Temporal and geographic changes in stage at diagnosis in England during 2008-2013: a population-based study of colorectal, lung and ovarian cancers

Patrick Muller, Laura Woods, Sarah Walters

Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT

Article word count: 3,254

Abstract word count: 250

Tables: 3

Figures: 2

Keywords: cancer; stage at diagnosis; early diagnosis of cancer; time trends; temporal changes; geographic inequalities; population surveillance; case-mix adjustment; missing data; multiple imputation

ABSTRACT

Background

Increasing diagnosis of cancer when the disease is still at early stages is a priority of cancer policy internationally. In England, reducing geographical inequalities in early diagnosis is also a key objective. Stage at diagnosis is not recorded for many patients, which may bias assessments of progress. We evaluate temporal and geographical changes in stage at diagnosis during 2008-2013 for colorectal, non-small cell lung, and ovarian cancers, using multiple imputation to minimise bias from missing data.

Methods

Population-based data from cancer registrations, routes to diagnosis, secondary care, and clinical audits were individually linked. Patient characteristics and recorded stage were summarised. Stage was imputed where missing using auxiliary information (including patient's survival time). Logistic regression was used to estimate temporal and geographical changes in early diagnosis adjusted for case mix using a multilevel model.

Results

We analysed 196,511 colorectal, 180,048 non-small cell lung, and 29,076 ovarian cancer patients. We estimate that there were very large increases in the percentage of patients diagnosed at stages I or II between 2008-09 and 2012-13: from 32% to 44% for colorectal cancer, 19% to 25% for non-small cell lung cancer, and 28% to 31% for ovarian cancer. Geographical inequalities reduced for colorectal and ovarian cancer.

Interpretation

Multiple imputation is an optimal approach to reduce bias from missing data, but residual bias may be present in these estimates. Increases in early-stage diagnosis coincided with increased diagnosis through the "two week wait" pathway and colorectal screening. Epidemiological analyses from 2013 are needed to evaluate continued progress.

1. INTRODUCTION

Diagnosis of cancer when the disease is still at an early stage is associated with markedly improved survival prospects [1, 2]. Increasing the proportion of patients diagnosed at early stages (often defined as stages I or II) is a focus of cancer policy in the UK and internationally [3-9].

In England, increased early diagnosis has been identified as one means to reduce the survival gap with other affluent countries [10]. Numerous early diagnosis targets and interventions have been initiated. In 2000 a target was introduced that no patient should have more than a two-week wait (2WW) to see a cancer specialist following general practitioner (GP) referral with possible cancer symptoms [7]. From 2007 that target was extended to include patients referred from a hospital or through screening [11]. In 2005 national guidance for GPs on referring patients with possible cancer symptoms to specialists was published [12]. Faecal Occult Blood Test (FOBT) screening for colorectal cancer was rolled out nationally during 2006-2009 [13], and from 2011 the 'Be Clear on Cancer' campaign has raised awareness of the symptoms from common and rarer cancers, and encouraged people to report them to their GP [14]. In 2015 it became national policy that by 2020 62% of staged cancers should be diagnosed at stages I or II; that the proportion of cancers staged should increase; and that inequalities between the local healthcare commissioners (Clinical Commissioning Groups - CCGs) should decrease [3].

To monitor progress against these targets, from 2016 Public Health England have produced a public-facing website of cancer statistics, the *CancerData* dashboard [15]. The dashboard presents the percentage of cancers with the disease stage recorded, and the percentage of those diagnosed at stages I or II, nationally and for each CCG, for each year from 2012. This "stages I or II" percentage is needed to monitor progress against the target set in 2015. However, it may be biased if used for analyses of changes in stage in the whole population, as it excludes patients whose stage was not ascertained or not collected centrally. Nationally, stage recording increased dramatically from 2008 but still only covered 71% of patients in 2013 [15]. The patients without recorded stage have poorer outcomes than patients with stage recorded, suggesting a less favourable underlying stage distribution [16, 17].

In addition to missing data, a consideration when interpreting stage trends is the extent to which observed changes are due changes in health services, for example the introduction of a screening programme, or patient case mix, such

as a decrease in the incidence of hard-to-detect tumours. Case-mix differences have been found to confound CCG rankings of early-stage diagnosis [18], and may also influence temporal comparisons.

In this study we analyse temporal and geographic differences in stage at diagnosis during 2008-2013, using statistical techniques to account for missing data and case mix differences. We analyse three malignancies commonly diagnosed late: colorectal cancer, non-small cell lung cancer (NSCLC), and ovarian cancer. We evaluate whether the number of patients diagnosed at stages I or II increased; whether geographic inequalities increased or decreased; and whether observed changes are associated with case-mix. Multiple imputation is employed to minimise bias from missing stage data [19-21].

2. MATERIALS AND METHODS

2.1 Data

Data on cancer registrations were obtained from the Office for National Statistics (ONS) for adults aged 15–99 years, diagnosed with colorectal cancer, NSCLC or ovarian cancer in England from 1 January 2008 to 31 December 2013 (ICD-10 codes C18-20, C21.8; C33-34; and C56-C57.7 [22]). Data on patient's vital status was complete up to 31 December 2014. Data were additionally linked to the national bowel and lung cancer audit datasets [23, 24], the Routes to Diagnosis (RtD) dataset [25], and Hospital Episodes Statistics (HES) records using patient's NHS number and postcode. The audit datasets were used to gain additional information on stage [26], whilst RtD records provided information about patient's interactions with the National Health Service (NHS) before diagnosis. The HES records provided information on receipt of major surgical treatment following diagnosis (based on OPCS Classification for Interventions and Procedures version 4 codes; full list in appendices) and Charlson Comorbidity Index (CCI - derived from HES records from 6 to 60 months prior to cancer diagnosis) [27].

Clinical Commissioning Group (CCG) areas were used to examine geographical inequalities. These territories were chosen as they have been responsible for commissioning cancer services from 2013, following the dissolution of the Primary Care Trusts which were previously responsible for cancer treatment. Differences in the proportion of patients diagnosed at an early stage were compared between three time periods: 2008-09, 2010-11, and 2012-13.

Two-year periods were chosen to ensure each had sufficient numbers of patients for a robust comparison of geographic inequalities.

2.2 Descriptive analysis

Temporal changes in the distribution of stage at diagnosis (I, II, III, IV, or missing) were evaluated. The percentage of patients diagnosed at stage I or II (of those with a recorded stage) was tabulated by patient characteristics. The association between each characteristic and missing stage was assessed.

2.3 Multiple imputation

Multiple imputation was conducted to estimate patient's stage of disease at diagnosis where unknown [28]. Imputation models including auxiliary patient information were fitted with the R package *jomo* [29], which accounts for the multi-level structure of the data (patients clustered within CCGs). It was assumed that stage was missing randomly conditional on variables strongly associated with either stage (I to IV) [16, 30-32], or with recording of stage [17]: quarter year of diagnosis, cancer registry area, CCG, age, sex, patient's Indices of Multiple deprivation (IMD) income quintile, Charlson comorbidity score, tumour topography, tumour morphology, route to diagnosis, receipt of major surgical treatment (yes/no), treatment admission method (elective/non-elective), time from diagnosis to censoring, and vital status at censoring. Cancer registry area of diagnosis was included as well as CCG, as historically the regional registries recorded stage at different levels of completeness for different tumours.

Tumour morphology and topography included categories which were uninformative as to the actual values ("non-specific", "miscellaneous and unspecified"). These values were re-coded to true missing and imputed using *jomo* alongside missing stage.

The number of imputation datasets created was equal to the percentage of missing data for each cancer: 39, 20, and 41 respectively for colorectal cancer, NSCLC, and ovarian cancer. These numbers are sufficient to achieve a <1% power reduction compared to using n=100 datasets [33]. Parameter estimates (percentages of patients at different

stages, and regression model parameters) were produced using each dataset and combined using Rubin's rules [34]. Full details of the imputation and examples of the R code used are in the supplementary appendices.

2.4 Regression modelling

The change in the odds of diagnosis at stages I or II between the two-year time periods was estimated using multilevel logistic regression models. Parameters to estimate the between-CCG variation in diagnosis at stages I or II in each time period were fitted and compared using Wald tests. These were also used to estimate odds ratios for CCGs at the 2.5th, 25th, 75th, and 97.5th percentiles, to illustrate the differences in early diagnosis odds between average CCGs and those with highest and lowest percentages of patients diagnosed early. The models were fitted in STATA using *meqrlogit* [35].

The first set of models included only time period and CCG as explanatory variables. A second set of case-mix adjusted variables were fitted including these variables along with age, sex, CCI, tumour topography, and tumour morphology. Further details on the model specification are provided the supplementary appendices.

3. RESULTS

We analysed cancer registrations of 196,511 colorectal, 180,048 NSCLC, and 29,076 ovarian cancer patients diagnosed during the period 2008-2013 (Table 1). On average in each time period and in each CCG there were 313, 287, and 46 new diagnoses of colorectal cancer, NSCLC, and ovarian cancer respectively (Appendix Table 1).

