

# Multilevel Latent Class Modelling of Colorectal Cancer Survival Status at Three Years and Socioeconomic Background Whilst Incorporating Stage of Disease

Wendy J. Harrison<sup>1</sup>, Mark S. Gilthorpe<sup>1</sup>, Amy Downing<sup>2</sup> & Paul D. Baxter<sup>1</sup>

<sup>1</sup> Division of Biostatistics, Centre for Epidemiology & Biostatistics, University of Leeds, Leeds, UK

<sup>2</sup> Cancer Epidemiology Group, Centre for Epidemiology & Biostatistics, University of Leeds, Leeds, UK

Correspondence: Wendy J. Harrison, Division of Biostatistics, Centre for Epidemiology & Biostatistics, Level 8 Worsley Building, University of Leeds, Leeds LS2 9JT, UK. Tel: 44-113-343-4831. E-mail: w.harrison@leeds.ac.uk

Received: May 22, 2013 Accepted: June 9, 2013 Online Published: July 10, 2013

doi:10.5539/ijsp.v2n3p85

URL: <http://dx.doi.org/10.5539/ijsp.v2n3p85>

## Abstract

Previous studies investigating survival from colorectal cancer have typically considered potential confounders to include stage of disease. Stage however may lie on the causal path and statistical adjustment with stage as a confounder can then introduce bias known as the reversal paradox. Classification of stage may also be imprecise and incomplete. Modelling using Latent Class Analysis (LCA) may minimise bias by including covariates on the causal path as 'class predictors' and by accommodating uncertainty associated with confounder values explicitly via the latent class part of the model. We construct multilevel latent class models to allow for the multilevel structure of the data: patients nested within NHS Trusts. We use a dataset of patients in a large UK regional population diagnosed with colorectal cancer between 1998 and 2004. Death within three years is the outcome. The optimum number of latent classes at patient and Trust level is determined with reference to likelihood-based model-fit criteria. The three-patient five-Trust class multilevel LCA model was chosen. Patient classes were identified as good, reasonable or poor prognosis groups. The impact of stage differed across the patient classes. Socioeconomic background and older age were clearly associated with increased odds of death in all patient classes. Females had significantly decreased odds of death compared with males in the good prognosis class. The five Trust classes identified outlying Trusts, indicating that the standard multilevel model would not have been sufficient to model these data.

**Keywords:** multilevel, latent class, confounding, mediation

## 1. Introduction

### 1.1 Background

Many factors influence survival from colorectal cancer, including diagnosis or treatment centre (see Kee et al., (1999) on the influence of hospital and clinician workload; McArdle & Hole (2004) on volume and specialisation; and Borowski et al. (2010) for a volume-outcome analysis), stage at diagnosis (see Woodman, Gibbs, Scott, Haboubi, & Collins (2001) on differences in stage at presentation; and Ciccolallo et al. (2005) on the role of stage and surgery), and associated risk factors such as socioeconomic background, age at diagnosis and sex (see Morris et al. (2011) for thirty-day postoperative mortality; Downing et al. (2013) for early mortality; Smith et al. (2006) on the impact of social deprivation; Widdison, Barnett, & Betambeau (2011) on age; and Hendifar et al. (2009) on gender disparities). Findings exploring the potential impact of socioeconomic background (SEB) vary, with some studies showing an impact of poor SEB on decreased colorectal cancer survival (Downing et al., 2013; Morris et al., 2011), whilst others do not find this association (Nur et al., 2008; Smith et al., 2006). We investigate the relationship between survival status from colorectal cancer and SEB, while accounting for other factors that may affect this relationship using a novel statistical approach that can account for both genuine confounding and what is effectively mediation, often mistakenly treated as confounding but which may in fact bias the estimated impact of SEB when adjusted for in standard regression models.

For instance, previous studies may have included stage of disease at diagnosis as a potential confounder to the

relationship between known potential risk factors and survival from colorectal cancer. However, a higher level of deprivation may provoke late presentation, perhaps due to the refusal of a screening invitation (Whynes, Frew, Manghan, Scholefield, & Hardcastle, 2003), which may result in a more advanced stage at diagnosis (Ionescu, Carey, Tait, & Steele, 1998). SEB therefore causally precedes stage at diagnosis and consequently stage does not qualify as a genuine confounder; it is a mediating factor. Bias may be introduced by the statistical adjustment for mediators on the causal path (Kirkwood & Sterne, 2003), termed the reversal paradox (Stigler, 1999), which may be a serious problem in epidemiology (Hernández-Díaz, Schisterman, & Hernán, 2006; Tu, West, Ellison, & Gilthorpe, 2005). Figure 1 shows a theorised diagram of causality using a Directed Acyclic Graph (DAG) (Pearl, 2000) to determine which variables in our study are confounders, proxies for confounding, competing exposures, or mediating variables that lie on the causal pathway.

