Cancer Epidemiology 52 (2018) 28–42

Contents lists available at ScienceDirect

Cancer Epidemiology

journal homepage: www.elsevier.com/locate/canep

Original Research Article

Missing data and chance variation in public reporting of cancer stage at diagnosis: Cross-sectional analysis of population-based data in England

Matthew E. Barclay[a,b], Georgios Lyratzopoulos[a,b,c,*], David C. Greenberg[a,b], Gary A. Abel[d]

[a] Cambridge Centre for Health Services Research, Department of Public Health and Primary Care, Forvie Site, Robinson Way, Cambridge, CB2 0SR, United Kingdom
[b] National Cancer Registration and Analysis Service, Public Health England, Victoria House, Capital Park, Fulbourn, Cambridge, CB21 5XA, United Kingdom
[c] Epidemiology of Cancer Healthcare and Outcomes (ECHO) Research Group, Department of Behavioural Science and Health, University College London, WC1E 7HB, United Kingdom
[d] University of Exeter Medical School (Primary Care), Smeall Building, St Luke's Campus, Exeter, EX1 2LU, United Kingdom

ABSTRACT

Background: The percentage of cancer patients diagnosed at an early stage is reported publicly for geographically-defined populations corresponding to healthcare commissioning organisations in England, and linked to pay-for-performance targets. Given that stage is incompletely recorded, we investigated the extent to which this indicator reflects underlying organisational differences rather than differences in stage completeness and chance variation.

Methods: We used population-based data on patients diagnosed with one of ten cancer sites in 2013 (bladder, breast, colorectal, endometrial, lung, ovarian, prostate, renal, NHL, and melanoma). We assessed the degree of bias in CCG (Clinical Commissioning Group) indicators introduced by missing-is-late and complete-case specifications compared with an imputed 'gold standard'. We estimated the Spearman-Brown (organisation-level) reliability of the complete-case specification. We assessed probable misclassification rates against current pay-for-performance targets.

Results: Under the missing-is-late approach, bias in estimated CCG percentage of tumours diagnosed at an early stage ranged from −2 to −30 percentage points, while bias under the complete-case approach ranged from −2 to +7 percentage points. Using an annual reporting period, indicators based on the least biased complete-case approach would have poor reliability, misclassifying 27/209 (13%) CCGs against a pay-for-performance target in current use; only half (53%) of CCGs apparently exceeding the target would be correctly classified in terms of their underlying performance.

Conclusions: Current public reporting schemes for cancer stage at diagnosis in England should use a complete-case specification (i.e. the number of staged cases forming the denominator) and be based on three-year reporting periods. Early stage indicators for the studied geographies should not be used in pay-for-performance schemes.

1. Introduction

The percentage of cancer patients diagnosed at an 'early stage' (i.e. TNM stages 1–2) has been routinely reported for National Health Service commissioning organisations (Clinical Commissioning Groups, CCGs) since 2014 [1], following recommendations in the 2011 national cancer strategy for England [2]. Recently, this indicator has been adopted into a pay-for-performance scheme for CCGs [3]. Typical CCGs meeting the relevant targets in a given year would receive a financial incentive of £250,000. The aim of these public reporting and pay-for-performance schemes is to promote diagnosis of cancer at an earlier stage and thereby improve outcomes for patients across England. We further summarise this policy context and the technical aspects of the indicator in Box 1.

Indicators used for comparing the performance of healthcare organisations should, among other considerations, be both valid and reliable. Valid indicators truly measure the intended construct of interest, while reliability indicates the precision by which the construct is measured. The validity of performance indicators based on routinely-collected healthcare data may be undermined by missing information [4,5]. Low reliability, where measures are not precise enough to distinguish organisational performance, is a prevailing concern when person-level measures are aggregated into organisation-level scores [6–9]. Frequently, indicators are published and used in pay-for-

* Corresponding author at: 1-19 Torrington Place, London, WC1E 7HB, United Kingdom.
E-mail address: y.lyratzopoulos@ucl.ac.uk (G. Lyratzopoulos).

https://doi.org/10.1016/j.canep.2017.11.005
Received 10 August 2017; Received in revised form 9 November 2017; Accepted 11 November 2017