3.1 Descriptive analyses

Amongst patients with stage recorded, the percentage diagnosed at stages I or II increased dramatically over time for colorectal cancer (from 31.0% in 2008-09 to 45.0% in 2012-13) and NSCLC (from 20.0% to 25.6%), whilst for ovarian cancer it remained similar (from 33.3% to 32.9%).

For colorectal cancer lower deprivation, higher Charlson score, diagnosis following screening or GP referral, and non-carcinoma disease morphology were all associated with early-stage diagnosis (Table 1). For NSCLC, factors associated with early diagnosis included female sex, higher Charlson score, diagnosis following referral from a GP or outpatient service, tumour origin in the lobe as opposed to main bronchus, and carcinoma morphology. For ovarian cancer early diagnosis was associated with younger age, lower Charlson score, type I epithelial or non-epithelial disease, and diagnosis following referral from a GP or outpatient service. For all cancers, receipt of major treatment and elective admission for treatment were strongly associated with early diagnosis.

Table 1 Numbers of patients and percentage diagnosed at stages I or II by age, sex, diagnosis period, deprivation, comorbidity, tumour topography, tumour morphology, route to diagnosis, cancer registry, major treatment and admission method

	Colorectal cancer			NSCLC			Ovarian cancer		
	Count	Stage I/II (%)*	Missing stage (%)	Count	Stage I/II (%)*	Missing stage (%)	Count	Stage I/II (%)*	Missing stage (%)
Total									
All patients	196,511	40.5	39.2	180,048	23.1	20.1	29,076	33.8	40.7
Age									
15-39	3,458	37.5	41.8	865	31.4	29.6	1,368	64.3	38.7
40-49	7,423	33.3	35.3	4,158	18.5	17.9	2,551	50.3	34.5
50-59	20,763	36.0	33.5	17,099	19.5	14.8	4,905	42.2	33.1
60-69	50,801	41.7	36.9	46,325	23.2	15.6	7,743	30.2	36.6
70-79	60,785	42.3	37.4	61,184	24.6	18.7	7,181	25.6	41.4
80-99	53,281	40.2	46.0	50,417	22.6	27.5	5,328	21.5	55.9
Sex									
Male	110,042	40.4	38.0	100,176	22.0	19.7			
Female	86,469	40.5	40.7	79,872	24.5	20.5			
Diagnosis period									
2008-2009	63,972	31.0	63.0	57,382	20.0	34.2	9,618	33.3	57.1
2010-2011	66,113	39.8	39.0	60,103	22.5	18.1	9,863	35.4	45.7
2012-2013	66,426	45.0	16.4	62,563	25.6	8.9	9,595	32.9	19.0
Deprivation quintile									
1 (Least deprived)	42,040	41.8	39.3	25,083	22.7	21.2	6,109	33.5	39.2
2	43,913	40.7	38.6	32,024	23.1	20.3	6,465	31.8	39.8
3	41,033	40.4	39.1	35,995	22.1	20.3	6,228	34.3	42.1
4	36,972	39.5	39.8	39,872	22.9	19.9	5,560	33.4	41.6
5 (Most deprived)	32,553	39.4	39.2	47,074	24.2	19.2	4,714	36.5	40.8
Charlson comorbidity Index									
0	156,968	39.9	38.4	123,622	20.7	19.2	24,896	34.5	39.6
1	18,596	43.1	41.8	28,211	29.0	21.1	2,244	29.6	44.5
2	11,347	41.3	41.1	13,873	28.1	21.9	1,145	32.3	47.0

3+	9,600	43.4	44.8		14,342	28.1	23.8		791	21.7	52.2
Topography										
Colon	127,152	39.4	40.5	Main Bronchus	8,953	7.8	18.8	Ovary	28,181	33.3	40.8
Rectum	69,359	42.2	36.9	Lobe	119,253	28.4	13.9	Fallopian tube	895	46.4	34.7
Missing				Missing	51,842	10.4	34.4				
Morphology										
Carcinoma	156,743	40.8	36.1	Carcinoma	91,659	26.6	12.4	Type I epithelial	5,350	77.1	28.4
Non-carcinoma	19,014	60.6	45.5	Non-carcinoma	70,198	17.9	24.1	Type II epithelial	21,066	19.7	39.9
Missing	20,754	12.8	56.9	Missing	18,191	22.5	43.2	Non-epithelial	873	76.7	52.3
								Missing	1,787	17.2	80.5
Route to diagnosis										
GP referral	45,760	43.7	37.6		37,907	30.5	18.1		6,319	42.5	39.5
Two-week wait	54,249	41.3	29.7		46,647	24.2	8.5		8,600	38.1	28.5
Emergency presentation	44,631	26.4	42.8		64,131	12.7	26.8		9,008	15.4	49.6
Inpatient elective	7,337	42.6	38.9		2,724	14.6	19.9		366	38.5	43.2
Other outpatient	14,027	44.9	41.5		19,510	36.9	18.8		3,172	45.8	40.8
Screening	16,557	59.6	32.6								
Unknown	13,950	32.7	75.5		9,129	24.8	42.7		1,611	33.4	58.9
Registry										
North & York	27,139	41.4	38.4		31,102	25.4	19.7		3,501	33.7	26.7
Trent	20,082	38.4	52.5		19,096	24.1	22.9		2,806	44.9	55.4
East Anglia	23,718	39.9	35.1		19,076	20.1	16.2		3,560	25.9	52.6
Thames	36,843	40.6	50.5		32,892	21.2	24.5		5,768	32.8	43.8
Oxford	10,360	42.7	48.3		7,987	25.1	24.9		1,657	40.8	44.2
South & West	31,065	41.5	29.1		23,551	21.1	18.4		4,636	29.1	28.7
West Midlands	21,324	39.7	30.5		18,412	21.9	16.2		3,290	33.6	38.2
Northwest & Mersey	25,980	39.7	33.0		27,932	26.0	18.5		3,858	39.1	41.8
Major treatment										
Yes	120,096	50.0	35.2		23,804	80.6	11.7		16,353	46.4	29.5
No	76,415	22.7	45.5		156,244	13.2	21.3		12,723	8.4	55.0
Treatment admission method**										
Elective	95,933	53.9	34.3		23,334	80.9	11.7		14,967	46.2	29.1
Non-elective	24,163	33.4	38.7		470	66.8	16.2		1,386	48.1	34.3

* Of patients with a recorded stage

** Of patients who received major treatment

Overall, stage was not recorded for 39.2%, 20.1%, and 40.7% of colorectal, NSCLC, and ovarian cancer patients respectively. The percentage of patients missing stage decreased dramatically over time for all three cancers, from 34.2-63.0% in 2008-09 to 8.9-19.0% in 2012-13. As stage recording improved, the prognostic characteristics of

patients without a recorded stage became less favourable: emergency presentation and pre-existing comorbidities became more common (Appendix Table 2).

Lack of a recorded stage was more common amongst patients who were very young or old compared to the rest of the cohort. Pre-diagnosis it was associated with pre-existing comorbidities and the emergency diagnosis. Post-diagnosis it was associated with a lower probability of receiving major treatment, and with non-elective (unplanned) admission. Lack of recorded stage was also associated with absence of records on tumour topography and morphology.

3.2 Temporal changes

Multiple imputation-based estimates indicate large increases in the percentage of colorectal cancer patients diagnosed at stages I or II nationally, from 32% in 2008-09 to 44% in 2012-13 (compared to 31% to 45% in the complete case analysis; Figure 1). For NSCLC the stages I or II percentage increased from 19% to 25% (compared to 20% to 26%). For ovarian cancer it rose from 28% to 31% (compared to remaining at 33%). These estimates also provide evidence of a stage shift from IV to III for colorectal and ovarian cancers: the percentage of stage III tumours amongst all stages III or IV rose from 37% to 48% for colorectal cancer and from 44% to 59% for ovarian cancer (Table 2).

Figure 1 Distribution of stage at diagnosis in England: comparison of crude results (left) and distribution based on multiple imputation (right) by quarter-year of diagnosis, 2008-2013.