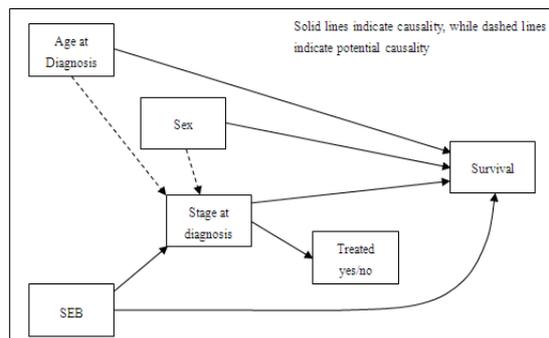


Figure 1. Directed Acyclic Graph (DAG) showing the inferred causal relationships amongst all available variables at the population level

In addition to the question of legitimate confounding, stage typically includes a large proportion of missing data (24.1% in the Northern and Yorkshire Cancer Registry and Information Service (NYCRIS) in 2008) (United Kingdom Association of Cancer Registries, 2008). Classification may also be imprecise as patients may be classified incorrectly due to the variable quality of pathology (Quirke & Morris, 2007) or be “understaged” (i.e. incorrectly assigned a earlier stage at diagnosis due to unidentified lymph node metastases) (Morris, Maughan, Forman, & Quirke, 2007). Statistical analyses using regression modelling may yield biased results where model covariates have measurement error (Greenwood, 2012) or missing values (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Fuller, 1987), and this bias is exacerbated when considering product interaction terms (Greenwood, Gilthorpe, & Cade, 2006). Models that incorporate staging data may therefore introduce bias due to the variable quality and completeness of pathology.

### 1.2 Modelling Approaches

Regression modelling (Normand et al., 2005) is often extended to a multilevel framework in order to incorporate differences across diagnosis or treatment centres (Leyland & Goldstein, 2001; Leyland & Groenewegen, 2003), such as NHS Trusts. This approach however assumes that a study sample is homogeneous at every level, the same model would be applied to all members of the sample, and the effects of covariates would thus be the same throughout. In a multilevel model, variation amongst the intercepts and slopes is assumed to be normally distributed and independent of the variation in the individual measurements. These assumptions may not be valid in observational health data; patients or Trusts are unlikely to be homogeneous where patients have not been randomly selected for inclusion and Trusts have not been randomly distributed geographically. As such, it may be inappropriate to apply one model to all individuals, and the effect of covariates therefore cannot be assumed to be the same throughout the sample.

Latent Class Analysis (LCA) could be used, and this also extends to a multilevel framework. Downing, Harrison, West, Forman and Gilthorpe (2010) incorporated stage at diagnosis as a “class predictor” in the LCA rather than as a covariate and found improvements in model fit using multilevel LCA (MLLCA) in comparison to single-level logistic regression modelling when studying risk factors related to breast cancer survival status. MLLCA is therefore proposed here to illustrate an original application in an area where its utility may be overlooked. It is important to consider alternative approaches to match the context of the data; we advocate an improved approach to analysis in cancer research data. We construct multilevel latent class models to identify subtypes of patients and

NHS Trusts, simultaneously, to model how patients may vary and how NHS Trusts may differ, based on survival status. We compare the MLLCA approach with standard multilevel models (MLMs) to examine improvements in model fit and model interpretation and hope to demonstrate the utility of the latent class approach.

## 2. Methods

### 2.1 The Colorectal Cancer Dataset

The Northern and Yorkshire Cancer Registry and Information Service (NYCRIS) database was used to identify cases of colorectal cancer (ICD-10 (World Health Organisation, 2005) codes C18, C19 and C20) diagnosed between 1998 and 2004, where the patient was resident in the Northern and Yorkshire regions. A description of the data extraction and exclusions is available in a previously published study (Gilthorpe, Harrison, Downing, Forman, & West, 2011). The outcome was whether or not the patient survived at three years following diagnosis, as this is clinically meaningful and facilitates ready comparison with other studies. Following exclusions 24640 records were available for analysis. Data include information on age at diagnosis, sex and SEB (using the Townsend Index (Townsend, Beattie, & Phillimore, 1987) recorded at the 2001 census, stage at diagnosis (using the Dukes classification (Dukes, 1949)), the ICD-10 diagnosis code for the tumour, its laterality (position in the body), and whether or not the patient was treated curatively. The diagnostic centre was defined as the NHS Trust where the latest staging took place; 19 Trusts were identified in the NYCRIS geographical area.

The colorectal cancer data are hierarchical since different groups of patients attend different diagnostic centres, dependent on factors such as their area of residence; patients are clustered within NHS Trusts and there will be variation at both patient and Trust level. A multilevel modelling framework would therefore seem appropriate. It is also important to account for stage at diagnosis, but this cannot be included as a covariate alongside SEB without the risk of introducing bias due to the reversal paradox.

