**(2), pp.95-98. http://dx.doi.org/10.1136/jcp.2.2.95

Efron, B and Tibshirani, RJ. 1993. *An introduction to the bootstrap. Monographs on statistics and applied probability.* Boca Raton, FL: Chapman and Hall/CRC.

Evans, L and Best, C. 2014. Accurate assessment of patient weight. *Nursing Times.* **110**(12), pp.12-14.

Faiz, O, Brown, T, Bottle, A, Burns, EM, Darzi, AW and Aylin, P. 2010. Impact of hospital institutional volume on postoperative mortality after major emergency colorectal surgery in English National Health Service Trusts, 2001 to 2005. *Diseases of the Colon and Rectum.* **53**(4), pp.393-401. http://dx.doi.org/10.1007/DCR.0b013e3181cc6fd2

Faiz, O, Haji, A, Bottle, A, Clark, SK, Darzi, AW and Aylin, P. 2011. Elective colonic surgery for cancer in the elderly: an investigation into postoperative mortality in English NHS hospitals between 1996 and 2007. *Colorectal Disease.* **13**(7), pp.779-785. http://dx.doi.org/10.1111/j.1463-1318.2010.02290.x

Fidler, MM, Ziff, OJ, Wang, S, Cave, J, Janardhanan, P, Winter, DL, Kelly, J, Mehta, S, Jenkinson, H, Frobisher, C, Reulen, RC and Hawkins, MM. 2015. Aspects of mental health dysfunction among survivors of childhood cancer. *British Journal of Cancer.* **113**(7), pp.1121-1132. https://dx.doi.org/10.1038/bjc.2015.310

Finch, W and French, BF. 2014. Multilevel latent class analysis: parametric and nonparametric models. *Journal of Experimental Education.* **82**(3), pp.307-333. http://dx.doi.org/10.1080/00220973.2013.813361

Fuller, WA. 1987. *Measurement error models.* New York: Wiley.

Geiser, C, Litson, K, Bishop, J, Keller, BT, Burns, G, Servera, M and Shiffman, S. 2015. Analyzing person, situation and person x situation interaction effects: latent state-trait models for the combination of random and fixed situations. *Psychological Methods.* **20**(2), pp.165-192. http://dx.doi.org/10.1037/met0000026

Gibson, WA. 1955. An extension of Anderson's solution for the latent structure equations. *Psychometrika.* **20**(1), pp.69-73. http://dx.doi.org/10.1007/BF02288961

Gibson, WA. 1959. Three multivariate models: factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika.* **24**(3), pp.229-252. http://dx.doi.org/10.1007/BF02289845

Gibson, WA. 1962. Extending latent class solutions to other variables. *Psychometrika.* **27**(1), pp.73-81. http://dx.doi.org/10.1007/BF02289666

Gill, MD, Bramble, MG, Hull, MA, Mills, SJ, Morris, E, Bradburn, DM, Bury, Y, Parker, CE, Lee, TJ and Rees, CJ. 2014. Screen-detected colorectal cancers are associated with an improved outcome compared with stage-matched interval cancers. *British Journal of Cancer.* **111**(11), pp.2076-2081. http://dx.doi.org/10.1038/bjc.2014.498

Gilthorpe, MS, Frydenberg, M, Cheng, Y and Baelum, V. 2009. Modelling count data with excessive zeros: the need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. *Statistics in Medicine.* **28**(28), pp.3539-3553. http://dx.doi.org/10.1002/sim.3699

Gilthorpe, MS, Harrison, WJ, Downing, A, Forman, D and West, RM. 2011. Multilevel latent class casemix modelling: a novel approach to accommodate patient casemix. *BioMed Central Health Services Research.* **11**(53). http://dx.doi.org/10.1186/1472-6963-11-53

Goldstein, H. 2011. Chapter 13: multiple membership models. In: Balding, DJ, Cressie, NAC, Fitzmaurice, GM, Goldstein, H, Johnstone, IM, Molenberghs, G, Scott, DW, Smith, AFM, Tsay, RS and Weisbery, S eds. *Multilevel statistical models.* 4th ed. Chichester: Wiley, pp.255-266.

Gomes, M, Gutacker, N, Bojke, C and Street, A. 2016. Addressing missing data in Patient-Reported Outcome Measures (PROMS): implications for the use of PROMS for comparing provider performance. *Health Economics.* **25**(5), pp.515-528. https://dx.doi.org/10.1002/hec.3173

Goodman, LA. 1974a. The analysis of systems of qualitative variables when some of the variables are unobservable. Part I-A modified latent structure approach. *American Journal of Sociology.* **79**(5), pp.1179-1259. http://dx.doi.org/10.1086/225676

Goodman, LA. 1974b. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika.* **61**(2), pp.215-231. http://dx.doi.org/10.1093/biomet/61.2.215

Goodman, LA. 1979. On the estimation of parameters in latent structure analysis. *Psychometrika.* **44**(1), pp.123-128. http://dx.doi.org/10.1007/BF02293792

Green, BF, Jr. 1951. A general solution for the latent class model of latent structure analysis. *Psychometrika.* **16**(2), pp.151-166. http://dx.doi.org/10.1007/BF02289112

Greenland, S, Pearl, J and Robins, JM. 1999. Causal diagrams for epidemiologic research. *Epidemiology.* **10**(1), pp.37-48.

Greenwood, DC. 2012. Chapter 3: measurement errors in epidemiology. In: Greenwood, DC and Tu, Y-K eds. *Modern methods for epidemiology.* London: Springer, pp.33-55.

Greenwood, DC, Gilthorpe, MS and Cade, JE. 2006. The impact of imprecisely measured covariates on estimating gene-environment interactions. *BioMed Central Medical Research Methodology.* **6**(21). http://dx.doi.org/10.1186/1471-2288-6-21

Hagenaars, JA. 1990. *Categorical longitudinal data: log-linear panel, trend, and cohort analysis.* Newbury Park: Sage.

Hanks, P, McLeod, WT and Urdang, L. 1986. *Collins dictionary of the English language.* 2nd revised ed. Collins.

Harrison, WJ, Gilthorpe, MS, Downing, A and Baxter, PD. 2013. Multilevel latent class modelling of colorectal cancer survival status at three years and socioeconomic background whilst incorporating stage of disease. *International Journal of Statistics and Probability.* **2**(3), pp.85-95. http://dx.doi.org/10.5539/ijsp.v2n3p85

Harrison, WJ, West, RM, Downing, A and Gilthorpe, M. 2012. Chapter 7: multilevel latent class modelling. In: Greenwood, DC and Tu, Y-K eds. *Modern methods for epidemiology.* London: Springer, pp.117-140.