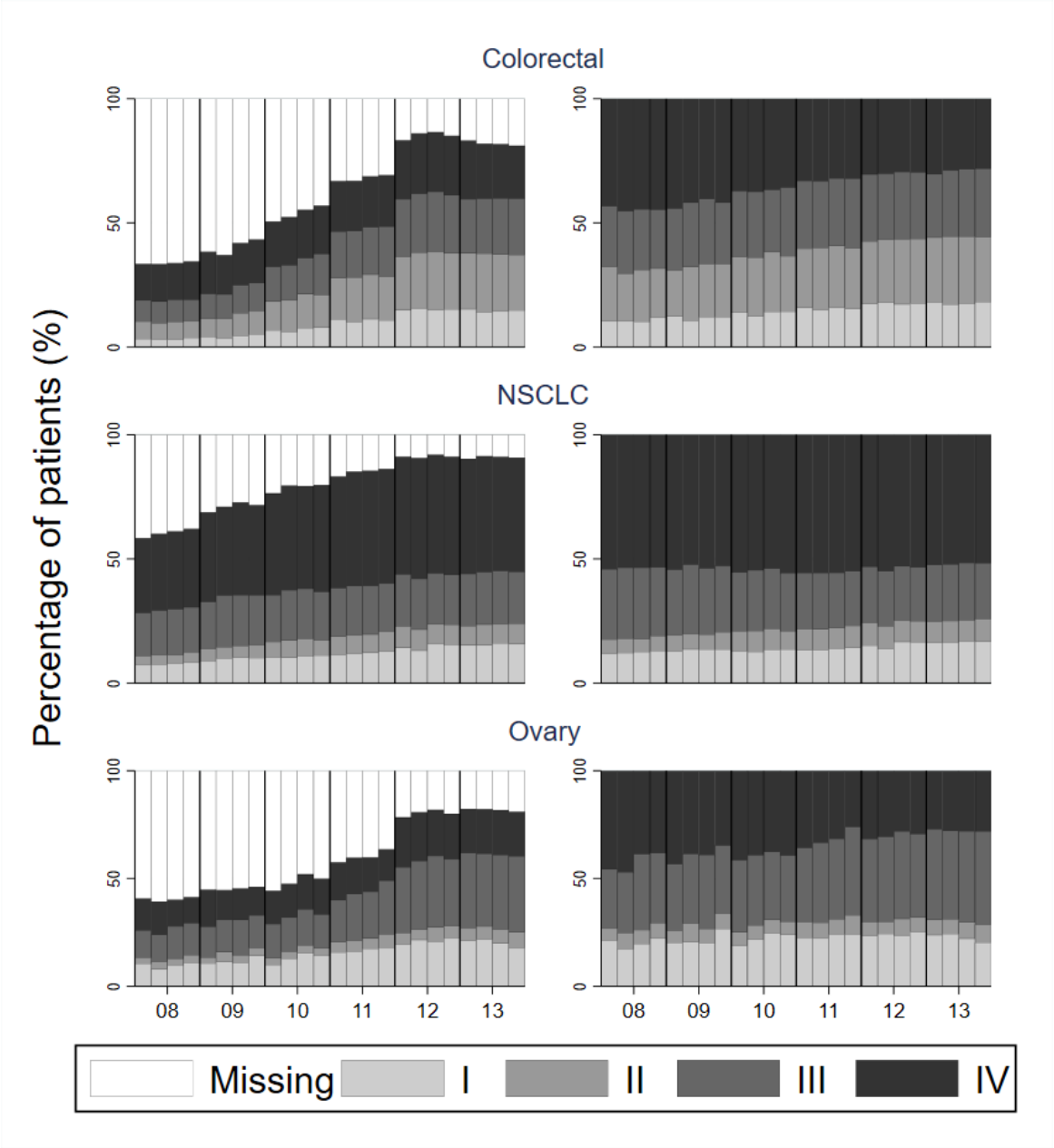


Table 2 Distribution of stage from multiple imputation estimates, by period and cancer

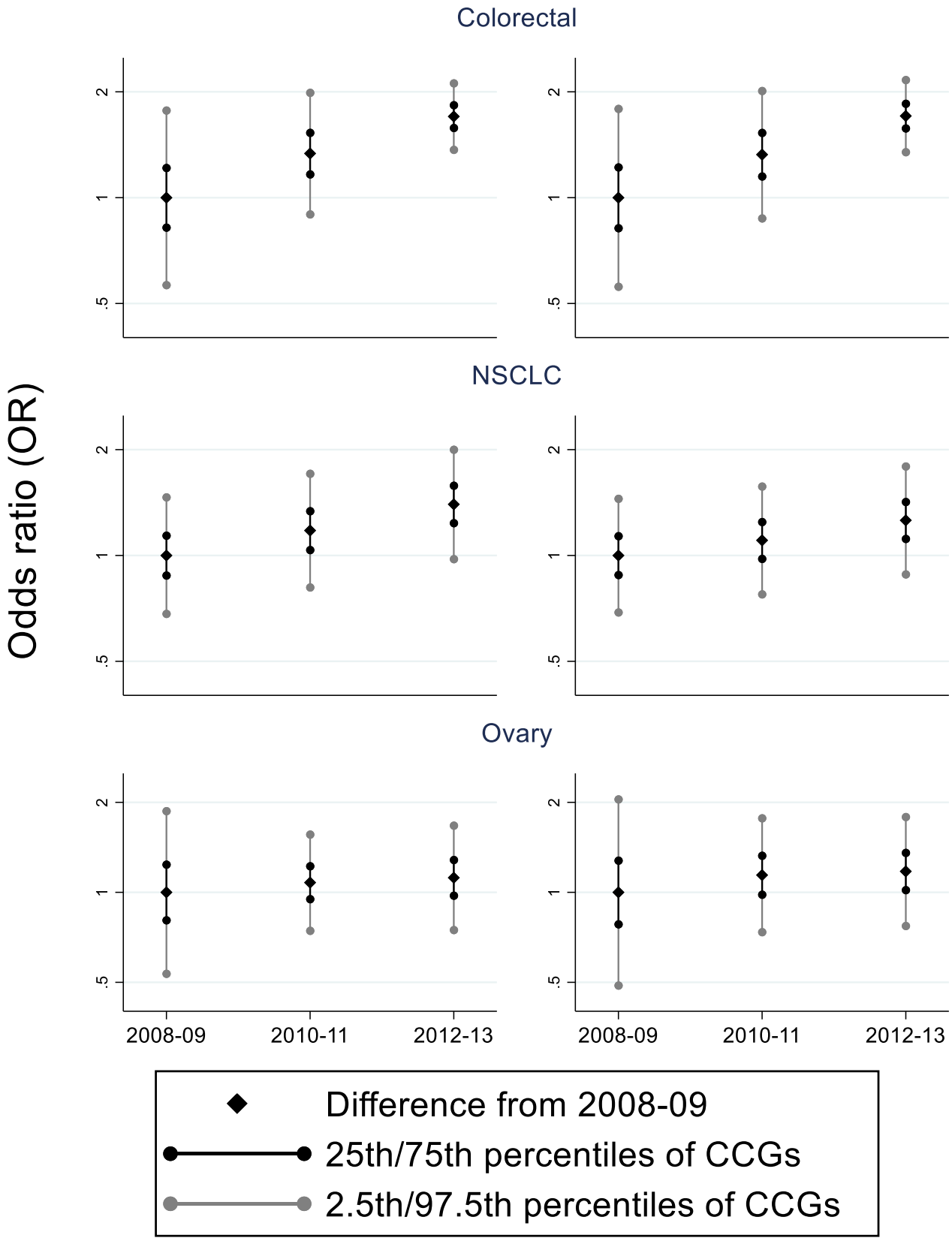
	2008-2009	2010-2011	2012-2013
		Colorectal cancer	
Patients diagnosed (N)	**63,972**	**66,113**	**66,426**
Stage I/II % (95% CI)	31.9 (31.4, 32.5)	38.5 (38.1, 39.0)	43.8 (43.4, 44.3)
Stage I % (95% CI)	11.3 (10.8, 11.8)	14.6 (14.3, 15.0)	17.7 (17.3, 18.0)
Stage II % (95% CI)	20.6 (20.0, 21.2)	23.9 (23.5, 24.3)	26.2 (25.8, 26.5)
Stage III % (95% CI)	24.9 (24.4, 25.5)	26.9 (26.4, 27.3)	26.8 (26.4, 27.1)
Stage IV % (95% CI)	43.1 (42.5, 43.8)	34.6 (34.1, 35.1)	29.4 (29.0, 29.8)
		NSCLC	
Patients diagnosed (N)	**57,382**	**60,103**	**62,563**
Stage I/II % (95% CI)	18.9 (18.5, 19.3)	21.7 (21.3, 22.0)	24.8 (24.5, 25.1)
Stage I % (95% CI)	12.9 (12.6, 13.2)	13.5 (13.2, 13.7)	16.1 (15.8, 16.4)
Stage II % (95% CI)	6.0 (5.8, 6.3)	8.2 (8.0, 8.5)	8.7 (8.5, 8.9)
Stage III % (95% CI)	27.7 (27.2, 28.2)	23.3 (22.9, 23.6)	22.5 (22.1, 22.8)
Stage IV % (95% CI)	53.4 (52.9, 53.9)	55.1 (54.6, 55.5)	52.7 (52.3, 53.1)
		Ovarian Cancer	
Patients diagnosed (N)	**9,618**	**9,863**	**9,595**
Stage I/II % (95% CI)	27.9 (26.7, 29.1)	29.7 (28.7, 30.8)	30.5 (29.5, 31.4)
Stage I % (95% CI)	21.1 (20.0, 22.2)	22.8 (21.9, 23.8)	23.4 (22.5, 24.3)
Stage II % (95% CI)	6.8 (6.2, 7.5)	6.9 (6.2, 7.6)	7.0 (6.5, 7.6)
Stage III % (95% CI)	31.7 (30.2, 33.2)	34.8 (33.6, 36.0)	40.7 (39.6, 41.8)
Stage IV % (95% CI)	40.4 (39.0, 41.9)	35.5 (34.3, 36.6)	28.8 (27.8, 29.8)

For NSCLC, early diagnosis is estimated to have increased between 2008-09 and 2012-13 in both models with and without case mix adjustment (Table 3, Figure 2). However, it was a smaller increase in the case-mix adjusted model (OR: 1.26 (95%CI: 1.20, 1.32) compared to 1.40 (95% CI: 1.33, 1.47), Table 3). During 2008-2013 the case mix for NSCLC shifted towards carcinomas (59.3% by 2012-13 compared to 49.5% in 2008-09, Appendix Table 3)), tumours originating in a lobe, and patients with pre-existing comorbidities; all characteristics associated with earlier diagnosis.