Box 1
Early stage at diagnosis indicator

In the English National Health Service (NHS), the planning, funding and monitoring of healthcare delivery is the responsibility of 'healthcare commissioning' organisations currently known as Clinical Commissioning Groups. These are responsible for geographically-defined populations. There are about 200 Clinical Commissioning Groups across England, covering an average general population of about 250,000 residents. To support and promote their planning, funding and monitoring function, high level performance indicators for Clinical Commissioning Groups are published annually, across different disease areas, including cancer. In England, a nationwide population-based cancer registration system has been in existence since 1971. In recent years, the modernisation of cancer registration systems has enabled the capturing of information on stage at diagnosis for a high proportion of patients. This has allowed for the introduction of the 'early diagnosis' indicator for Clinical Commissioning Groups studied in our paper. This indicator relates to the stage at diagnosis of 10 different solid tumour sites, and can be met by a Clinical Commissioning Group if either of the following criteria apply: a) 60% or greater proportion of all registered cases with relevant tumours are known to have been diagnosed in TNM stages 1 or 2; or b) there has been a 4% or greater absolute increase within a year in the proportion of all registered cases with relevant tumours known to have been diagnosed in TNM stages 1 or 2.

performance schemes without these concerns being examined or addressed.

The validity and reliability of the early stage indicator for CCGs as currently specified have not been evaluated. Currently, patients with cancer with no recorded stage are treated as though they had late stage cancer, but an alternate specification excluding such patients may be more appropriate. Furthermore, the annual reporting period may be either unnecessarily long or too short to allow for reliable estimation of performance. In this article, we demonstrate how appropriate statistical techniques may be used to examine the properties of this indicator, and identify specific improvements to reduce bias and improve its reliability.

2. Materials and methods

2.1. Data sources

We used population-based data (Public Health England National Cancer Registration and Analysis Service) on TNM stage at diagnosis and other patient and tumour characteristics of patients diagnosed during 2013 with 10 common cancers: bladder (ICD10 C67); female breast (C50); colorectal (C18–C20); endometrial (C54); lung (C33–C34); ovarian (C56–C574); prostate (C61); and renal (C64) cancers; melanoma (C43); and non-Hodgkin lymphoma (C82–C85). The choice of cancer sites and definition of early stage (TNM stages 1–2) reflected those included in the Public Health Outcomes Framework and the CCG Quality Premium; for both, data relating to patients diagnosed in 2013 was reported in 2014 [1,3,10,11].

2.2. Analysis

2.2.1. Examining bias arising from missing data in indicators of early stage at diagnosis

In the study year (2013) stage completeness across all 10 cancer sites was 82%, ranging from 71% to 91% for renal and endometrial cancer, respectively. We used multiple imputation by chained equations (MI) to produce a 'best estimate' early stage indicator, which we treated as the gold standard. Separately by cancer site, a binary early stage indicator for each patient was imputed with logistic regression [12], using auxiliary information on important patient and tumour characteristics associated with stage at diagnosis including patient age, sex, tumour grade (partially missing), CCG, and survival time from diagnosis [13–16]. The MI indicator for each CCG was estimated as the mean percentage of tumours diagnosed at early stage over ten imputed datasets [17]. Appendix A contains further details of the imputation model.

We judged *a priori* that indicators based on the MI approach were not suitable for routine use in public reporting, primarily due to the need for follow-up periods to have elapsed to obtain survival information for use in imputation models, as well as the computational complexity and lack of end-user familiarity with the underlying statistical methods. Instead simpler approaches would be preferable if they are not associated with a substantial degree of bias. We therefore investigated the degree of bias in CCG scores using two simpler approaches for producing early stage indicators. First, the 'missing-is-late' indicator, where the percentage of all tumours with recorded early stage is estimated assuming that those without recorded stage information are advanced stage tumours. The missing-is-late approach is currently used to produce early stage indicators [1,3,10]. Second, the 'complete-case' indicator, where the percentage of staged tumours diagnosed at early stage is estimated based only on tumours with observed stage. We described the degree of bias in either missing-is-late or complete-case indicators by comparing organisational estimates against the 'best estimate' MI indicator.