Haynes, K, Forde, KA, Schinnar, R, Wong, P, Strom, BL and Lewis, JD. 2009. Cancer incidence in the Health Improvement Network. *Pharmacoepidemiology and Drug Safety.* **18**(8), pp.730-736. https://dx.doi.org/10.1002/pds.1774

He, Y, Landrum, MB and Zaslavsky, AM. 2014. Combining information from two data sources with misreporting and incompleteness to assess hospice-use among cancer patients: a multiple imputation approach. *Statistics in Medicine.* **33**(21), pp.3710-3724. http://dx.doi.org/10.1002/sim.6173

Hendifar, A, Yang, D, Lenz, F, Lurje, G, Pohl, A, Lenz, C, Ning, Y, Zhang, W and Lenz, HJ. 2009. Gender disparities in metastatic colorectal cancer survival. *Clinical Cancer Research.* **15**(20), pp.6391-6397. http://dx.doi.org/10.1158/1078-0432.CCR-09-0877

Henry, KL and Muthén, B. 2010. Multilevel latent class analysis: an application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling.* **17**(2), pp.193-215. http://dx.doi.org/10.1080/10705511003659342

Hernández-Díaz, S, Schisterman, EF and Hernán, MA. 2006. The birth weight "paradox" uncovered? *American Journal of Epidemiology.* **164**(11), pp.1115-1120. http://dx.doi.org/10.1093/aje/kwj275

Hill, C. 1995. *Cancer incidence in five continents.* International Agency for Research on Cancer (IARC) Scientific Publication no. 120. Lyon: IARC.

Hox, JJ. 2013. Multilevel regression and multilevel Structural Equation Modeling. In: Little, TD ed. *The Oxford handbook of quantitative methods in psychology: vol. 2: statistical analysis.* New York: Oxford University Press.

Hox, JJ and Bechger, TM. 1998. An introduction to Structural Equation Modeling. *Family Science Review.* **11**(4), pp.354-373. Available from: http://www.joophox.net/publist/semfamre.pdf

Hwang, MJ, Evans, T, Lawrence, G and Karandikar, S. 2014. Impact of bowel cancer screening on the management of colorectal cancer. *Colorectal Disease.* **16**(6), pp.450-458. http://dx.doi.org/10.1111/codi.12562

Ihedioha, U, Gravante, G, Lloyd, G, Sangal, S, Sorge, R, Singh, B and Chaudhri, S. 2013. Curative colorectal resections in patients aged 80 years and older: clinical characteristics, morbidity, mortality and risk factors. *International Journal of Colorectal Disease.* **28**(7), pp.941-947. http://dx.doi.org/10.1007/s00384-012-1626-0

International Agency for Research on Cancer (IARC). 2010. *Glossary of terms.* [Online]. [Accessed 23 August 2016]. Available from: http://www-dep.iarc.fr/WHOdb/glossary.htm

Jeffreys, M, Redaniel, MT and Martin, RM. 2015. The effect of pre-diagnostic vitamin D supplementation on cancer survival in women: a cohort study within the UK Clinical Practice Research Datalink. *BioMed Central Cancer.* **15**, p670. http://dx.doi.org/10.1186/s12885-015-1684-0

Jones, AP, Haynes, R, Sauerzapf, V, Crawford, SM, Zhao, H and Forman, D. 2008. Travel time to hospital and treatment for breast, colon, rectum,

lung, ovary and prostate cancer. *European Journal of Cancer.* **44**(7), pp.992-999. http://dx.doi.org/10.1016/j.ejca.2008.02.001

Kalmijn, M and Vermunt, JK. 2007. Homogeneity of social networks by age and marital status: a multilevel analysis of ego-centered networks. *Social Networks.* **29**(1), pp.25-43. http://dx.doi.org/10.1016/j.socnet.2005.11.008

Kaplan, D. 2009. *Structural Equation Modeling. Foundations and extensions.* 2nd ed. Advanced Quantitative Techniques in the Social Sciences. Volume: 10. Thousand Oaks, CA: Sage.

Kirkwood, BR and Sterne, JA. 2003. *Essential medical statistics.* 2nd ed. Oxford, UK: Blackwell.

Knaus, WA, Zimmerman, JE, Wagner, DP, Draper, EA and Lawrence, DE. 1981. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Critical Care Medicine.* **9**(8), pp.591-597. http://dx.doi.org/10.1097/00003246-198108000-00008

Koch, T, Schultze, M, Jeon, M, Nussbeck, FW, Praetorius, A-K and Eid, M. 2016. A cross-classified CFA-MTMM model for structurally different and nonindependent interchangeable methods. *Multivariate Behavioral Research.* **51**(1), pp.67-85. http://dx.doi.org/10.1080/00273171.2015.1101367

Koo, JH, Jalaludin, B, Wong, SK, Kneebone, A, Connor, SJ and Leong, RW. 2008. Improved survival in young women with colorectal cancer. *American Journal of Gastroenterology.* **103**(6), pp.1488-1495. http://dx.doi.org/10.1111/j.1572-0241.2007.01779.x

Lazarsfeld, PF. 1950. Chapter 10: the logical and mathematical foundations of latent structure analysis. In: Stouffer, SA ed. *Studies in social psychology in World War II. Volume 4. Measurement and prediction.* Princeton NJ: Princeton University Press, pp.362-412.

Lazarsfeld, PF. 1959. Latent structure analysis. In: Koch, S ed. *Psychology: a study of a science. Study I. Conceptual and systematic. Volume 3. Formulations of the person and the social context.* New York: McGraw-Hill, pp.476-543.

Lazarsfeld, PF and Henry, NW. 1968. *Latent structure analysis.* Boston: Houghton Mifflin.

Lee, Y-C, Lee, Y-L, Chuang, J-P and Lee, J-C. 2013. Differences in survival between colon and rectal cancer from SEER data. *PLoS ONE [Electronic Resource].* **8**(11), e78709. http://dx.doi.org/10.1371/journal.pone.0078709

Lejeune, C, Sassi, F, Ellis, L, Godward, S, Mak, V, Day, M and Rachet, B. 2010. Socio-economic disparities in access to treatment and their impact on colorectal cancer survival. *International Journal of Epidemiology.* **39**(3), pp.710-717. http://dx.doi.org/10.1093/ije/dyq048

Leyland, AH and Goldstein, H. 2001. *Multilevel modelling of health statistics.* Chichester, UK: Wiley.