By contrast, for ovarian cancer early diagnosis is estimated to have increased in both models, but it was a greater increase in the model in which case mix is adjusted for (OR: 1.17 (95%CI: 1.05, 1.31) compared to 1.12 (95%CI: 1.02, 1.22)).

For colorectal cancer estimates were similar between the models in which case mix was and wasn't adjusted for.

Figure 2 Model-based estimates for change in odds of diagnosis at stages I or II, and change in between-CCG variation, in England during 2008-2013: comparison of un-adjusted (left) and case-mix adjusted (right) estimates.



3.3 Geographic inequalities

Geographic inequalities in early diagnosis decreased over time for colorectal cancer and ovarian cancer in models where case-mix was not considered (both $p<0.05$, Table 3). Geographic inequalities for NSCLC were smaller than for the other two cancers in 2008-09, but there is no evidence that they decreased. Case-mix adjustment had little impact on the magnitude of inequalities, or on changes in inequalities over time.

Table 3: Multi-level logistic regression results: Odds ratios (OR) for change in geographic inequalities and in probability of diagnosis at stages I or II during 2008-13

	No case mix adjustment			Case mix adjustment done***		
	2008-09	2010-11	2012-13	2008-09	2010-11	2012-13
Colorectal cancer						
OR for difference between time periods (95% CI)	1.00	1.34 (1.26, 1.42)	1.70 (1.61, 1.80)	1.00	1.33 (1.25, 1.40)	1.71 (1.62, 1.80)
Between-CCG variation (95% CI)*	0.09 (0.06, 0.12)	0.04 (0.03, 0.06)	0.01 (0.01, 0.02)	0.09 (0.07, 0.11)	0.05 (0.03, 0.06)	0.01 (0.01, 0.02)
P-value for differences in between-CCG variability**	.	<0.01	<0.01	.	<0.01	<0.01
OR for CCG at 2.5th percentile in period	0.56	0.67	0.80	0.56	0.66	0.79
OR for CCG at 97.5th percentile in period	1.77	1.49	1.24	1.79	1.52	1.27
NSCLC						
OR for difference between time periods (95% CI)	1.00	1.18 (1.12, 1.24)	1.40 (1.33, 1.47)	1.00	1.10 (1.05, 1.16)	1.26 (1.20, 1.32)
Between-CCG variation (95% CI)*	0.04 (0.03, 0.06)	0.04 (0.03, 0.05)	0.03 (0.02, 0.05)	0.04 (0.03, 0.05)	0.03 (0.02, 0.05)	0.03 (0.02, 0.04)
P-value for differences in between-CCG variability**	.	0.84	0.61	.	0.76	0.68
OR for CCG at 2.5th percentile in period	0.68	0.69	0.70	0.69	0.70	0.70
OR for CCG at 97.5th percentile in period	1.46	1.45	1.43	1.45	1.42	1.42
Ovarian cancer						
OR for difference between time periods (95% CI)	1.00	1.08 (0.98, 1.18)	1.12 (1.02, 1.23)	1.00	1.14 (1.02, 1.28)	1.18 (1.05, 1.31)
Between-CCG variation (95% CI)*	0.10 (0.06, 0.17)	0.04 (0.01, 0.10)	0.04 (0.02, 0.09)	0.13 (0.08, 0.23)	0.05 (0.02, 0.13)	0.05 (0.02, 0.12)
P-value for differences in between-CCG variability**	.	0.03	0.04	.	0.05	0.04
OR for CCG at 2.5th percentile in period	0.53	0.69	0.67	0.49	0.65	0.66
OR for CCG at 97.5th percentile in period	1.87	1.45	1.5	2.05	1.55	1.52

* Estimated between CCG-variance on log scale

** Comparing between-CCG variation between 2008-09 and 2010-11; and variability between 2008-09 and 2012-13

*** Factors adjusted for: age, sex, comorbidity status, tumour morphology, tumour topography

In 2008-09, patients in CCGs with the lowest percentages of patients diagnosed early had 30-50% lower odds of early diagnosis compared to patients in an average CCG, even after adjustment for case mix (Table 3). By 2012-13 the gap had reduced to 20-30%. For colorectal cancer the reduction in CCG inequalities equate to an approximate between-CCG range in diagnosis at stages I or II of 21-46% in 2008-09, reducing to 38-50% in 2012-13 (Table 2, Table 3). For ovarian cancer they equate to a range of 16-44% in 2008-09 reducing to 22-40% in 2012-13

4. DISCUSSION

We report evidence for substantial increases in the percentage of patients diagnosed early in England during 2008-2013. Geographic inequalities in early diagnosis between CCGs were present in all time periods, but reduced substantially during 2008-2013 for colorectal cancer and ovarian cancer. Case-mix differences did not account for the changes we observed.

4.1 Strengths

We used multiple imputation – a gold-standard approach to minimise bias in cases where a fraction of data is irretrievably missing - in estimates of national changes in stage at diagnosis [19, 20].

The patients missing stage data had poorer outcomes [16, 17], were older, had more comorbidities, were more commonly diagnosed as an emergency, and less likely to receive major treatment. We estimate that excluding them leads to overstatement of early-stage diagnosis by 1-5 percentage points. For colorectal cancer and NSCLC estimates of improvements were similar whether or not patients missing stage were included, but for ovarian cancer their exclusion lead to a different conclusion of no improvement. This finding shows that patients missing recorded stage need to be considered when evaluating progress in early-stage diagnosis, to avoid bias.

We also found that emergency presentation became more common amongst patients missing stage during 2008-2013, whilst nationally it became less common. This indicates that the increase in stage recording has been skewed towards patients with better prognostic characteristics. If this trend continues surveillance that excludes patients

missing stage may become less representative of changes in the population, as patients with a recorded stage become less similar to patients without one.

We were able to exclude case-mix factors as an explanation for improvements over time. The changes in case mix also provide contextual information: the observed (unadjusted) increases in early diagnosis over time for NSCLC occurred as the case-mix skewed towards more carcinomas and increased comorbidities (factors associated with earlier diagnosis; comorbidities potentially due to incidental detection during X-ray for another condition), whilst for ovarian cancer early diagnosis increased despite a shift towards increased comorbidities and type II epithelial disease (factors associated with later diagnosis).

4.2 Limitations

We imputed missing stage information by assuming that it is missing randomly conditional on all the other information available, including on patient's subsequent survival ("missing at random", MAR). It is likely that this assumption is not entirely met, and that our approach reduced but did not eliminate bias. More work is needed to understand the mechanisms for missing data in England and evaluate how bias from it can be reduced. For example, misspecification of the imputation model could affect the magnitude of increases reported. However, given the very large effect estimates for changes in early diagnosis and geographic inequalities we found it unlikely that residual bias would change our overall conclusions.

Another restriction of this study is that data after reform of the NHS in 2013 were not available. This reform may have had a positive or negative impact on early diagnosis. Additionally, from 2010 NHS funding increases were lower in real terms than in previous years, and failed to keep pace with increases in demand, resulting in the need for efficiency savings and reductions in per-head spend on cancer by 2011 [36]. Increased waits for GP appointments and at A&E departments have also been documented [37, 38], and the pressure on these gateway services may have affected early diagnosis. Our analysis can't be used to assess the long-term impact from these changes.

4.3 Effect of early diagnosis interventions in England

The introduction of FOBT screening and increasing referrals under the two-week wait (TWW) urgent GP referral route to diagnosis are likely to have played a role in the large increases in early diagnosis for colorectal cancer. Between 2008-09 and 2012-13 the percentage of patients diagnosed through screening rose from 6% to 10% and diagnoses through TWW rose from 26% to 30%, corresponding to 2,600 more patients per annum diagnosed through these routes (Appendix Table 5).