2.2.2. Examining the reliability of early stage indicators

The statistical reliability of a measure indicates its reproducibility (consistency) in repeated measurement and its robustness to random measurement error. Here we are concerned with organisation-level (or Spearman-Brown) reliability which represents the extent to which organisational measures (in our case the measured percentages of cancer patients diagnosed in early stage) reflect true differences between organisations, as opposed to random (i.e. chance) variation [7,18–20]. For further details of the calculation of reliability for binary indicators, see Appendix B.

Mixed effects logistic regression models were used to model variation in the percentage of tumours diagnosed at early stage estimated using the complete-case indicator. Our main focus was the composite (all 10 cancers) indicator for CCGs, but we performed similar analyses for each individual cancer site (see Appendix B) and for local government organisations (local authorities) and general practices. These models produced an estimate of the organisation-level variance on the log-odds scale. The estimated variance was used to calculate odds ratios for diagnosis at early rather than late stage comparing the 75th/25th and 95th/5th percentiles of the distribution to illustrate the variation between organisations. Importantly, this was the underlying (true) variation which can be thought of as that which would be seen with very large sample sizes in each organisation, such that the influence of sampling variation would be minimal. This underlying (true) variation will be less than the variation in observed stage metrics as the latter will also include a contribution from chance/sampling [19]. The organisation-level variance on the log-odds scale was also used to calculate the reliability for each indicator based on the number of cases in the study year.

In addition to estimating the reliability of the observed data, model outputs were used to estimate the number of tumours required for each

organisation to have a reliable estimate of the percentage diagnosed at an early stage based on reliability thresholds of 0.7 and 0.9. A reliability of 0.7 or higher is commonly required in public reporting, while a reliability of 0.9 may be required for high-stakes reporting, including pay-for-performance schemes [6,19–21]. Following this we calculated the number of years of data required for reliable reporting at current completeness levels.

To illustrate the direct impact of low reliability, we used the estimated distribution of CCG performance in 2013 to evaluate expected misclassification rates for CCGs on the Quality Premium pay-for-performance thresholds. Estimating the overall CCG misclassification rate (in respect of both targets combined) was not possible using one year of data. We therefore performed two similar simulation processes, one for investigating the 60% criterion and one for the ≥4% change criterion (Appendix D). This proceeded as follows. We started with a list of 209 CCGs and the number of staged tumours (N_i) in 2013 for each CCG. We simulated plausible values of the true performance of each CCG, P_i, using the intercept and random effect from our multi-level model, and mapping back from the logistic to the probability scale. We used the binomial distribution with probability of success P_i and number of trials N_i to generate plausible observed performances for each CCG, given the simulated underlying performance and actual number of staged tumours. For the ≥4% change criterion we simulated two years of data for each CCG with a true, uniform change in performance between the two years, repeated for true changes between −4% and +12%, in steps of 0.1%. We repeated each simulation 10,000 times, examining the sensitivity, specificity, and positive and negative predictive values of both the 60% and ≥4% change criteria. All analyses were carried out in Stata 13 [22].

3. Results

Of 208,112 diagnoses of relevant tumours in 2013, 98,218 (47%) were diagnosed in early stage (1–2), 71,809 (35%) were diagnosed in stages 3–4, and 38,085 (18%) had no recorded stage information (Fig. A1).

3.1. Bias arising from missing data in indicators of early stage at diagnosis

Comparing with the 'best estimate' indicator based on multiply imputed data for CCGs (median 55% early stage, range 45%–66%), the missing-is-late indicator underestimated true performance (median 48%, range 25%–62%), while the complete-case indicator overestimated true performance (median 57%, range 48%–70%).

There was little association between CCG early stage percentages estimated using the indicator based on multiply imputed data and CCG percentages of tumours with missing stage (Fig. 1 panel A). In contrast, when using the missing-is-late specification, we observed a very strong negative relationship between early stage and missing stage percentages (panel B). The complete-case specification did not show a clear association of these two measures (panel C).