Leyland, AH and Groenewegen, PP. 2003. Multilevel modelling and public health policy. *Scandinavian Journal of Public Health.* **31**(4), pp.267-274. http://dx.doi.org/10.1080/14034940210165028

Logan, RF, Patnick, J, Nickerson, C, Coleman, L, Rutter, MD, von Wagner, C and English Bowel Cancer Screening Evaluation Committee. 2012. Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests. *Gut.* **61**(10), pp.1439-1446.

Lord, FM. 1952. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika.* **17**(2), pp.181-194. http://dx.doi.org/10.1007/BF02288781

Lord, FM and Novick, MR. 1968. *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Macdonald, L, Cummins, S and Macintyre, S. 2007. Neighbourhood fast food environment and area deprivation - substitution or concentration? *Appetite.* **49**(1), pp.251-254. http://dx.doi.org/10.1016/j.appet.2006.11.004

Magidson, J and Vermunt, JK. 2004. Chapter 10: latent class models. In: Kaplan, D ed. *The Sage handbook of quantitative methodology for the social sciences.* Thousand Oaks, CA: Sage, pp.175-198.

Mansouri, D, McMillan, DC, Grant, Y, Crighton, EM and Horgan, PG. 2013. The impact of age, sex and socioeconomic deprivation on outcomes in a

colorectal cancer screening programme. *PLoS ONE [Electronic Resource].* **8**(6), pe66063. http://dx.doi.org/10.1371/journal.pone.0066063

Maringe, C, Li, R, Mangtani, P, Coleman, MP and Rachet, B. 2015. Cancer survival differences between South Asians and non-South Asians of England in 1986-2004, accounting for age at diagnosis and deprivation. *British Journal of Cancer.* **113**(1), pp.173-181. http://dx.doi.org/10.1038/bjc.2015.182

Maringe, C, Walters, S, Rachet, B, Butler, J, Fields, T, Finan, P, Maxwell, R, Nedrebo, B, Pahlman, L, Sjovall, A, Spigelman, A, Engholm, G, Gavin, A, Gjerstorff, ML, Hatcher, J, Johannesen, TB, Morris, E, McGahan, CE, Tracey, E, Turner, D, Richards, MA, Coleman, MP and Icbp Module 1 Working Group. 2013. Stage at diagnosis and colorectal cancer survival in six high-income countries: a population-based study of patients diagnosed during 2000-2007. *Acta Oncologica.* **52**(5), pp.919-932. http://dx.doi.org/10.3109/0284186X.2013.764008

Marshall, CE and Spiegelhalter, DJ. 1998. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *British Medical Journal.* **316**(7146), pp.1701-1705. http://dx.doi.org/10.1136/bmj.316.7146.1701

Martins, T, Hamilton, W and Ukoumunne, OC. 2013. Ethnic inequalities in time to diagnosis of cancer: a systematic review. *BioMed Central Family Practice.* **14**, p197. http://dx.doi.org/10.1186/1471-2296-14-197

Matsueda, RL. 2012. Key advances in the history of Structural Equation Modeling. In: Hoyle, R ed. *Handbook of Structural Equation Modeling.* New York, NY: Guilford Press.

McArdle, CS, McMillan, DC and Hole, DJ. 2003. Male gender adversely affects survival following surgery for colorectal cancer. *British Journal of Surgery.* **90**(6), pp.711-715. http://dx.doi.org/10.1002/bjs.4098

McClements, PL, Madurasinghe, V, Thomson, CS, Fraser, CG, Carey, FA, Steele, RJ, Lawrence, G and Brewster, DH. 2012. Impact of the UK colorectal cancer screening pilot studies on incidence, stage distribution and

mortality trends. *Cancer Epidemiology.* **36**(4), pp.e232-242. http://dx.doi.org/10.1016/j.canep.2012.02.006

McCowan, C, Munro, AJ, Donnan, PT and Steele, RJ. 2013. Use of aspirin post-diagnosis in a cohort of patients with colorectal cancer and its association with all-cause and colorectal cancer specific mortality. *European Journal of Cancer.* **49**(5), pp.1049-1057. http://dx.doi.org/10.1016/j.ejca.2012.10.024

McMillan, DC and McArdle, CS. 2009. The impact of young age on cancer-specific and non-cancer-related survival after surgery for colorectal cancer: 10-year follow-up. *British Journal of Cancer.* **101**(4), pp.557-560. http://dx.doi.org/10.1038/sj.bjc.6605222

McPhail, S, Elliss-Brookes, L, Shelton, J, Ives, A, Greenslade, M, Vernon, S, Morris, EJ and Richards, M. 2013. Emergency presentation of cancer and short-term mortality. *British Journal of Cancer.* **109**(8), pp.2027-2034. http://dx.doi.org/10.1038/bjc.2013.569

McPhail, S, Johnson, S, Greenberg, D, Peake, M and Rous, B. 2015. Stage at diagnosis and early mortality from cancer in England. *British Journal of Cancer.* **112 Suppl 1**, pp.S108-115. http://dx.doi.org/10.1038/bjc.2015.49

Meng, X-L. 1994. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science.* **9**(4), pp.538-558.

Morris, EJ, Maughan, NJ, Forman, D and Quirke, P. 2007a. Identifying stage III colorectal cancer patients: the influence of the patient, surgeon and pathologist. *Journal of Clinical Oncology.* **25**(18), pp.2573-2579. http://dx.doi.org/10.1200/JCO.2007.11.0445

Morris, EJ, Maughan, NJ, Forman, D and Quirke, P. 2007b. Who to treat with adjuvant therapy in Dukes B/Stage II colorectal cancer? The need for high quality pathology. *Gut.* **56**(10), pp.1419-1425. http://dx.doi.org/10.1136/gut.2006.116830

Morris, EJ, Taylor, EF, Thomas, JD, Quirke, P, Finan, PJ, Coleman, MP, Rachet, B and Forman, D. 2011. Thirty-day postoperative mortality after colorectal cancer surgery in England. *Gut.* **60**(6), pp.806-813. http://dx.doi.org/10.1136/gut.2010.232181

Morris, EJ, Whitehouse, LE, Farrell, T, Nickerson, C, Thomas, JD, Quirke, P, Rutter, MD, Rees, C, Finan, PJ, Wilkinson, JR and Patnick, J. 2012. A retrospective observational study examining the characteristics and outcomes of tumours diagnosed within and without of the English NHS Bowel Cancer Screening Programme. *British Journal of Cancer.* **107**(5), pp.757-764. http://dx.doi.org/10.1038/bjc.2012.331