We estimate that by 2012-13 there were almost 4,000 fewer new colorectal cancer diagnoses at stage IV annually (Table 2), and corresponding increases in diagnoses at stages I, II and, to an extent, III. One concern about screening programmes is the increased risk of overdiagnosis and corresponding increase in unnecessary treatment of low-grade/benign tumours [39]. If overdiagnosis increases, there be increases in incidence, early diagnosis, and survival without benefit to patients. Our estimates indicate that early diagnosis increases for colorectal cancer during 2008-2013 are unlikely to be due to increased overdiagnosis. This is because incidence rose only slightly, whilst the absolute number of diagnoses at stage IV dropped substantially.

For NSCLC and ovarian cancers there were also large increases in 2WW diagnoses in this period. These may have resulted from the introduction of GP referral guidelines and symptom awareness campaigns.

The changes in early diagnosis during 2008-2013 occurred following sustained government investment in cancer control initiated through the national Cancer Plan in 2000 (which promised an additional £570 million for cancer by 2003-04) [7]. Though it is probable that the increased spending coupled with this plan (and subsequent extensions to it in 2007 and 2011 [8, 11]) led to improvements in early-stage diagnosis, empirical data supporting it have thus far been sparse. Our study provides evidence for a stark improvement. It seems likely that these cancer plans have at least in part led to this, as well as contributing to reduced geographic inequalities, if national referral guidelines and targets for cancer have helped standardise patient pathways across the country.

4.4 Conclusion and recommendations

We report very large increases in the percentage of patients diagnosed at stages I or II for colorectal cancer and NSCLC during 2008-2013, and a smaller increase for ovarian cancer. The increases we report may be subject to residual bias from missing stage data, however the overall conclusion of large improvements is robust to some misestimation. Increased investment and more frequent diagnoses through screening (for colorectal cancer only) and the two-week wait route to diagnosis are likely to have contributed to the increases. Geographic inequalities reduced considerably for colorectal and ovarian cancer over the same time period.

Though useful for rapid surveillance and evaluation of success against government targets, two measures currently used by Public Health England, the "complete case" early stage percentage and missing stage percentage, give an incomplete picture of changes in early diagnosis in the population. Epidemiological analyses of stage trends are needed in addition to these in order to evaluate progress. Patient records missing stage should be included in surveillance through an imputation approach as done here, or prognostic measures based on estimated stage or survival could be used [16]. This recommendation is based on analysis of patients in England, but is likely to be equally relevant to the *Detect Cancer Early* programme in Scotland [40], and other stage surveillance programmes internationally.

Our findings are based on a gold-standard approach to reduce bias when stage data are missing but auxiliary information is available. They concord with improvements in survival during this period [41]. However, further research is needed to better understand the mechanisms by which stage is missing and to optimise imputation models; to replicate our finding of a very large increase in colorectal cancer early diagnosis; and to understand the drivers of improvement.

Our analysis concludes in 2013. It is important that epidemiological analyses of trends in early stage diagnosis after this period are done in order to understand the impact of health service reform and financial austerity on cancer control.

REFERENCES

[1] Office for National Statistics, Cancer survival by stage at diagnosis for England (experimental statistics): Adults diagnosed 2012, 2013 and 2014 and followed up to 2015, London, UK, 2016.

[2] C. Maringe, S. Walters, B. Rachet, J. Butler, T. Fields, P.J. Finan, R. Maxwell, B. Nedrebø, L. Påhlman, A. Sjövall, A. Spigelman, G. Engholm, A. Gavin, M.L. Gjerstorff, J. Hatcher, T. Borge Johannesen, E.J. Morris, C.E. McGahan, E. Tracey, D. Turner, M.A. Richards, M.P. Coleman, ICBP Module 1 Working Group, Stage at diagnosis and colorectal cancer survival in six high-income countries: a population-based study of patients diagnosed during 2000-7, Acta Oncologica 52 (2013) 919-932.

[3] Independent Cancer Taskforce, Achieving World-Class Cancer Outcomes: A Strategy for England 2015-2020, London, UK, 2015.

[4] Cancer Australia, Cancer Australia Strategic Plan 2014-2019, Surry Hills, New South Wales, 2014, p. 40.

[5] Cancer Care Ontario, Ontario Cancer Plan 2008-2011, Toronto, Canada, 2008.

[6] Danish National Board of Health, The National Cancer Plan Summary, Copenhagen, Denmark, 2000.

[7] Department of Health, The NHS Cancer Plan, London, UK, 2000.

[8] Department of Health, Improving Outcomes: A Strategy for Cancer, London, UK, 2011.

[9] Scottish Government, Detect Cancer Early (programme summary).
https://www2.gov.scot/Topics/Health/Services/Cancer/Detect-Cancer-Early.

[10] M.A. Richards, The size of the prize for earlier diagnosis of cancer in England, British journal of cancer 101 Suppl 2 (2009) S125-9.

[11] Department of Health, Cancer Reform Strategy, London, UK, 2007.

[12] R. Roope, New NICE GP guidelines have huge ambition and potential, 2015.
https://scienceblog.cancerresearchuk.org/2015/06/23/new-nice-gp-guidelines-have-huge-ambition-and-potential/?_ga=2.101901505.1789612612.1564331422-821341162.1564331422.

[13] R.F.A. Logan, J. Patnick, C. Nickerson, L. Coleman, M.D. Rutter, C. von Wagner, Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests, Gut 61(10) (2011) 1439-46.

[14] Public Health England, Be Clear On Cancer - Overview, UK, 2019.

[15] Public Health England, CancerData dashboard, 2019.
https://www.cancerdata.nhs.uk/dashboard#?tab=Overview.

[16] P. Muller, S. Walters, M.P. Coleman, L. Woods, Which indicators of early cancer diagnosis from population-based data sources are associated with short-term mortality and survival?, Cancer epidemiology 56 (2018) 161-170.

[17] C. Di Girolamo, S. Walters, S. Benitez Majano, B. Rachet, M.P. Coleman, E.N. Njagi, M. Morris, Characteristics of patients with missing information on stage: a population-based study of patients diagnosed with colon, lung or breast cancer in England in 2013, BMC cancer 18 (2018) 492.

[18] M.E. Barclay, G.A. Abel, L. Elliss-Brookes, D.C. Greenberg, G. Lyratzopoulos, The influence of patient case mix on public health area statistics for cancer stage at diagnosis: a cross-sectional study, European Journal of Public Health (2019).

[19] Q. Luo, S. Egger, X.Q. Yu, D.P. Smith, D.L. O'Connell, Validity of using multiple imputation for "unknown" stage at diagnosis in population-based cancer registry data, PLOS ONE 12(6) (2017) e0180033.

[20] M. Falcaro, U. Nur, B. Rachet, J.R. Carpenter, Estimating excess hazard ratios and net survival when covariate data are missing: strategies for multiple imputation, Epidemiology (Cambridge, Mass.) 26(3) (2015) 421-8.

[21] M.E. Barclay, G. Lyratzopoulos, D.C. Greenberg, G.A. Abel, Missing data and chance variation in public reporting of cancer stage at diagnosis: Cross-sectional analysis of population-based data in England, Cancer epidemiology 52(Supplement C) (2018) 28-42.

[22] World Health Organisation, International statistical classification of diseases and related health problems - 10th revision (ICD-10), Geneva, Switzerland, 2011.

[23] Royal College of Physicians, National Lung Cancer Audit, 2015. https://www.rcplondon.ac.uk/projects/national-lung-cancer-audit. (Accessed 03/11 2016).

[24] Healthcare Quality Improvement Partnership, National Bowel Cancer Audit report 2014, Leeds, 2014.

[25] L. Elliss-Brookes, S. McPhail, A. Ives, M. Greenslade, J. Shelton, S. Hiom, M. Richards, Routes to diagnosis for cancer - determining the patient journey using multiple routine data sets, British journal of cancer 107(8) (2012) 1220-6.

[26] S. Benitez Majano, H. Fowler, C. Maringe, C. Di Girolamo, B. Rachet, Deriving stage at diagnosis from multiple population-based sources: colorectal and lung cancer in England, British journal of cancer 115(3) (2016) 391-400.