### 2.2 Latent Class Analysis

LCA is also known as discrete latent variable modelling, or mixture modelling (Goodman, 1974; Magidson & Vermunt, 2004). In LCA, a number of latent classes, or subgroups, are identified, the optimum choice of which is selected by the researcher (usually informed by log-likelihood statistics). Units of analysis are assigned to a latent class based on similarities in their characteristics and latent classes are therefore homogeneous, with similar effects of each covariate on units in the same latent class, though 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.

LCA has the utility to model covariates as “class predictors”. This may be either in addition to or instead of their inclusion as standard covariates along with the main exposure under investigation. For confounders that are also potential effect modifiers (i.e. they exhibit an interaction with the main exposure), modelling these variables as class predictors yields an implicit interaction, since the outcome-exposure relationship may vary across latent classes. This averts the need to include an explicit confounder-exposure product term in the standard part of the regression model, which would otherwise exacerbate any bias introduced if the confounder is measured with error or has missing values. Modelling effect modification this way minimises bias; uncertainty associated with confounder values is explicitly accommodated via the latent class part of the model.

### 2.3 Confounding and Mediation

If an alleged confounder lies on the causal path between exposure and outcome, it is a mediator, and its statistical adjustment in the standard regression model introduces bias; it is then wise to discard the mediator as a model covariate. This does not preclude the mediator becoming a “class predictor”, though one then has to ensure there is no remaining implicit bias. Modelling a mediator as class predictor yields the potential for implicit interaction, as before, where the exposure-outcome relationship may vary across latent classes. The exposure may thus 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 outcome-exposure 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. Although the causal circularity is avoided, this 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 the mediator as a class predictor. We nevertheless explore the notion that variables which lie on the causal path between exposure and outcome may be considered as class predictors instead of being incorrectly adjusted for as

alleged confounders within the standard regression model.

#### 2.4 Multilevel Latent Class Analysis

MLLCA is an extension of LCA with latent classes determined at more than one level, where classes at the lower level are based on similarities in characteristics (Skrondal & Rabe-Hesketh, 2005). Latent classes at the lower level are homogeneous, while those at the upper level can be homogeneous or heterogeneous, dependent on model specification and research question. An optimum model is sought for all classes at all levels simultaneously. 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). If intercepts and slopes are fixed within classes at all upper levels, no distributional assumptions are required. As within single-level LCA, covariates can be modelled separately to the main association under investigation, as ‘class predictors’.

We use MLLCA to analyse the colorectal cancer data. Ultimately, no assumption of normality is made at the upper level, though initially a continuous latent variable for the upper level is adopted (as an approximation) while the latent class structure is explored for the lower level. Once the lower-level optimum number of classes is determined, the upper level latent variable is switched to categorical and the optimum upper-level latent class structure is determined. Stage at diagnosis is considered as a class predictor with the SEB-survival relationship held constant across the patient-level latent classes, thereby minimising the effect of the reversal paradox and potential bias introduced due to measurement error in the stage variable. Trust classes are homogeneous with respect to both patient outcome and its relationship with model covariates. This places the focus on patients and allows us to determine what kind of patient is potentially susceptible to the adverse impact of SEB in terms of their cancer survival status (i.e. we determine patient casemix characteristics in relation to outcome). An alternative approach of grouping Trusts according to differences in characteristics is discussed by Gilthorpe et al. (2011), where differences in survival at the Trust level may be as a result of underlying differences in Trust performance, rather than patient casemix.

#### 2.5 The Modelling

We use both likelihood-based model-fit criteria and a graphical method to determine the optimum number of latent classes at each level; we consider both the Bayesian Information Criterion (BIC) (Schwarz, 1978) (for reasons of parsimony) and the change in log likelihood (LL). We also examine and report classification error (CE), but do not use it to inform our model choice. CE reflects the proportion of misclassified observations (at each level separately) when comparing the modal and probabilistic assignment to classes; a lower CE signifies that the latent classes are more “real”, i.e. observations are almost entirely assigned to single classes. Models include adjustment for age at diagnosis, sex and SEB. An age-squared term is also included as age was found to have a non-linear relationship with survival; the inclusion of age-squared allows for an adjustment to the linear effect of age. Stage at diagnosis is a ‘class predictor’, and the ICD-10 diagnosis code, laterality and whether treatment is curative or not are ‘inactive’ covariates at the patient level, i.e. not used to estimate the model, though used to partition the findings by these variables for descriptive purposes. We generate 200 bootstrapped datasets and analyse each similarly in order to generate 95% confidence intervals (CIs). The software Stata (StataCorp, 2011) was used for data manipulation, summary statistics, tabulation and charts, while the statistical software LatentGOLD (Vermunt & Magidson, 2005) was used for the latent class analyses.

### 3. Results

#### 3.1 MLM Analysis and LCA Approach

Table 1. Results from MLM analysis (multilevel logistic regression): odds of death within 3 years

Model Statistics	Prevalence
Overall	51.6%
Reference Group	49.3%
Model Covariates	OR (95% CI)
SEB (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; the reference group consisted of males of mean age, diagnosed with Stage A colorectal cancer and with a zero Townsend score; LL = -11 985.