Morrison, DS, Batty, GD, Kivimaki, M, Davey Smith, G, Marmot, M and Shipley, M. 2011. Risk factors for colonic and rectal cancer mortality: evidence from 40 years' follow-up in the Whitehall I study. *Journal of Epidemiology and Community Health.* **65**(11), pp.1053-1058. http://dx.doi.org/10.1136/jech.2010.127555

Morselli, D and Passini, S. 2012. Disobedience and support for democracy: Evidences from the World Values Survey. *The Social Science Journal.* **49**(3), pp.284-294. http://dx.doi.org/10.1016/j.soscij.2012.03.005

Mumford, EA, Liu, W, Hair, EC and Yu, TC. 2013. Concurrent trajectories of BMI and mental health patterns in emerging adulthood. *Social Science and Medicine.* **98**, pp.1-7. http://dx.doi.org/10.1016/j.socscimed.2013.08.036

Muthén, BO. 1984. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika.* **49**(1), pp.115-132. http://dx.doi.org/10.1007/BF02294210

Muthén, BO. 2002. Beyond SEM: General latent variable modeling. *Behaviormetrika.* **29**(1), pp.81-117. http://doi.org/10.2333/bhmk.29.81

Muthén, LK and Muthén, BO. 1998-2015. *Mplus user's guide. Seventh edition*. Los Angeles, CA: Muthén and Muthén.

Mutz, R and Daniel, HD. 2013. University and student segmentation: multilevel latent-class analysis of students' attitudes towards research methods and statistics. *British Journal of Educational Psychology.* **83**(Pt 2), pp.280-304. http://dx.doi.org/10.1111/j.2044-8279.2011.02062.x

Nachiappan, S, Askari, A, Mamidanna, R, Munasinghe, A, Currie, A, Stebbing, J and Faiz, O. 2015. The impact of adjuvant chemotherapy timing on overall survival following colorectal cancer resection. *European Journal of*

Surgical Oncology. **41**(12), pp.1636-1644. http://dx.doi.org/10.1016/j.ejso.2015.09.009

National Cancer Institute. 2016. *Age standards for survival.* [Online]. [Accessed 23 August 2016]. Available from: http://seer.cancer.gov/stdpopulations/survival.html

National Cancer Intelligence Network. 2010. *Cancer Outcomes and Services Dataset (COSD).* [Online]. [Accessed 6 February 2017]. Available from: http://www.ncin.org.uk/collecting_and_using_data/data_collection/cosd

National Institute for Clinical Excellence. 2004. *Guidance on cancer services: improving outcomes in colorectal cancers - manual update.* London: National Institute for Clinical Excellence.

NHS Digital. 2017. *Data sets.* [Online]. [Accessed 6 February 2017]. Available from: http://content.digital.nhs.uk/datasets

National Institute for Clinical Excellence. 2017a. *NICE pathways: colorectal cancer overview.* [Online]. [Accessed 27 January 2017]. Available from: https://pathways.nice.org.uk/pathways/colorectal-cancer

National Institute for Clinical Excellence. 2017b. *NICE pathways: suspected cancer recognition and referral overview.* [Online]. [Accessed 27 January 2017]. Available from: https://pathways.nice.org.uk/pathways/suspected-cancer-recognition-and-referral

Nicholl, J. 2007. Case-mix adjustment in non-randomised observational evaluations: the constant risk fallacy. *Journal of Epidemiology and Community Health.* **61**(11), pp.1010-1013. http://dx.doi.org/10.1136/jech.2007.061747

Nicholson, GA, Finlay, IG, Diament, RH, Molloy, RG, Horgan, PG and Morrison, DS. 2011. Mechanical bowel preparation does not influence outcomes following colonic cancer resection. *British Journal of Surgery.* **98**(6), pp.866-871. http://dx.doi.org/10.1002/bjs.7454

Noble, M, Wright, G, Dibben, C, Smith, GAN, McLennan, D, Anttila, C, Barnes, H, Mokhtar, C, Noble, S, Avenell, D, Gardner, J, Covizzi, I and Lloyd, M. 2004. *The English indices of deprivation 2004 (revised).* London: Office of the Deputy Prime Minister.

Norat, T, Scoccianti, C, Boutron-Ruault, MC, Anderson, A, Berrino, F, Cecchini, M, Espina, C, Key, T, Leitzmann, M, Powers, H, Wiseman, M and Romieu, I. 2015. European code against cancer 4th edition: diet and cancer. *Cancer Epidemiology.* **39 Suppl 1**, pp.S56-66. http://dx.doi.org/10.1016/j.canep.2014.12.016

Normand, S-LT, Sykora, K, Li, P, Mamdani, M, Rochon, PA and Anderson, GM. 2005. Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. *British Medical Journal.* **330**(7498), pp.1021-1023. http://dx.doi.org/10.1136/bmj.330.7498.1021

Nur, U, Lyratzopoulos, G, Rachet, B and Coleman, MP. 2015. The impact of age at diagnosis on socioeconomic inequalities in adult cancer survival in England. *Cancer Epidemiology.* **39**(4), pp.641-649. http://dx.doi.org/10.1016/j.canep.2015.05.006

Nur, U, Rachet, B, Parmar, MK, Sydes, MR, Cooper, N, Lepage, C, Northover, JM, James, R, Coleman, MP and collaborators, A. 2008. No socioeconomic inequalities in colorectal cancer survival within a randomised clinical trial. *British Journal of Cancer.* **99**(11), pp.1923-1928. http://dx.doi.org/10.1038/sj.bjc.6604743

NYCRIS (Northern and Yorkshire Cancer Registry and Information Service). 2007. *Cancer in the 21st century: NYCRIS statistical report 2000-2004.* Leeds, UK: NYCRIS. Available from: http://www.nycris.nhs.uk/uploads/doc55_1_nycris_stats_report_06.pdf.

Office for National Statistics. 2016a. *Cancer registration statistics, England.* Series MB1. Newport, UK: Office for National Statistics. Available from: http://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/qmis/cancerregistrationstatisticsqmi.