[27] C. Maringe, H. Fowler, B. Rachet, M.A. Luque-Fernandez, Reproducibility, reliability and validity of population-based administrative health data for the assessment of cancer non-related comorbidities, PLoS One 12(3) (2017) e0172814.
[28] J.R. Carpenter, M.G. Kenward, Multiple imputation and its application, Wiley, Chichester, 2013.
[29] M. Quartagno, J. Carpenter, jomo: A package for Multilevel Joint Modelling Multiple Imputation, 2017.
[30] J. Gurney, D. Sarfati, J. Stanley, The impact of patient comorbidity on cancer stage at diagnosis, British journal of cancer 113(9) (2015) 1375-1380.
[31] D.H. Brewster, C.S. Thomson, D.J. Hole, R.J. Black, P.L. Stroner, C.R. Gillis, Relation between socioeconomic status and tumour stage in patients with breast, colorectal, ovarian, and lung cancer: results from four national, population based studies, Bmj 322(7290) (2001) 830-1.
[32] S. Benitez Majano, C. Di Girolamo, B. Rachet, C. Maringe, M.G. Guren, B. Glimelius, L.H. Iversen, E.A. Schnell, K. Lundqvist, J. Christensen, M. Morris, M.P. Coleman, S. Walters, Surgical treatment and survival from colorectal cancer in Denmark, England, Norway, and Sweden: a population-based study, The Lancet Oncology 20(1) (2019) 74-87.
[33] J.W. Graham, A.E. Olchowski, T.D. Gilreath, How many imputations are really needed? Some practical clarifications of multiple imputation theory, Prevention science : the official journal of the Society for Prevention Research 8(3) (2007) 206-13.
[34] D.B. Rubin, Multiple imputation for nonresponse in surveys, John Wiley and Sons, New York, 1987.
[35] S. Rabe-Hesketh, A. Skrondal, Multilevel and longitudinal modeling using Stata., 3rd ed., Stata Press2012.
[36] Cancer Research UK, Cancer Services: Reverse, Pause or Progress?, Cancer Research UK, London, 2012.
[37] More patients waiting longer than a week for GP appointments, The Guardian, London, 2017.
[38] I. Blunt, N. Edwards, L. Merry, What's behind the A&E 'crisis'?, Nuffield Trust, London, 2015.
[39] Margaret McCartney, Patients deserve the truth: heatlh screening can do more harm than good, The Guardian, 2014.
[40] Scottish Government, Detect Cancer Early: NHS Scotland performance against Local Delivery Plan (LDP) standards, 2019. https://www.gov.scot/publications/nhsscotland-performance-against-ldp-standards/pages/detect-cancer-early/.
[41] S. Walters, S. Benitez Majano, P. Muller, M.P. Coleman, C. Allemani, J. Butler, M. Peake, M.G. Guren, B. Glimelius, S. Bergstrom, L. Pahlman, B. Rachet, Is England closing the international gap in cancer survival?, British journal of cancer 113(5) (2015) 848-60.

SUPPLEMENTARY APPENDIX

Appendix Table 1 Counts of patients in each diagnosis period by CCG and stage at diagnosis

	Colorectal cancer			NSCLC			Ovarian cancer		
	2008-2009	2010-2011	2012-2013	2008-2009	2010-2011	2012-2013	2008-2009	2010-2011	2012-2013
Count of patients by CCG									
CCG Average	306	316	318	275	288	299	46	47	45
Minimum	36	43	36	62	79	57	3	8	8
Maximum	1,355	1,402	1,295	951	1,044	1,060	216	213	195
Count (%) of patients by stage at diagnosis									
I	2,480 (3.9)	5,970 (9.0)	9,899 (14.9)	5,116 (8.9)	6,879 (11.4)	9,467 (15.1)	1,041 (10.8)	1,472 (14.9)	1,979 (20.6)
II	4,859 (7.6)	10,052 (15.2)	15,074 (22.7)	2,435 (4.2)	4,219 (7.0)	5,109 (8.2)	333 (3.5)	426 (4.3)	576 (6.0)
III	6,214 (9.7)	11,247 (17.0)	15,262 (23.0)	10,893 (19.0)	11,803 (19.6)	12,995 (20.8)	1,397 (14.5)	1,875 (19.0)	3,171 (33.0)
IV	10,112 (15.8)	13,028 (19.7)	15,291 (23.0)	19,301 (33.6)	26,329 (43.8)	29,395 (47.0)	1,356 (14.1)	1,582 (16.0)	2,048 (21.3)
Missing	40,307 (63.0)	25,816 (39.0)	10,900 (16.4)	19,637 (34.2)	10,873 (18.1)	5,597 (8.9)	5,491 (57.1)	4,508 (45.7)	1,821 (19.0)

Appendix Table 2 Characteristics of the patients without a recorded stage at diagnosis

	Colorectal cancer			NSCLC			Ovarian cancer		
	2008-2009	2010-2011	2012-2013	2008-2009	2010-2011	2012-2013	2008-2009	2010-2011	2012-2013
Patients missing stage (%)	40,307 (63.0)	25,816 (39.0)	10,900 (16.4)	19,637 (34.2)	10,873 (18.1)	5,597 (8.9)	5,491 (57.1)	4,508 (45.7)	1,821 (19.0)
Average age	72	73	74	75	75	76	68	69	70
Female (%)	45.2	45.5	47.8	44.7	45.6	47.2	-	-	-
1+ comorbidities (%)	19.2	22.6	29	31.6	36.6	39.9	15	16.8	20.2
Emergency presentation (%)	23.5	24.8	29.6	46.1	48.3	50.9	36.8	37.8	40.7
Deprivation quintile 4/5 (%)	36.1	35.4	34.5	47.8	46.4	45.7	35.5	36.6	34.9

Appendix Table 3 Distribution of case-mix characteristics in each time period (with multiple imputation estimates used for topography and morphology to account for the missing data)

	Colorectal cancer				NSCLC				Ovarian cancer		
	2008-09	2010-11	2012-13		2008-09	2010-11	2012-13		2008-09	2010-11	2012-13
Age											
15-39 (%)	1.5	1.6	2.2		0.5	0.5	0.4		4.8	4.5	4.8
40-49 (%)	3.7	3.7	3.9		2.4	2.3	2.2		8.6	9.1	8.6
50-59 (%)	10.1	10.3	11.2		9.8	9.4	9.2		17.0	16.4	17.2
60-69 (%)	26.4	26.5	24.7		25.2	25.7	26.3		26.2	26.6	27.1
70-79 (%)	31.3	31.0	30.5		34.3	33.7	34.0		24.8	24.7	24.6
80-99 (%)	27.1	26.8	27.4		27.8	28.4	27.9		18.6	18.7	17.6
Sex											
Male (%)	55.7	56.0	56.3		56.5	55.8	54.7				
Female (%)	44.3	44.0	43.7		43.5	44.2	45.3				
Charlson comorbidity Index											
0 (%)	81.9	80.1	77.7		71.7	68.6	65.9		87.0	85.5	84.4
>0 (%)	18.1	19.9	22.3		28.3	31.4	34.1		13.0	14.5	15.6
Topography											
Colon (%)	64.3	64.6	65.2	Main bronchus	9.0	7.6	6.1	Ovary	97.7	97.2	95.8
Rectum (%)	35.7	35.4	34.8	Lobe	91.0	92.4	93.9	Fallopian tube	2.3	2.8	4.2
Morphology											
Carcinoma (%)	90.2	89.9	88.2	Carcinoma	49.5	57.1	59.3	Type I epithelial	19.4	18.9	18.9
Non-carcinoma (%)	9.8	10.1	11.8	Non-carcinoma	50.5	42.9	40.7	Type II epithelial	77.2	78.0	78.2
								Non-epithelial	3.4	3.1	2.9

Appendix Table 4 Multi-level logistic regression results: Odds ratios (and 95% CIs) for association between patient factors and probability of diagnosis at stages I or II

	Colorectal cancer		NSCLC		Ovarian cancer	
Diagnosis period						
2008-2009	1.00		1.00		1.00	
2010-2011	1.33 (1.25, 1.40)		1.10 (1.05, 1.16)		1.14 (1.02, 1.28)	
2012-2012	1.71 (1.62, 1.80)		1.26 (1.20, 1.32)		1.18 (1.05, 1.31)	
Age						
15-39	1.00		1.00		1.00	
40-49	0.90 (0.82, 0.98)		0.36 (0.30, 0.43)		0.74 (0.58, 0.93)	
50-59	1.02 (0.95, 1.10)		0.35 (0.30, 0.41)		0.61 (0.50, 0.75)	
60-69	1.29 (1.20, 1.39)		0.42 (0.35, 0.49)		0.45 (0.37, 0.55)	
70-79	1.30 (1.21, 1.39)		0.43 (0.36, 0.50)		0.36 (0.29, 0.44)	
80-99	1.05 (0.97, 1.13)		0.40 (0.34, 0.47)		0.27 (0.22, 0.34)	
Sex						
Male	1.00		1.00		NA	
Female	1.00 (0.98, 1.02)		1.19 (1.19, 1.19)			
Charlson comorbidity Index						
0	1.00		1.00		1.00	
1	1.06 (1.02, 1.09)		1.55 (1.50, 1.60)		0.93 (0.81, 1.07)	
2	0.97 (0.93, 1.01)		1.52 (1.45, 1.59)		1.04 (0.87, 1.23)	
3+	1.06 (1.02, 1.11)		1.55 (1.48, 1.62)		0.69 (0.52, 0.91)	
Topography*						
Colon	1.00	Lobe	1.00		Ovary	1.00
Rectum	1.12 (1.10, 1.14)	Main Bronchus	0.25 (0.23, 0.28)		Fallopian tube	3.37 (2.84, 4.01)
Morphology						
Carcinoma	1.00	Carcinoma	1.00		Type I epithelial	1.00
Non-carcinoma	2.10 (2.03, 2.16)	Non-carcinoma	0.60 (0.58, 0.62)		Type II epithelial	0.07 (0.07, 0.08)
					Non-epithelial	0.58 (0.46, 0.74)