Table 1 shows the results of the MLM analysis. Overall, 12 708 patients (51.6%) died within three years. The reference group comprised males aged 71.5 years (the mean age), diagnosed with Stage A colorectal cancer and with a zero Townsend score. Substantial and statistically significant associations were found between increasing deprivation and increased odds of death (OR = 1.18, 95% CI = 1.15 to 1.21 per SD increase in Townsend 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).

With a continuous latent variable at the upper level, the MLLCA approach suggests that three patient classes are optimum by both the BIC and change in LL. Table 2 summarises the model-fit criteria for the multilevel latent-class models on switching the upper level latent variable to a categorical to determine the optimum upper-level latent class structure, and shows the optimum models identified by each criterion.

Table 2. Model-fit criteria for the three-patient multilevel latent-class models with a categorical upper level latent variable

Trust Classes	LL	BIC	Number of Parameters	Patient CE	Trust CE
1 class	-11 988	24 209	23	22.7%	0.0%
2 classes	-11 983	24 240	27	23.2%	10.6%
3 classes	-11 981	24 275	31	23.1%	10.1%
4 classes	-11 980	24 313	35	23.9%	12.9%
5 classes	-11 978	24 351	39	23.2%	17.8%
6 classes	-11 978	24 390	43	24.1%	21.4%
7 classes	-11 978	24 431	47	24.1%	30.5%
8 classes	-11 978	24 471	51	24.1%	36.5%

LL–Log likelihood; BIC–Bayesian Information Criterion; CE–Classification Error.

Table 2 shows that one Trust class is optimum by the BIC. More than one class at the Trust level is required to explain Trust differences however, therefore we consider also the -2LL plot shown in Figure 2, which shows model fit improving as the number of Trust classes increases. The standard MLM showed a LL of -11 985 which is surpassed by using two Trust classes. In order to model fully Trust variability and to improve patient class estimates we choose the model with five Trust classes, which lies at the point where there is little further improvement in model fit.



Figure 2. -2LL plot used to determine the optimum number of Trust classes in the MLLCA approach

## 3.2 Patient Classes

Table 3. Results for the patient classes in the three-patient, five-Trust-class multilevel latent-class model

Model Summary Statistics by class	Good	Reasonable	Poor	p-value
	Prognosis	Prognosis	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 Prevalence	9.4 (2.2,17.4)	58.3 (49.3,72.9)	93.2 (92.0, 99.6)	
Reference Group Prevalence	8.0 (0.1, 16.5)	57.8 (36.7, 78.6)	94.1 (90.8,100.0)	
% patients (bootstrapped 95% CI)				
Model Class Profiles				
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)	
Model Covariates	Good	Reasonable	Poor	
	Prognosis	Prognosis	Prognosis	
OR of death within three years (95% CI)				
SEB (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 reference group comprised males, aged 71.5 years, classified as Stage A at diagnosis and attributed a Townsend score of zero; the Wald p-value indicates levels of statistical significance for differences in effect across the classes. CIs from bootstrapping calculated using percentiles.

Table 3 summarises the patient classes from the chosen three-patient five-Trust-class MLLCA model, where patients were apportioned into one of three groups, labeled post-hoc as: good prognosis, reasonable prognosis, or poor prognosis. The good prognosis class contained 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.

The profile of stage differed across the patient classes. The good prognosis class corresponds to early-stage diagnosis with 70.8% of the stage A/B patients. The reasonable prognosis class corresponds to mid-stage diagnosis with 59.1% of the stage B/C patients, and a large proportion of patients with missing values for stage (30.5%). The poor prognosis class corresponds to late-stage diagnosis with 82.6% of the stage C/D patients. The good prognosis class contains the highest proportion of patients treated curatively (98.8%), which may be partly due to their stage at diagnosis as early-stage patients commonly receive curative, instead of palliative, treatment (National Institute for Clinical Excellence, 2004). There is little indication that either the type or position of the tumour is associated with survival status, as the proportions are broadly similar across the patient classes.

Across all patient classes, SEB was clearly associated with increased odds of death (OR = 1.33, 95% CI = 1.26 to 1.41). In the good prognosis patient class, females had significantly decreased odds of death compared with males (OR = 0.59, 95% CI = 0.40 to 0.87); in the reasonable and poor prognosis classes the association was 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). This indicates that women may fare better than men for early-stage disease, with diminishing differentiation for mid- to late-stage disease. Across all classes, older age was 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, the age-squared term was substantially and significantly 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 results do not differ markedly from those obtained when different numbers of Trust classes were considered. We also investigated models with three patient classes and one to six Trust classes. Stage remained an important predictor of survival status in every model with similar proportions of early-, mid- and late- stage diagnoses in the good, reasonable and poor prognosis groups respectively. SEB remained clearly associated with increased odds of death across all classes and in all models. Females maintained decreased odds of death in the good prognosis classes, although this did not reach statistical significance when considering four or six Trust classes. Finally, older age remained substantially and significantly associated with increased odds of death across all classes in models containing three or more Trust classes; when considering one or two Trust classes the association was less clear in the good and poor prognosis classes.