Office for National Statistics. 2016b. *Population estimates analysis tool.* [Online]. [Accessed 30 August 2016]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesanalysistool

Oliphant, R, Mansouri, D, Nicholson, GA, McMillan, DC, Horgan, PG, Morrison, DS and West of Scotland Colorectal Cancer Managed Clinical

Network. 2014a. Emergency presentation of node-negative colorectal cancer treated with curative surgery is associated with poorer short and longer-term survival. *International Journal of Colorectal Disease.* **29**(5), pp.591-598. http://dx.doi.org/10.1007/s00384-014-1847-5

Oliphant, R, Nicholson, GA, Horgan, PG, McMillan, DC, Morrison, DS and West of Scotland Colorectal Cancer Managed Clinical Network. 2014b. The impact of surgical specialisation on survival following elective colon cancer surgery. *International Journal of Colorectal Disease.* **29**(9), pp.1143-1150. http://dx.doi.org/10.1007/s00384-014-1965-0

Oliphant, R, Nicholson, GA, Horgan, PG, Molloy, RG, McMillan, DC, Morrison, DS and West of Scotland Colorectal Cancer Managed Clinical Network. 2013a. Contribution of surgical specialization to improved colorectal cancer survival. *British Journal of Surgery.* **100**(10), pp.1388-1395. http://dx.doi.org/10.1002/bjs.9227

Oliphant, R, Nicholson, GA, Horgan, PG, Molloy, RG, McMillan, DC, Morrison, DS and West of Scotland Colorectal Cancer Managed Clinical Network. 2013b. Deprivation and colorectal cancer surgery: longer-term survival inequalities are due to differential postoperative mortality between socioeconomic groups. *Annals of Surgical Oncology.* **20**(7), pp.2132-2139. http://dx.doi.org/10.1245/s10434-013-2959-9

Paterson, HM, Mander, BJ, Muir, P, Phillips, HA and Wild, SH. 2014. Deprivation and access to treatment for colorectal cancer in Southeast Scotland 2003-2009. *Colorectal Disease.* **16**(2), pp.O51-57. http://dx.doi.org/10.1111/codi.12442

Pearl, J. 2000. *Causality: models, reasoning and inference.* Cambridge, UK: Cambridge University Press.

Pearl, J. 2009. Remarks on the method of propensity score. *Statistics in Medicine.* **28**(9), pp.1415-1416; author reply pp.1420-1413. http://dx.doi.org/10.1002/sim.3521

Pearl, J. 2011. Invited commentary: understanding bias amplification. *American Journal of Epidemiology.* **174**(11), pp.1223-1227; discussion pp. 1228-1229. http://dx.doi.org/10.1093/aje/kwr352

Pearl, J, Glymour, M and Jewell, NP. 2016. *Causal inference in statistics: a primer.* Chichester, UK: Wiley.

Pirani, E. 2013. Evaluating contemporary social exclusion in Europe: a hierarchical latent class approach. *Quality and Quantity: International Journal of Methodology.* **47**(2), pp.923-941. http://dx.doi.org/10.1007/s11135-011-9574-2

Preacher, KJ, Zhang, Z and Zyphur, MJ. 2016. Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods.* **21**(2), pp.189-205. http://dx.doi.org/10.1037/met0000052

Quirke, P and Morris, E. 2007. Reporting colorectal cancer. *Histopathology.* **50**(1), pp.103-112. http://dx.doi.org/10.1111/j.1365-2559.2006.02543.x

R Development Core Team. 2010. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from: http://www.r-project.org/

Raleigh, VS, Frosini, F, Sizmur, S and Graham, C. 2012. Do some trusts deliver a consistently better experience for patients? An analysis of patient experience across acute care surveys in English NHS trusts. *British Medical Journal Quality and Safety.* **21**(5), pp.381-390. https://dx.doi.org/10.1136/bmjqs-2011-000588

Redaniel, MT, Martin, RM, Blazeby, JM, Wade, J and Jeffreys, M. 2014. The association of time between diagnosis and major resection with poorer colorectal cancer survival: a retrospective cohort study. *BioMed Central Cancer.* **14**, p642. http://dx.doi.org/10.1186/1471-2407-14-642

Rees, CJ and Bevan, R. 2013. The National Health Service Bowel Cancer Screening Program: the early years. *Expert review of gastroenterology and hepatology.* **7**(5), pp.421-437. http://dx.doi.org/10.1586/17474124.2013.811045

Reeves, GK, Pirie, K, Beral, V, Green, J, Spencer, E, Bull, D and Million Women Study Collaboration. 2007. Cancer incidence and mortality in relation to body mass index in the Million Women Study: cohort study. *British*

Medical Journal. **335**(7630), p1134.
http://dx.doi.org/10.1136/bmj.39367.495995.AE

Rindskopf, D. 2006. Heavy alcohol use in the "fighting back" survey sample: separating individual and community level influences using multilevel Latent Class Analysis. *Journal of Drug Issues.* **36**(2), pp.441-462.
http://dx.doi.org/10.1177/002204260603600210

Robinson, WS. 1950. Ecological correlations and the behaviour of individuals. *American Sociological Review.* **15**(3), pp.351-357.

Roden, DM and George Jr, AL. 2002. The genetic basis of variability in drug responses. *Nature Reviews Drug Discovery.* **1**, pp.37-44.
http://dx.doi.org/10.1038/nrd705

Rosenbaum, PR and Rubin, DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika.* **70**(1), pp.41-55.
http://dx.doi.org/10.1093/biomet/70.1.41

Rosenbaum, PR and Rubin, DB. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association.* **79**(387), pp.516-524.
http://dx.doi.org/10.2307/2288398

Rothman, K, Greenland, S and Lash, T. 1986. *Modern epidemiology.* Boston: Little, Brown and Co.

Rubin, DB. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association.* **74**(366), pp.318-324.
http://dx.doi.org/10.2307/2286330

Sanfelix-Gimeno, G, Rodriguez-Bernal, CL, Hurtado, I, Baixauli-Perez, C, Librero, J and Peiro, S. 2015. Adherence to oral anticoagulants in patients with atrial fibrillation-a population-based retrospective cohort study linking health information systems in the Valencia region, Spain: a study protocol. *British Medical Journal Open.* **5**(10), e007613.
http://dx.doi.org/10.1136/bmjopen-2015-007613

Schneider, C, Bevis, PM, Durdey, P, Thomas, MG, Sylvester, PA and Longman, RJ. 2013. The association between referral source and outcome

in patients with colorectal cancer. *Surgeon Journal of the Royal Colleges of Surgeons of Edinburgh and Ireland.* **11**(3), pp.141-146. http://dx.doi.org/10.1016/j.surge.2012.10.004

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics.* **6**(2), pp.461-464. http://dx.doi.org/10.1214/aos/1176344136

Seaman, SR and White, IR. 2013. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research.* **22**(3), pp.278-295. http://dx.doi.org/10.1177/0962280210395740