* Cancer sites are defined by the following ICD codes: Colon=C18; Rectum=C19,C20,C21.8; Main Bronchus=C34.0; Lobe=C34.1,C34.2,C34.3; Ovary=C56; Fallopian tube=C57

Appendix Table 5 Counts of patient's route to diagnosis, cancer, period of diagnosis, 2008-2013

	Colorectal cancer			NSCLC			Ovarian cancer		
Route to diagnosis (%)	2008-09	2010-11	2012-13	2008-09	2010-11	2012-13	2008-09	2010-11	2012-13
Emergency presentation	14,743 (23.1)	14,563 (22.0)	15,325 (23.1)	21,230 (37.0)	21,284 (35.4)	21,617 (34.6)	3,133 (32.6)	3,030 (30.7)	2,845 (29.7)
GP referral	15,343 (24.0)	15,120 (22.9)	15,297 (23.0)	12,135 (21.2)	12,326 (20.5)	13,446 (21.5)	2,173 (22.6)	2,077 (21.1)	2,069 (21.6)
Inpatient Elective	2,841 (4.4)	2,322 (3.5)	2,174 (3.3)	937 (1.6)	894 (1.5)	893 (1.4)	161 (1.7)	114 (1.2)	91 (1.0)
Other outpatient	5,135 (8.0)	4,588 (6.9)	4,304 (6.5)	6,050 (10.5)	6,258 (10.4)	7,202 (11.5)	1,156 (12.0)	1,104 (11.2)	912 (9.5)
Screening	3,738 (5.8)	6,415 (9.7)	6,404 (9.6)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
TWW	16,502 (25.8)	17,981 (27.2)	19,766 (29.8)	13,688 (23.9)	15,340 (25.5)	17,619 (28.2)	2,316 (24.1)	2,874 (29.1)	3,410 (35.5)
Unknown or missing	5,670 (8.9)	5,124 (7.8)	3,156 (4.8)	3,342 (5.8)	4,001 (6.7)	1,786 (2.9)	679 (7.1)	664 (6.7)	268 (2.8)

Appendix: List of OPCS codes used to determine receipt of major surgical treatment. This list was produced by staff in the Cancer Survival Group at LSHTM and kindly provided to the authors by Helen Fowler.

Code	Description	Code	Description	Code	Description
Colorectal cancer		**Colorectal cancer (continued)**		**Ovarian cancer**	
H011	Emergency excision of abnormal appendix and drainage HFQ	H115	Colectomy and exteriorisation of bowel (CODE COLOSTOMY SEPERATELY)	Q071	Radical Hysterectomy (removes uterus + cervix + vagina). Wertheims hysterectomy
H012	Emergency excision of abnormal appendix NEC	H118	Other excision of colon, other specified	Q072	Abdominal Hysterectomy and excision of periuterine tissue NEC.Radical Hysterectomy
H013	Emergency excision of normal appendix	H119	Hemicolectomy NEC; Colectomy NEC, Other excision of colon, unspecified;	Q074	TAH, Panhysterectomy, hysterectomy NEC (removes uterus + cervix). Total abdominal hysterectomy NEC
H018	Other specified emergency excision of appendix	H121	Excision of diverticulum of colon	Q221	Bilateral salpingoophorectomy
H019	Emergency appendicectomy NEC, unspecified	H122	Polypectomy NEC, Excision of lesion NEC	Q223	Bilateral oophorectomy, excision of gonads
H021	Interval appendicectomy	H123	Destruction of lesion of colon NEC	Q231	Unilateral salpingoophorectomy NEC
H022	Planned delayed appendicectomy NEC	H128	Other specified extirpation of lesion of colon	Q232	Salpingoophorectomy of remaining solitary fallopian tube and ovary
H023	Prophylactic appendicectomy NEC	H129	Unspecified extirpation of lesion of colon	Q235	Unilateral oophorectomy NEC
H024	Incidental appendicectomy	H291	Subtotal excision of colon and rectum and creation of colonic pouch and anastomosis of colon to anus	Q241	Salpingoophorectomy NEC
H028	Other specified other excision of appendix	H292	Subtotal excision of colon and rectum and creation of colonic pouch NEC	Q243	Oophorectomy NEC
H029	Appendicectomy NEC, unspecified;	H293	Subtotal excision of colon and creation of colonic pouch and anastomosis of colon to rectum	Q431	Excision of wedge of ovary
H041	Proctocolectomy NEC, Panproctocolectomy and Ileostomy	H294	Subtotal excision of colon and creation of colonic pouch NEC	Q432	Excision of lesion of ovary - cystectomy
H042	Panproctocolectomy and anastomosis of ileum to anus and creation of pouch HFQ	H298	Subtotal excision of colon, Other specified	Q433	Marsupialisation of lesion of ovary
H043	Panproctocolectomy and anastomosis of ileum to anus NEC	H299	Subtotal excision of colon, Unspecified	Q439	Unspecified partial excision of ovary
H048	Other specified total excision of colon and rectum	H331	Abdominoperineal excision of rectum and end colostomy; APR; SCAPER	Q441	Open cauterisation of lesion of ovary
H049	Panproctocolectomy NEC, Total excision of colon and rectum, unspecified-	H332	Proctectomy and anastomosis of colon to anus	Q449	Unspecified open destruction of lesion of ovary
H051	Total colectomy and anastomosis of ileum to rectum	H333	Anterior resection of rectum and anastomosis of colon to rectum using staples	Q478	Other specified other open operations on ovary
H052	Total colectomy and ileostomy and creation of rectal fistula HFQ	H334	Anterior resection of rectum and anastomosis NEC	Q479	Unspecified other open operations on ovary
H053	Total colectomy and ileostomy NEC	H335	Hartmann procedure, Rectosigmoidectomy and closure of rectal stump and exteriorisation of bowel (CODE COLOSTOMY SEPERATELY)	Q498	Other specified therapeutic endoscopic operations on ovary
H058	Total excision of colon, other specified	H336	Anterior resection of rectum and exteriorisation, (CODE COLOSTOMY SEPARATELY)	Q499	Unspecified therapeutic endoscopic operations on ovary
H059	Total excision of colon, Unspecified	H337	Perineal resection of rectum HFQ	Q518	Other operations on ovary, Other specified
H061	Extended right hemicolectomy and end to end anastomosis	H338	Anterior Resection of Rectum NEC, Rectosigmoidectomy and anastomosis of colon to rectum Excision of rectum, other specified	Q519	Other operations on ovary, Unspecified
H062	Extended right hemicolectomy and anastomosis of ileum to colon	H339	Rectosigmoidectomy NEC, Excision of rectum, unspecified;	T361	Omentectomy – Complete
H063	Extended right hemicolectomy and anastomosis NEC	H341	Open excision of lesion of rectum: Open removal of polyp; Yorke Mason	T865	Para-aortic lymph node sampling
H064	Extended right hemicolectomy and ileostomy HFQ	H342	Open cauterisation of lesion of rectum, Diathermy	T868	Pelvic lymph node sampling
H068	Other specified extended excision of right hemicolon	H343	Open cryotherapy to lesion of rectum	T875	Para-aortic lymphadenectomy
H069	Extended excision of Right hemicolon, unspecified, excision of Right colon and surrounding tissue	H344	Open laser destruction of lesion of rectum	T878 + Z941	Bilateral pelvic lymphadenectomy
H071	Right hemicolectomy and end to end anastomosis of ileum to colon, Ileocaecal resection	H345	Open destruction of lesion of rectum NEC	T878 + Z942	Right pelvic lymphadenectomy
H072	Right hemicolectomy and side to side anastomosis of ileum to transverse colon,	H348	Open removal of lesion of rectum, other specified	T878 + Z943	Left pelvic lymphadenectomy
H073	Right hemicolectomy and anastomosis NEC	H349	Open removal of lesion of rectum, unspecified	X141	Total exenteration of pelvis
H074	Right hemicolectomy and ileostomy HFQ	H401	Trans-sphincteric excision of mucosa of rectum	X142	Anterior exenteration of pelvis
H078	Other specified other excision of right hemicolon	H402	Trans-sphincteric excision of lesion of rectum	X143	Posterior exenteration of pelvis
H079	Other excision of right hemicolon, unspecified; Right hemicolectomy NEC	H403	Trans-sphincteric destruction of lesion of rectum	X148	Clearance of Pelvis OS
H081	Transverse colectomy and end to end anastomosis	H404	Trans-sphincteric anastomosis of colon to anus	X149	Clearance of Pelvis unspecified
H082	Transverse colectomy and anastomosis of ileum to colon	H408	Other specified operations on rectum through anal sphincter		
H083	Transverse colectomy and anastomosis NEC	H409	Unspecified operations on rectum through anal sphincter		
H084	Transverse colectomy and ileostomy HFQ	X141	Total exenteration of pelvis		
H085	Transverse colectomy and exteriorisation of bowel NEC (CODE COLOSTOMY SPERATELY)	X142	Anterior exenteration of pelvis		
H088	Other specified excision of transverse colon	X143	Posterior exenteration of pelvis		
H089	Excision of transverse colon, unspecified	X148	Other specified clearance of pelvis		
H091	Left hemicolectomy and end to end anastomosis of colon to rectum	X149	Clearance of pelvis, unspecified		
H092	Left hemicolectomy and end to end anastomosis of colon to colon	**NSCLC**			
H093	Left hemicolectomy and anastomosis NEC	E391	Open excision of lesion of trachea		
H094	Left hemicolectomy and ileostomy HFQ	E398	Other specified partial excision of trachea		
H095	Left hemicolectomy and exteriorisation of bowel NEC (CODE COLOSTOMY SEPERATELY)	E399	Unspecified partial excision of trachea		
H098	Excision of left hemicolon, Other specified	E441	Excision of carina		
H099	Left hemicolectmy NEC, Excision of left hemicolon, Unspecified	E461	Sleeve resection of bronchus and anastomosis HFQ		
H101	Sigmoid colectomy and end to end anastomosis of ileum to rectum	E541	Total pneumonectomy, total removal of lung, Pneumonectomy NEC		
H102	Sigmoid colectomy and anastomosis of colon to rectum	E542	Bilobectomy of lung		
H103	Sigmoid colectomy and anastomosis NEC	E543	Lobectomy of lung		
H104	Sigmoid colectomy and ileostomy HFQ	E544	Excision of segment of lung		
H105	Sigmoid colectomy and exteriorisation of bowel NEC	E545	Partial lobectomy of lung NEC		
H108	Other specified excision of sigmoid colon	E548	Excision of lung, other specified		
H109	Unspecified excision of sigmoid colon	E549	Excision of lung, Unspecified		
H111	Colectomy and end to end anastomosis of colon to colon NEC	E552	Open excision of lesion of lung		
H112	Colectomy and side to side anastomosis of ileum to colon NEC	E559	Open removal of lesion of lung, unspecified		
H113	Colectomy and anastomosis NEC	T013	Excision of lesion of chest wall		
H114	Colectomy and ileostomy NEC	T023	Insertion of prothesis into chest wall NEC		