### 3.3 Trust Classes

Table 4. Results for the Trust classes in the three-patient, five-Trust-class multilevel latent-class model

Model Summary	Trust class 1	Trust class 2	Trust class 3	Trust class 4	Trust class 5
Statistics by Class	% patients (bootstrapped 95% CI)				
Class Size	37.3 (26.3, 64.4)	26.9 (15.7, 36.6)	14.3 (6.1, 24.9)	11.1 (3.7, 17.9)	10.4 (3.5, 14.7)
Prevalence	52.1 (50.0, 53.7)	50.9 (49.0, 54.2)	50.7 (48.2, 55.4)	49.6 (47.5, 57.5)	54.5 (47.2, 59.6)
Model Class Profiles	mean (bootstrapped 95% CI)				
Mean SEB	0.05 (-0.38, 0.41)	-0.05 (-0.60, 0.66)	0.38 (-0.99, 1.14)	-0.39 (-1.33, 1.26)	-0.45 (-1.82, 1.47)
Mean Age (years)	71.5 (71.2, 71.8)	71.8 (71.1, 72.1)	71.6 (70.9, 72.2)	71.2 (70.7, 72.6)	71.4 (70.8, 73.2)
Model Class Profiles	% patients (bootstrapped 95% CI)				
% Female	44.0 (42.8, 45.1)	44.3 (42.6, 45.5)	44.1 (41.4, 47.1)	43.4 (41.2, 47.6)	44.6 (41.2, 49.7)
Stage A	11.6 (10.7, 12.8)	11.6 (10.1, 13.3)	11.6 (9.6, 13.6)	12.2 (9.0, 13.7)	10.8 (9.0, 14.9)
Stage B	28.0 (26.3, 28.8)	27.1 (25.4, 28.9)	26.9 (25.1, 30.5)	26.9 (24.5, 31.0)	28.6 (24.0, 34.3)
Stage C	24.7 (23.5, 26.6)	26.6 (22.8, 27.9)	24.0 (21.8, 28.4)	25.3 (21.1, 29.4)	22.9 (18.6, 32.2)
Stage D	22.5 (21.7, 23.9)	23.0 (21.3, 24.5)	23.7 (20.6, 24.5)	22.7 (19.7, 25.0)	22.0 (17.8, 25.4)
Missing Stage	13.2 (11.8, 14.5)	11.7 (10.9, 15.1)	13.8 (10.5, 15.8)	12.9 (10.5, 16.9)	15.7 (8.8, 18.7)
Patients receiving treatment	83.0 (82.1, 84.8)	84.7 (81.9, 85.7)	82.7 (81.2, 86.5)	84.5 (80.0, 87.2)	81.7 (78.8, 89.1)
ICD-10 C18 (colon)	58.5 (27.2, 60.1)	58.9 (56.4, 60.9)	57.3 (55.0, 63.2)	59.1 (55.3, 63.4)	61.8 (55.3, 64.8)
ICD-10 C19 (rectosigmoid junction)	10.9 (9.2, 11.8)	10.2 (7.7, 12.0)	10.9 (7.4, 12.6)	11.1 (5.5, 13.0)	8.9 (5.2, 13.0)
ICD-10 C20 (rectum)	30.7 (29.4, 32.0)	30.9 (28.8, 32.9)	31.8 (27.1, 34.0)	29.8 (25.3, 34.5)	29.2 (24.4, 35.6)
Tumour on left side	65.7 (64.3, 67.6)	67.1 (64.0, 68.5)	67.2 (62.9, 69.2)	64.8 (62.2, 69.4)	63.9 (60.4, 70.3)
Tumour on right side	27.7 (26.3, 28.6)	27.7 (25.5, 29.0)	26.7 (25.2, 29.2)	27.3 (24.3, 29.8)	25.3 (23.1, 30.2)
Tumour across both sides	6.6 (5.1, 8.0)	5.2 (4.6, 9.0)	6.1 (4.3, 11.0)	7.9 (4.2, 11.4)	10.8 (4.1, 11.5)

CI—Confidence Interval; mean Townsend score over all classes -0.04; mean age over all classes 71.5 years. CIs from bootstrapping calculated using percentiles.

Table 4 summarises the chosen model for the Trust classes, where Trusts were apportioned into one of five groups. According to modal assignment, class 1 contained seven Trusts (37.3% of patients); class 2 contained five Trusts (26.9% of patients); class 3 contained three Trusts (14.3% of patients); and classes 4 and 5 contained two Trusts each (11.1% and 10.4% of patients respectively). According to probabilistic assignment, the prevalence rates ranged from 49.6% of patients dying within three years (in class 4) to 54.5% (in class 5).