Sheppard, JP, Stevens, R, Gill, P, Martin, U, Godwin, M, Hanley, J, Heneghan, C, Hobbs, FD, Mant, J, McKinstry, B, Myers, M, Nunan, D, Ward, A, Williams, B and McManus, RJ. 2016. Predicting Out-of-Office Blood Pressure in the clinic (PROOF-BP): derivation and validation of a tool to improve the accuracy of blood pressure measurement in clinical practice. *Hypertension.* **67**(5), pp.941-950. https://dx.doi.org/10.1161/HYPERTENSIONAHA.115.07108

Sheridan, J, Walsh, P, Kevans, D, Cooney, T, O'Hanlon, S, Nolan, B, White, A, McDermott, E, Sheahan, K, O'Shea, D, Hyland, J, O'Donoghue, D, O'Sullivan, J, Mulcahy, H and Doherty, G. 2014. Determinants of short- and long-term survival from colorectal cancer in very elderly patients. *Journal of Geriatric Oncology.* **5**(4), pp.376-383. http://dx.doi.org/10.1016/j.jgo.2014.04.005

Simmonds, SJ, Syddall, HE, Walsh, B, Evandrou, M, Dennison, EM, Cooper, C and Aihie Sayer, A. 2014. Understanding NHS hospital admissions in England: linkage of Hospital Episode Statistics to the Hertfordshire Cohort Study. *Age and Ageing.* **43**(5), pp.653-660. https://dx.doi.org/10.1093/ageing/afu020

Skrondal, A and Rabe-Hesketh, S. 2004. *Generalized latent variable modeling: multilevel, longitudinal, and Structural Equation Models.* Interdisciplinary Statistics Series. Boca Raton, FL: Chapman and Hall/CRC.

Smith, JD, Van Ryzin, MJ, Fowler, JC and Handler, L. 2014. Predicting response to intensive multimodal inpatient treatment: a comparison of

single- and multiple-class growth modeling approaches. *Journal of Personality Assessment.* **96**(3), pp.306-315. http://dx.doi.org/10.1080/00223891.2013.834439

Smith, JJ, Tilney, HS, Heriot, AG, Darzi, AW, Forbes, H, Thompson, MR, Stamatakis, JD, Tekkis, PP, Association of Coloproctology of Great, B and Ireland. 2006. Social deprivation and outcomes in colorectal cancer. *British Journal of Surgery.* **93**(9), pp.1123-1131.

Snijders, T and Bosker, R. 1999. Chapter 14: discrete dependent variables. *Multilevel analysis.* London: Sage, p.223.

Spearman, C. 1904. "General Intelligence", objectively determined and measured *American Journal of Epidemiology.* **15**(2), pp.201-292. http://dx.doi.org/10.2307/1412107

StataCorp. 2015. *Stata statistical software: release 14*. College Station, TX: StataCorp LP.

Sterne, JA, White, IR, Carlin, JB, Spratt, M, Royston, P, Kenward, MG, Wood, AM and Carpenter, JR. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal.* **338**, b2393. http://dx.doi.org/10.1136/bmj.b2393

Stigler, SM. 1999. *Statistics on the table: the history of statistical concepts and methods.* Cambridge, MA: Harvard University Press.

Textor, J, Hardt, J and Knüppel, S. 2011. DAGitty: a graphical tool for analyzing causal diagrams. *Epidemiology.* **5**(22), p745. Available from: http://dagitty.net/

Textor, J, van der Zander, B, Gilthorpe, MS, Liśkiewicz, M and Ellison, GTH. 2016. Robust causal inference using Directed Acyclic Graphs: the R package 'dagitty'. *[in press].*

The National Cancer Registration and Analysis Service. 2017. *National cancer registration for England.* [Online]. [Accessed 7 February 2017]. Available from: http://www.ncras.nhs.uk/

Thurstone, LL. 1947. *Multiple-factor analysis: a development and expansion of the vectors of mind.* Chicago, IL: University of Chicago Press.

Townsend, P, Beattie, A and Phillimore, P. 1987. *Health and deprivation: inequality and the north.* London, UK: Routledge.

Tu, Y-K, Tilling, K, Sterne, JA and Gilthorpe, MS. 2013. A critical evaluation of statistical approaches to examining the role of growth trajectories in the developmental origins of health and disease. *International Journal of Epidemiology.* **42**(5), pp.1327-1339. http://dx.doi.org/10.1093/ije/dyt157

Tu, Y-K, West, R, Ellison, GTH and Gilthorpe, MS. 2005. Why evidence for the fetal origins of adult disease might be a statistical artifact: the "reversal paradox" for the relation between birth weight and blood pressure in later life. *American Journal of Epidemiology.* **161**(1), pp.27-32. http://dx.doi.org/10.1093/aje/kwi002

University of Leeds. 2017. *Leeds Institute for Data Analytics.* [Online]. [Accessed 6 February 2017]. Available from: http://lida.leeds.ac.uk/

Van Horn, M, Fagan, AA, Jaki, T, Brown, EC, Hawkins, J, Arthur, MW, Abbott, RD and Catalano, RF. 2008. Using multilevel mixtures to evaluate intervention effects in group randomized trials. *Multivariate Behavioral Research.* **43**(2), pp.289-326. http://dx.doi.org/10.1080/00273170802034893

van Lettow, B, Vermunt, JK, de Vries, H, Burdorf, A and van Empelen, P. 2013. Clustering of drinker prototype characteristics: what characterizes the typical drinker? *British Journal of Psychology.* **104**(3), pp.382-399. http://dx.doi.org/10.1111/bjop.12000

Vanderweele, TJ. 2015. *Explanation in causal inference: methods for mediation and interaction.* New York, NY: Oxford University Press.

Varriale, R and Vermunt, JK. 2012. Multilevel mixture factor models. *Multivariate Behavioral Research.* **47**(2), pp.247-275. http://dx.doi.org/10.1080/00273171.2012.658337

Vermunt, JK. 1997. *Log-linear models for event histories.* Thousand Oaks, CA: Sage.

Vermunt, JK. 2003. Multilevel latent class models. *Sociological Methodology.* **33**(1), pp.213-239. http://dx.doi.org/10.1111/j.0081-1750.2003.t01-1-00131.x

Vermunt, JK. 2008a. Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research.* **17**(1), pp.33-51. http://dx.doi.org/10.1177/0962280207081238

Vermunt, JK. 2008b. Multilevel latent variable modeling: an application in education testing. *Austrian Journal of Statistics.* **37**(3&4), pp.285-299.

Vermunt, JK and Magidson, J. 2003. Latent class models for classification. *Computational Statistics and Data Analysis.* **41**(3-4), pp.531-537. http://dx.doi.org/10.1016/S0167-9473(02)00179-2

Vermunt, JK and Magidson, J. 2005. *Latent GOLD 4.0 user's guide*. Belmont, MA: Statistical Innovations Inc.