Supplementary Appendix: multiple imputation

Missing stage was imputed on the basis of auxiliary patient information using joint modelling with the R package *jomo* (subcommand *jomo1rancat*) [29], treating stage as a categorical variable with no ordering and accounting for the multi-level structure of the data.

The modelling assumes that stage is missing randomly conditional on variables that are highly associated with either stage (I to IV) [1, 2, 30-32] or recording of stage [19]: quarter year of diagnosis, cancer registry area, CCG, age, sex, patient's Indices of Multiple deprivation (IMD) income quintile, Charlson comorbidity score [27], tumour topography, tumour morphology, route to diagnosis, receipt of major treatment (yes/no), major treatment admission method (elective/non-elective), time from diagnosis to censoring in days, and vital status at censoring.

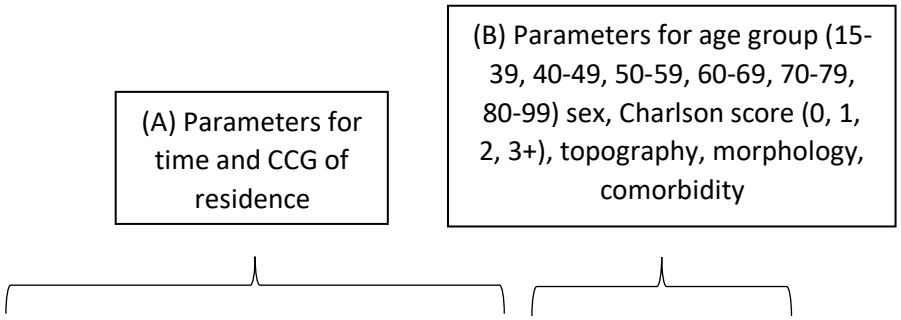
Information on survival time was included in the imputation model as i) the Nelson Aalen cumulative hazard estimate ii) a binary indicator for whether censored or died. Three random effects were included, one for CCG in each of the three time periods (i.e. treated like three independent random intercepts, each estimated from the subset of patients diagnosed in that time period). Cancer registry, deprivation, charlson score (0, 1, 2, 3+), age group (15-39, 40-49, 50-59, 60-69, 70-79, 80-99), topography, morphology, route to diagnosis, and quarter year of diagnosis (Jan-Mar 2008, Apr-Jun 2008, … , Jul-Sep 2013, Oct-Dec 2013) were all included as categorical variables. For colorectal cancer only the sensitivity of the imputation to the addition of interaction terms (including between age and time period, and between cumulative hazard and the binary indicator for death) was evaluated. Their inclusion was not found to materially change the results.

The imputation therefore allowed for differences in the stage distribution between periods as small as three months. Descriptive analyses using the imputation datasets included estimates of the stage distribution in each of these three-month periods, whilst the modelling analysis only compared differences between larger two-year time periods in order to achieve greater power to detect differences between the start and end of 2008-2013.

The number of imputation datasets created was 39, 20, and 41 respectively for colorectal cancer, NSCLC, and ovarian cancer, chosen to reduce the computational burden whilst achieving minimal power reduction compared to using n=100 datasets (an estimated reduction of <1% for all cancers [33]). Parameter estimates (percentages of patients at different stages and regression model parameters) were produced using each dataset and combined using Rubin's rules [34].

Supplementary Appendix: modelling

The analysis models were specified as follows, where i denotes a patient, j their CCG of residence; y_{ij} whether a patient was diagnosed at stage I or II, $\boldsymbol{X_{ij}}$ the vector of patient and tumour covariables, and μ_{jk} the variance associated with CCG effects on early diagnosis in period k:



$$logit\left(\Pr\left(y_{ij}=1\middle|\boldsymbol{X_{ij}},\boldsymbol{U_{jk}}\right)\right)=\left(\beta_1+\mu_{j1}\right)+T_2(\beta_2+\mu_{j2})+T_3(\beta_3+\mu_{j3})+\beta_4x_{4i}+\cdots+\beta_nx_{ni}$$

$$\mu_{jk}\sim N(0,\varphi_k^2)$$

$$T_2=\begin{cases}1\ if\ diagnosis\ date\ is\ in\ 2010\ or\ 2011;\\ 0\ \ otherwise\end{cases}$$

$$T_3=\begin{cases}1\ if\ diagnosis\ date\ is\ in\ 2012\ or\ 2013;\\ 0\ otherwise\end{cases}$$

The models were run once with just the (A) parameters (no case-mix adjustment) then run with both the (A) and (B) parameters (case-mix adjustment), and results compared. The (B) parameters were specified in the analysis model as categorical variables, identical to their specification in the imputation model.

The odds ratios exp(β_2) and exp(β_3) provide information on changes in the probability of diagnosis at stage I/II at the national level. The total estimated between-CCG variance at each time period ($\mu_{j1},\mu_{j2},\mu_{j3}$), and the derived odds ratios for 2.5th , 25th, 75th, and 97.5th percentiles of CCG effects, provide information on geographic inequalities.

The command meqrlogit was used to fit the regression models in STATA, as this command typically has fewer problems with convergence compared the alternative melogit when estimated random effects are small [4]. The model estimates distinct CCG effects for each period as separate levels in the model (with no overlap in records used to estimate effects at different levels), in an analogous specification to the multilevel heterogeneity models used in longitudinal studies which allow variability in growth curves between boys and girls (e.g. those described in Rabe-Hesketh and Skrondal, *Multilevel and Longitudinal Modelling Using Stata Vol 1: Continuous Responses* (page 362)).

ACKNOWLEDGEMENTS

PM and SW received funding from a Cancer Research UK (CRUK) Early Diagnosis Policy Research grant to the Cancer Policy Programme at the London School of Hygiene & Tropical Medicine (award number C7923/A18348), and LW is funded by a Cancer Research UK Population Research Postdoctoral Fellowship (award number C23409/A11415). The authors would like to thank attendees of the CRUK Early Diagnosis Indicators Workshop (2017) for insightful discussion regarding measurement of early diagnosis. We owe thanks to Chiara Di Girolamo, Melanie Morris, Michel Coleman, Bernard Rachet, and Sara Benitez Majano for informative discussions and support; to Matteo Quartagno for assistance and advice on the use of *jomo*; and to Helen Fowler for providing the algorithm to determine receipt of major treatment. We additionally thank Aurelien Belot and Edmund Njeru Njagi for advice on the specification of the multilevel models and analysis strategy, Adrian Turculet for performing the data cleaning and linkage, and Public Health England for providing the data used in the analysis.