Fig. 2 shows the bias associated with the amount of missing stage information compared with the 'best estimate' MI indicator (i.e. where bias is the difference between the 'best estimate' MI indicator and the indicator of interest). Bias in the missing-is-late specification increased in magnitude rapidly as the percentage of tumours with missing stage information increased; median bias across all CCGs was −6% (range −30% to −2%). Using a complete case specification typically produced less biased estimates than the missing-is-late approach across all CCGs, irrespectively of the degree of data completeness. There was a slight positive association between the degree of bias and the percentage of patients with missing stage among CCGs with < 20% missing stage data, and no apparent association among CCGs with > 20% missing stage data. Median bias in the complete-case specification across all CCGs was +2% (range −2% to +7%). Importantly, between-CCG variation in bias due to missing data under the missing-is-late specification (observed range of bias: 28%) was larger than observed variation in early stage on the 'best estimate' (observed range of performance: 21%), while this was not the case for the complete-case indicator (observed range of bias: 9%).

3.2. Reliability of the complete-case indicator

The median reliability of the early stage indicator for CCGs was 0.66 (Table 1), despite strong evidence of variation between CCGs ($p < 0.0001$) and moderate sample sizes for each CCG (median 691 staged tumours). This is below levels of reliability required for use in public reporting or pay-for-performance schemes. The aggregation of three years of data would suffice to produce indicators suitable for public reporting ($\lambda \geq 0.7$) for 90% of CCGs. Indicators for 90% of CCGs with sufficient reliability for use in pay-for-performance schemes ($\lambda \geq 0.9$) would require aggregation of nine years of data. Reliability estimates for individual sites are given in Table C1. For breast and lung cancer, indicators based on three and four years of incident cases respectively would allow for adequate reliability ($\lambda \geq 0.7$) for about 70% of all CCGs, respectively. For other cancer sites, eight (renal cancer) to 35 (endometrial cancer) years would be required. Results for local authorities were similar, while general practice indicators had very low reliability (Table C2).

3.3. Probable misclassification on CCG Quality Premium targets for reporting periods of varying length

Considering the CCG Quality Premium criterion providing financial incentives to CCGs which have 60% of tumours diagnosed at stage 1 or 2 in a single year, based on our simulation (which assumes the complete-case indicator is used), we would expect 40 of the 209 CCGs to appear to meet this 60% target, of which only 21 would have an underlying or long-run performance of 60% or higher, giving a positive predictive value of 53% (Fig. 3). We would expect 29 CCGs to have underlying performance above the 60% target, of which one quarter (eight of 29) would appear to miss the target, giving a sensitivity of 74%. Aggregating multiple years of data reduces expected misclassification rates. Using 2.5 (9) years of data, giving reliability of 0.7 (0.9) for more than 90% of CCGs, increases the expected number of true positives to 23 (25) and reduces the expected number of false positives to 11 (5) (Table C3).

For the 4% year-on-year increase criterion of the CCG Quality Premium, misclassification rates depend on the size of underlying changes in performance expected in the long-term for individual CCGs as well as CCG size. If the CCGs' underlying performance did not change, then with very large sample sizes we would not expect to see any CCGs meet this target. However, based on the actual sample sizes for one year of data we would expect 8% of CCGs to be misclassified as meeting the target if the underlying performance did not change for any CCG (Fig. 4). Furthermore, for a CCG to have an 80% chance of meeting the 4% improvement target they would have to improve their underlying performance such that they increased the percentage of cases diagnosed at early stage by 6.2% (Fig. 4).

4. Discussion

The current specification of the early stage indicator for English commissioning organisations is biased due to organisational variation in stage completeness. For the period we examine, the degree of bias is so large that it dominates the variability in this indicator. An alternative specification of the indicator based only on tumours with recorded stage is substantially less biased. Nonetheless, such complete-case indicators will not be reliable when based on one year of data, and will be associated with a high degree of random misclassification if used in pay-for-performance schemes. Complete-case indicators will be suitable for public reporting if based on three-year reporting periods. Timely