Vermunt, JK and Magidson, J. 2013. *LG-syntax user's guide: manual for Latent GOLD 5.0 syntax module*. Belmont, MA: Statistical Innovations Inc.

Vermunt, JK and Magidson, J. 2016. *Technical guide for Latent GOLD 5.1: basic, advanced and syntax.* Belmont, MA: Statistical Innovations Inc.

von Hippel, PT. 2009. How to impute interactions, squares, and other transformed variables. *Sociological Methodology.* **39**(1), pp.265-291. http://dx.doi.org/10.1111/j.1467-9531.2009.01215.x

von Wagner, C, Baio, G, Raine, R, Snowball, J, Morris, S, Atkin, W, Obichere, A, Handley, G, Logan, RF, Rainbow, S, Smith, S, Halloran, S and Wardle, J. 2011. Inequalities in participation in an organized national colorectal cancer screening programme: results from the first 2.6 million invitations in England. *International Journal of Epidemiology.* **40**(3), pp.712-718. http://dx.doi.org/10.1093/ije/dyr008

Walker, AJ, Grainge, MJ and Card, TR. 2012. Aspirin and other non-steroidal anti-inflammatory drug use and colorectal cancer survival: a cohort study. *British Journal of Cancer.* **107**(9), pp.1602-1607. http://dx.doi.org/10.1038/bjc.2012.427

Walker, JJ, Brewster, DH, Colhoun, HM, Fischbacher, CM, Lindsay, RS, Wild, SH and Scottish Diabetes Research Network Epidemiology Group. 2013. Cause-specific mortality in Scottish patients with colorectal cancer with and without type 2 diabetes (2000-2007). *Diabetologia.* **56**(7), pp.1531-1541. http://dx.doi.org/10.1007/s00125-013-2917-x

Wallis, KL, Foxhall, M, Warner, RM and Jaffe, W. 2015. Clinical coding inaccuracies in skin cancer surgery: the financial implications for a plastic surgery service. *Journal of Plastic, Reconstructive and Aesthetic Surgery.* **68**(4), pp.582-583. https://dx.doi.org/10.1016/j.bjps.2014.11.015

West, RM. 2012. Chapter 15: generalised additive models. In: Greenwood, DC and Tu, Y-K eds. *Modern methods for epidemiology.* London: Springer, pp.261-278.

White, IR, Royston, P and Wood, AM. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine.* **30**(4), pp.377-399. http://dx.doi.org/10.1002/sim.4067

Widdison, AL, Barnett, SW and Betambeau, N. 2011. The impact of age on outcome after surgery for colorectal adenocarcinoma. *Annals of the Royal College of Surgeons of England.* **93**(6), pp.445-450. http://dx.doi.org/10.1308/003588411X587154

Wiesel, J, Arbesfeld, B and Schechter, D. 2014. Comparison of the Microlife blood pressure monitor with the Omron blood pressure monitor for detecting atrial fibrillation. *American Journal of Cardiology.* **114**(7), pp.1046-1048. https://dx.doi.org/10.1016/j.amjcard.2014.07.016

Williams, J, Scarborough, P, Townsend, N, Matthews, A, Burgoine, T, Mumtaz, L and Rayner, M. 2015. Associations between food outlets around schools and BMI among primary students in England: a cross-classified multi-level analysis. [Erratum appears in PLoS One. 2016;11(1):e0147164; PMID: 26752415]. *PLoS ONE [Electronic Resource].* **10**(7), e0132930. https://dx.doi.org/10.1371/journal.pone.0132930

World Health Organisation. 2005. *The International statistical Classification of Diseases and related health problems (ICD-10): tenth revision.* 2nd ed. Geneva, Switzerland: World Health Organisation.

Yu, H-T and Park, J. 2014. Simultaneous decision on the number of latent clusters and classes for multilevel latent class models. *Multivariate Behavioral Research.* **49**(3), pp.232-244. http://dx.doi.org/10.1080/00273171.2014.900431

Zafar, A, Mak, T, Whinnie, S and Chapman, MA. 2012. The 2-week wait referral system does not improve 5-year colorectal cancer survival. *Colorectal Disease.* **14**(4), pp.e177-180. http://dx.doi.org/10.1111/j.1463-1318.2011.02826.x

Zgaga, L, Theodoratou, E, Farrington, SM, Din, FV, Ooi, LY, Glodzik, D, Johnston, S, Tenesa, A, Campbell, H and Dunlop, MG. 2014. Plasma vitamin D concentration influences survival outcome after a diagnosis of colorectal cancer. *Journal of Clinical Oncology.* **32**(23), pp.2430-2439. http://dx.doi.org/10.1200/JCO.2013.54.5947

Zhang, X, van der Lans, I and Dagevos, H. 2012. Impacts of fast food and the food retail environment on overweight and obesity in China: a multilevel latent class cluster approach. *Public Health Nutrition.* **15**(1), pp.88-96. http://dx.doi.org/10.1017/S1368980011002047

Appendix A
Literature search strategies for the review of comparable latent variable approaches

Searches performed 16 September 2016; for articles from 1996 to September Week 1 2016.

Index	Search Terms	Results (N)	
		Medline	PsycINFO
1	Multilevel OR multi-level	17,416	15,409
2	MeSH: Multilevel Analysis	1,041	-
3	1 OR 2	17,416	15,409
4	Latent variable OR latent class	3,869	4,884
5	Latent AND (mixture model*)	511	719
6	4 OR 5	4,095	5,161
7	3 AND 6	60	220
8	Limit 7 to (abstracts AND English language AND humans AND year=”2006-current” AND journals only)	52	122
	TOTAL	**174**	

Appendix B

Literature search strategy for the review of risk factors associated with survival (or mortality) from colorectal cancer

Search performed 23 September 2016; for articles from 1996 to September Week 2 2016. Medline only.