The remainder of the results pertain to probabilistic assignment. SEB differed somewhat across the Trust classes, with the highest value seen in class 3 (mean SEB = 0.38), indicating that Trusts in this class receive patients on average from more deprived areas. In contrast, the lowest values of mean SEB are seen in classes 4 and 5 (mean SEB = -0.39 and -0.45 respectively), indicating that Trusts in these classes receive patients on average from more affluent areas. The mean age of patients remains fairly constant across the classes, ranging from 71.2 years (in class 4) to 71.8 years (in class 2). The proportion of females also remains fairly constant across the classes, ranging from 43.4% (in class 4) to 44.6% (in class 5). No substantial trend is seen across the Trust classes by stage, although

class 5 contains slightly more patients with missing values for stage (15.7%) compared with the other classes. The fewest patients received curative treatment in Trust class 5 (81.7% compared with 84.7%, the highest proportion, in class 2), which perhaps indicates that the two Trusts in this class are not treating as many early-stage patients as other Trusts. There are some modest differences seen across the Trust classes in type of tumour: class 5 has the highest proportion of colon tumours (61.8%) and the lowest proportion of rectosigmoid junction tumours (8.9%); while class 3 has the highest proportion of rectum tumours (31.8%). There are also some small differences seen in the laterality of the tumour: class 3 has the highest proportion of tumours on the left side (67.2%); classes 1 and 2 have the highest proportion of tumours on the right side (27.7%); and class 5 has the highest proportion of tumours across both sides (10.8%).

#### 4. Discussion

##### 4.1 Findings

The MLM analysis found sizeable and significant associations between increasing deprivation and increased odds of death, between being female and decreased odds of death, and between older age and increased odds of death. The MLLCA categorised patients into three latent classes, labelled as good, poor and reasonable prognosis (relating to stage at diagnosis) and in all classes, the impact of the covariates was found to agree with that seen in the MLM analysis. There were some differences across the prognosis groups, with the good prognosis group showing clearly decreased odds of death for females compared with males; and the reasonable prognosis group showing a greater impact of older age. The association between SEB and survival status was clear, with higher deprivation associated with increased odds of death in all patient classes (as the impact of SEB was deliberately held constant across the classes, i.e. there was no stage-SEB interaction, because this would otherwise have introduced circularity in the causal relationships amongst SEB, stage at diagnosis, and the latent classes within which the SEB-survival relationship varied). As discussed, previous findings into the association of SEB with survival from colorectal cancer have been seen to vary and this may in part depend upon whether these studies have undertaken appropriate statistical adjustment for alleged confounders, or introduced bias due to the reversal paradox.

MLLCA has considerable utility to account for issues of structure, non-homogeneity, inferred causality, missing values and measurement error, while improving upon the MLM approach by producing models that are better fitting to the data and provide an enhanced interpretation of the data. It allows for the hierarchical structure of the data without imposing any distributional assumptions. It accounts for non-homogeneity at all levels by categorising both patients and Trusts into latent classes and allowing the relationship between survival status and associated risk factors to vary across these classes (where appropriate), rather than modelling the relationship over all patients and across all Trusts. By modelling stage at diagnosis as a class predictor, bias due to the reversal paradox is minimised. As patient classes depend on stage, we investigate covariate-outcome associations within sub-categories of stage without introducing product interaction terms, thereby minimising bias due to measurement error. We take account of incomplete data within stage by categorising missing values and the modelling assigns patients with missing stage values to the most appropriate patient class according to similarities in their other characteristics compared with other patients.

The determination of Trust classes in the MLLCA was not straightforward. The best fit according to the more parsimonious measure of the BIC was one Trust class but this would not be sufficient to describe fully the natural structure of the data. Model fit according to the LL improved as the number of Trust classes increased, surpassing the fit of the standard multilevel model when considering only two Trust classes; the -2LL plot showed that the model fit continued to improve up to around five Trust classes (Figure 2). These five classes identified outlying Trusts, with predominantly three Trusts in Trust class 3 and two Trusts in each of Trust classes 4 and 5, indicating that the normal approximation at the upper level would not be ideal. This confirms our suspicion that the standard multilevel model would not be the most appropriate to model these data.

The five Trust classes differ only due to patient survival status and the relationship between survival and the covariates, potentially enabling us to highlight differences in patient care (e.g. treatment pathways or hospital characteristics, such as size or speciality) that might explain the differences and so be worthy of further investigation. It should be noted that the differences seen across the Trust classes are not statistically significant. Nevertheless, their inclusion was necessary in order to account for variability at the Trust level and so to model best the corresponding patient classes. No Trust level covariates were available for inclusion in the modelling, meaning that this potential aspect of investigation could not be addressed in this study.

## 4.2 Limitations

Although we minimise any potential bias that arises due to the reversal paradox by including stage at diagnosis as a class predictor, and by holding constant the SEB-survival relationship across classes, this bias may not be entirely eradicated and it is unclear how much bias could remain. No methodology can completely eliminate all bias and the use of MLLCA will have reduced the risk of bias in comparison with standard MLM techniques.