Index	Search Terms	Results (N)
1	MeSH: Colorectal Neoplasms	61,980
2	Cancer AND (colorectal OR colon OR rectum OR rectosigmoid OR bowel)	111,723
3	1 OR 2	124,273
4	Survival OR MeSH: survival	718,775
5	Mortality OR MeSH: mortality	421,963
6	4 OR 5	1,036,600
7	MeSH: (risk factors OR prognosis OR diagnosis-related groups)	817,794
8	Factor*	3,124,064
9	7 OR 8	3,276,853
10	3 AND 6 AND 9 (i.e. colorectal cancer + survival + risk factors)	17,853
11	MeSH: (socioeconomic factors OR poverty)	111,547
12	Deprivation OR poverty OR socioeconomic status OR socio-economic status OR SES OR socioeconomic background OR socio-economic background OR SEB	112,286
13	11 OR 12	182,657
14	Stage	415,920
15	3 AND 13 AND 14 (i.e. colorectal cancer + deprivation + stage)	343
16	**10 OR 15**	**18,018**

Continued on next page

Index	Search Terms	Results (N)
17	MeSH: (Great Britain OR United Kingdom OR England OR Wales OR Scotland OR Northern Ireland)	192,994
18	Great Britain OR United Kingdom OR England OR Wales OR Scotland OR Northern Ireland	227,356
19	17 OR 18	227,356
20	16 AND 19	385
21	**Limit 20 to (abstracts AND English language AND humans AND year="2006-current")**	**263**

Appendix C

Stata code for data simulation

This code is in two parts: code is defined in the first program, and run in a loop in the second to generate 100 similarly defined datasets. * or /* preceding text indicates a comment.

1. Define programs

```
*******************************************************************************
* SIMULATE TRUST MEMBERSHIP
* allocate patients to Trusts by blocks of a random normal variable (p)
*******************************************************************************

capture program drop hosmembership
program define hosmembership
	* generate p = random uniform distribution (by patient)
	gen p = uniform()

	* generate 19 hospitals, of differing sizes
	qui gen HosID = .
	local pstart = 0
	local bit = -0.04
	local pend = `pstart' + (1/19) + `bit'
	* loop to create HosID for values of p
	forvalues i = 1/19 {
		qui replace HosID = `i' if p>`pstart' & p<`pend'
		local bit = `bit'  + (0.08/18)
		local pstart = `pend'
		local pend = `pstart' + (1/19) + `bit'
	}

	* randomise hospitals & generate pT (for continuous outcome)
	qui sort HosID p
	qui by HosID: gen oldcode = HosID if _n==1
	sort oldcode HosID p
	qui gen flag=uniform() in 1/19
	sort flag in 1/19
	qui gen newcode = _n in 1/19
	qui gen pTcode=uniform() in 1/19
	sort HosID p
```

```
	* set HosID = new (randomised) HosID code
	by HosID: egen newHosID = max(newcode)
	* generate pT = random uniform distribution (by Trust)
	by HosID: egen pT = max(pTcode)

	* tidy dataset
	drop HosID oldcode flag newcode pTcode
	rename newHosID HosID
	move HosID p
end

*************************************************************
* GENERATE SEX, DEP & AGE VARIABLES
* uses trivariate covariance matrix
* deprivation SD=3.18, age SD=11.6
* draws from a normal distribution, converts sex to 0/1
*************************************************************

capture program drop demog
program define demog
	matrix A = (1,0,0 \ 0,3.18^2,0 \ 0,0,11.6^2)
	drawnorm sex dep age, cov(A)
	qui recode sex (min/0 = 0)(nonmiss = 1)
end

*****************************************************************************
* BINARY TRUST-LEVEL COVARIATE EFFECT (HGrp)
* equals 0.5 or -0.5 (centred on zero) with error SD=0.01
* sets local betas, calculates Trust linear predictor (hlp1)
* simulates binary outcome (dth_bin) and continuous outcome (oc_bin)
*****************************************************************************

capture program drop dth_bin
program define dth_bin
	* generate blank binary Trust-level covariate
	qui gen HGrp=.

	* calculate 19 binary Trust-level covariates (SD=0.01)
	forvalues i = 1/19 {
		local mult = rnormal(0.5,0.01) - rbinomial(1,0.5)
		qui replace HGrp = `mult' if HosID == `i'
	}

	* set local betas (described in section 5.2.2)
	local b0 = -0.0265
	local b1 = -0.1368
	local b2 = 0.0527
	local b3 = 0.0547
```

```
        local HBeta = -`b1'          /*Trust-level coefficient effect*/

        * calculate linear predictor
        gen hlp1 = `b0' + `b1'*sex + `b2'*dep + `b3'*age + `HBeta'*HGrp
        * calculate binary outcome
        gen dth_bin = rbinomial(1,invlogit(hlp1))

        * set error for continuous outcome (error variance = 0.225 (50%))
        gen error=(sqrt(0.225))*invnormal(uniform())
        * calculate continuous outcome
        gen oc_bin=hlp1+error

        * tidy dataset
        drop hlp1 error
end

*******************************************************************************
* CONTINUOUS TRUST-LEVEL COVARIATE EFFECT (HVal)
* ranges from -0.5 to +0.5 (centred on zero)
* sets local betas, calculates Trust linear predictor (hlp2)
* simulates binary outcome (dth_con) and continuous outcome (oc_con)
*******************************************************************************

capture program drop dth_con
program define dth_con
        * generate continuous Trust-level covariate (by Trust, uses pT)
        gen HVal = 0.5-pT

        * set local betas (described in section 5.2.2)
        local b0 = -0.0265
        local b1 = -0.1368
        local b2 = 0.0527
        local b3 = 0.0547
        local HBeta = `b2'           /*Trust-level coefficient effect*/

        * calculate linear predictor
        gen hlp2 = `b0' + `b1'*sex + `b2'*dep + `b3'*age + `HBeta'*HVal
        * calculate binary outcome
        gen dth_con = rbinomial(1,invlogit(hlp2))

        * set error for continuous outcome (error variance = 0.218 (50%))
        gen error=(sqrt(0.218))*invnormal(uniform())
        * calculate continuous outcome
        gen oc_con=hlp2+error

        * tidy dataset
        drop hlp2 error p pT
end
```

2. Run programs

```
*************************************************************************
* RUN PROGRAMS TO SIMULATE DATA FOR MLLC MODELLING
* stores 100 datasets per run
*************************************************************************

version 13.1
clear
set more off
* set working directory
cd "N:\...\2016 Trust level covars\Data simulation\Datasets\Run 1"

* set and record seeds (use different seed for multiple runs)
capture log close
capture log using "seeds.log", replace
set seed 1073741823
*set seed 484848484        /*not active for this run*/
*set seed 8493829          /*not active for this run*/

* loop to generate 100 datasets with same specification
forvalues i = 1/100 {
        * set up Stata with 24,640 observations (patients)
        di `i'
        clear
        qui set obs 24640
        gen patID = _n

        * simulate dataset based on defined programs
        hosmembership
        demog
        dth_bin
        dth_con

        * export to csv
        qui export delimited using "simulated`i'.csv"

        * increment seed
        set seed `c(seed)'
        di c(seed)
        }
capture log close
```