Although interest may lie in establishing *treatment* centre characteristics that may have an impact on survival from colorectal cancer, we have modelled *diagnostic* centre at the upper level. This allowed us to include all patients regardless of whether or not they received treatment. We modelled Trust of diagnosis at this level to minimise error that could be introduced by patients receiving treatment at different hospitals during their care, as a higher proportion of patients receiving treatment were treated within the same Trust at diagnosis (90.2%), than within the same hospital at diagnosis (81.7%).

We have included SEB at the patient level to simplify the analysis, though it is derived at the small area level. Individual measures of deprivation are rarely available, especially when using routine data. Extrapolating area-based findings to individuals, however, can lead to the ecological fallacy (Robinson, 1950). An additional upper level could be introduced relating to patients' area of residence and this would be cross-classified with Trusts, since patients attending hospitals in one Trust may be resident of different areas. We could also consider survival as a continuous measure, as within Cox proportional hazard modelling. Both these extensions could be accommodated within a MLLCA framework, though alternative software such as WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) or MPlus (L. K. Muthén & B. O. Muthén, 1998-2011) would then be required (although cross-classified extensions are still under development for MPlus). For simplicity of illustration of the methodology, however, we did not pursue these options within this study.

For this paper, we chose to focus on the utility and interpretability of an alternative modelling approach that has the potential to model appropriately both confounders and mediators while preserving the clinical message. Simulation studies may be beneficial to gain insight into the sensitivity of the data to model choice, and these could be considered as an extension to this study.

## 5. Conclusion

The MLLCA modelling approach illustrated a better fit to the data and showed new insights that were not previously apparent using the MLM approach. The impact of covariates on survival status differed across latent classes defined by stage at diagnosis. By tailoring treatments and pathways according to patients' profiles, there might be opportunities in the future to optimise patient care. This analytical strategy has prognostic utility to inform health service providers of disparities within patient care.

## Acknowledgements

The authors would 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.

## References

- Borowski, D. W., Bradburn, D. M., Mills, S. J., Bharathan, B., Wilson, R. G., Ratcliffe, A. A., & Kelly, S. B. (2010). Volume-outcome analysis of colorectal cancer-related outcomes. *British Journal of Surgery*, *97*, 1416-1430. <http://dx.doi.org/10.1002/bjs.7111>
- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.). London, UK: Chapman & Hall.
- Ciccolallo, L., Capocaccia, R., Coleman, M. P., Berrino, F., Coebergh, J. W., Damhuis, R. A., ... Williams, E. M. I. (2005). Survival differences between European and US patients with colorectal cancer: role of stage at diagnosis and surgery. *Gut*, *54*, 268-273. <http://dx.doi.org/10.1136/gut.2004.044214>
- Downing, A., Aravani, A., Macleod, U., Oliver, S., Finan, P. J., Thomas, J. D., ... Morris, E. J. (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*, 681-685. <http://dx.doi.org/10.1038/bjc.2012.585>
- Downing, A., Harrison, W. J., West, R. M., Forman, D., & Gilthorpe, M. S. (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, 772-776. <http://dx.doi.org/10.1136/jech.2008.085852>
- Dukes, C. E. (1949). The surgical pathology of rectal cancer. *Journal of Clinical Pathology*, 2, 95-98.
- Fuller, W. A. (1987). *Measurement Error Models*. New York, NY: Wiley.
- 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). <http://dx.doi.org/10.1186/1472-6963-11-53>
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215-231.
- Greenwood, D. C. (2012). Measurement Errors in Epidemiology. In D. C. Greenwood, & Y.-K. Tu (Eds.), *Modern Methods for Epidemiology* (pp. 33-55). London, UK: Springer.
- Greenwood, D. C., Gilthorpe, M. S., & Cade, J. E. (2006). The impact of imprecisely measured covariates on estimating gene-environment interactions. *BMC Medical Research Methodology*, 6(21), 4 May 2006. <http://dx.doi.org/10.1186/1471-2288-6-21>
- Hendifar, A., Yang, D., Lenz, F., Lurje, G., Pohl, A., Lenz, C., ... Lenz, H. J. (2009). Gender disparities in metastatic colorectal cancer survival. *Clinical Cancer Research*, 15, 6391-6397. <http://dx.doi.org/10.1158/1078-0432.CCR-09-0877>
- Hernández-Díaz, S., Schisterman, E. F., & Hernán, M. A. (2006). The birth weight “paradox” uncovered? *American Journal of Epidemiology*, 164, 1115-1120. <http://dx.doi.org/10.1093/aje/kwj275>
- Ionescu, M. V., Carey, F., Tait, I. S., & Steele, R. J. (1998). Socioeconomic status and stage at presentation of colorectal cancer. *The Lancet*, 352, 1439. [http://dx.doi.org/10.1016/S0140-6736\(98\)00052-X](http://dx.doi.org/10.1016/S0140-6736(98)00052-X)
- Kee, F., Wilson, R. H., Harper, C., Patterson, C. C., McCallion, K., Houston, R. F., ... Rowlands, B. J. (1999). Influence of hospital and clinician workload on survival from colorectal cancer: cohort study. *BMJ*, 318, 1381-1385. <http://dx.doi.org/10.1136/bmj.318.7195.1381>
- Kirkwood, B. R., & Sterne, J. A. (2003). *Essential Medical Statistics* (2nd ed.). Oxford, UK: Blackwell.
- Leyland, A. H., & Goldstein, H. (2001). *Multilevel Modelling of Health Statistics*. Chichester, UK: Wiley.
- Leyland, A. H., & Groenewegen, P. P. (2003). Multilevel modelling and public health policy. *Scandinavian Journal of Public Health*, 31, 267-274.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325-337.
- Magidson, J., & Vermunt, J. K. (2004). Latent class models. In D. Kaplan (Ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA: Sage Publications.
- McArdle, C. S., & Hole, D. J. (2004). Influence of volume and specialization on survival following surgery for colorectal cancer. *British Journal of Surgery*, 91, 610-617. <http://dx.doi.org/10.1002/bjs.4476>
- Morris, E. J., Maughan, N. J., Forman, D., & Quirke, P. (2007). Identifying stage III colorectal cancer patients: the influence of the patient, surgeon and pathologist. *Journal of Clinical Oncology*, 25, 2573-2579.
- Morris, E. J., Taylor, E. F., Thomas, J. D., Quirke, P., Finan, P. J., Coleman, M. P., ... Forman, D. (2011). Thirty-day postoperative mortality after colorectal cancer surgery in England. *Gut*, 60, 806-813. <http://dx.doi.org/10.1136/gut.2010.232181>
- Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus User's Guide* [Computer programme] (6th ed.). Los Angeles, CA: Muthén & Muthén.
- National Institute for Clinical Excellence. (2004). *Guidance on Cancer Services: Improving Outcomes in Colorectal Cancers-Manual Update*. London, UK: National Institute for Clinical Excellence.
- Normand, S.-L. T., Sykora, K., Li, P., Mamdani, M., Rochon, P. A., & Anderson, G. M. (2005). Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. *BMJ*, 330, 1021-1023.
- Nur, U., Rachet, B., Parmar, M. K., Sydes, M. R., Cooper, N., Lepage, C., ... Coleman, M. P. (2008). No

- socioeconomic inequalities in colorectal cancer survival within a randomised clinical trial. *British Journal of Cancer*, 99, 1923-1928. <http://dx.doi.org/10.1038/sj.bjc.6604743>
- Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge, UK: Cambridge University Press.
- Quirke, P., & Morris, E. (2007). Reporting colorectal cancer. *Histopathology*, 50, 103-112.
- Robinson, W. S. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review*, 15, 351-357.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Skrondal, A., & Rabe-Hesketh, S. (2005). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.
- Smith, J. J., Tilney, H. S., Heriot, A. G., Darzi, A. W., Forbes, H., Thompson, M. R., ... Tekkis, P. P. (2006). Social deprivation and outcomes in colorectal cancer. *British Journal of Surgery*, 93, 1123-1131. <http://dx.doi.org/10.1002/bjs.5357>
- StataCorp. (2011). *Stata statistical software: Release 12* [Computer programme]. College Station, TX: StataCorp LP.
- Stigler, S. M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*. Cambridge, MA: Harvard University Press.
- Townsend, P., Beattie, A., & Phillimore, P. (1987). *Health and Deprivation: Inequality and the North*. London, UK: Routledge.
- Tu, Y.-K., West, R., Ellison, G. T., & Gilthorpe, M. S. (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, 27-32.
- United Kingdom Association of Cancer Registries. (2008). *UKACR Quality and Performance Indicators 2008: Final*. UK: United Kingdom Association of Cancer Registries.
- Vermunt, J. K., & Magidson, J. (2005). *Latent GOLD 4.0 User's Guide* [Computer programme]. Belmont, Massachusetts: Statistical Innovations Inc.
- Whynes, D. K., Frew, E. J., Manghan, C. M., Scholefield, J. H., & Hardcastle, J. D. (2003). Colorectal cancer, screening and survival: the influence of socio-economic deprivation. *Public Health*, 117, 389-395. [http://dx.doi.org/10.1016/S0033-3506\(03\)00146-X](http://dx.doi.org/10.1016/S0033-3506(03)00146-X)
- Widdison, A. L., Barnett, S. W., & Betambeau, N. (2011). The impact of age on outcome after surgery for colorectal adenocarcinoma. *Annals of the Royal College of Surgeons of England*, 93, 445-450. <http://dx.doi.org/10.1308/003588411X587154>
- Woodman, C. B., Gibbs, A., Scott, N., Haboubi, N. Y., & Collins, S. (2001). Are differences in stage at presentation a credible explanation for reported differences in the survival of patients with colorectal cancer in Europe? *British Journal of Cancer*, 85, 787-790.
- 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.

## Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).