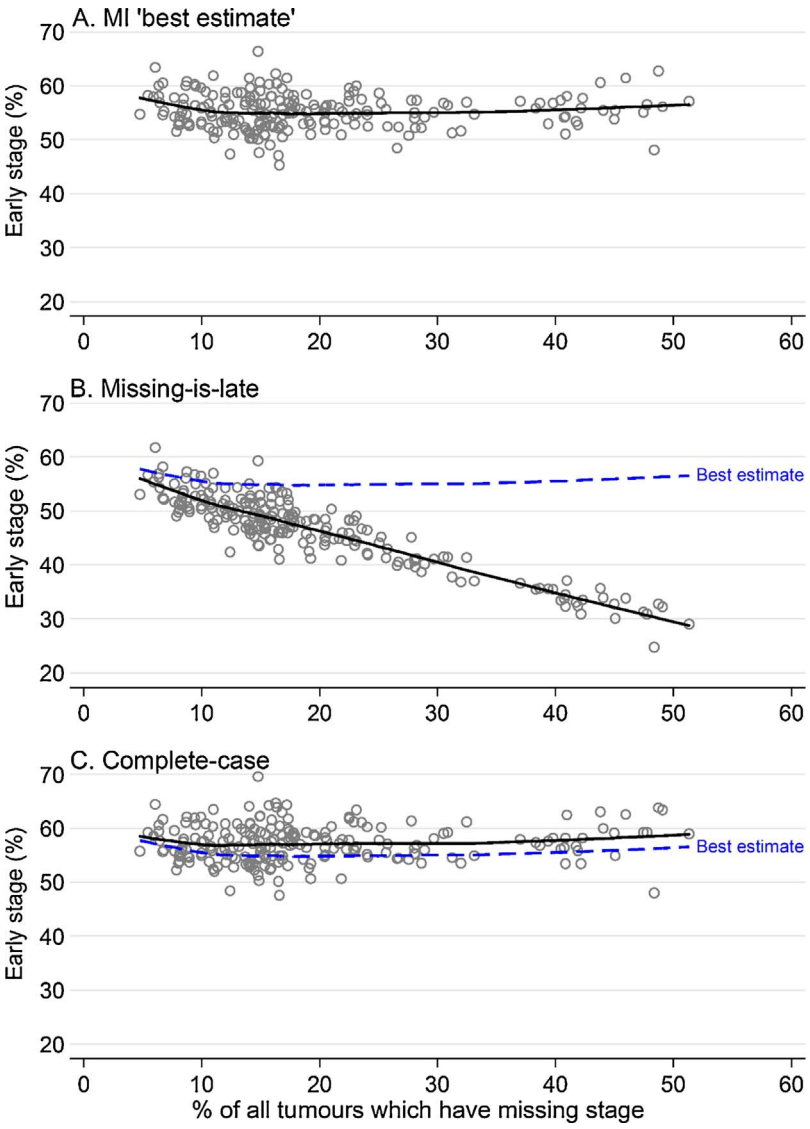

Fig. 1. Observed early-stage percentage calculated using: A. the 'best estimate' multiple imputation approach; B. the missing-is-late approach; and C. the complete-case approach, plotted against the percentage of tumours with no recorded stage information, CCGs, England 2013.

Fig. 2. Bias in scores calculated using the complete-case and missing-is-late approaches when compared with the 'best estimate' MI indicator, plotted against the percentage of tumours with no recorded stage information, CCGs, England 2013.

Table 1
Number of CCGs, staged tumours per CCG, odds ratios over estimated underlying distribution of CCG performance, quartiles of the reliability of the complete-case early stage indicator, and the number of tumours and associated aggregated years of data for 50%, 70%, 90% and 100% of CCGs to have reliability of 0.7 or higher or of 0.9 or higher.

CCGs		209
Number of staged tumours per CCG	Minimum	125
	25th percentile	479
	Median	691
	75th percentile	943
	Maximum	3575
Odds ratio over CCG distribution*	75th/25th percentiles	1.16
	95th/5th percentiles	1.43
Reliability	Minimum	0.26
	25th percentile	0.58
	Median	0.66
	75th percentile	0.73
	Maximum	0.91
Number of tumours per CCG required for reliability 0.7	50% of units	803
	70% of units	812
	90% of units	833
	All units	926
Data years required for reliability 0.7	50% of units	1.2
	70% of units	1.5
	90% of units	2.3
	All units	6.6
Number of tumours per CCG required for reliability 0.9	50% of units	3095
	70% of units	3132
	90% of units	3210
	All units	3570
Data years required for reliability 0.9	50% of units	4.5
	70% of units	5.6
	90% of units	8.7
	All units	25.3

* $p < 0.0001$. Odds ratio calculated directly from the estimated variance of the random intercept from the mixed-effects logistic regression ($\widehat{\sigma^2} = 0.012$) using the appropriate centiles of the standard normal distribution. The 75th/25th percentile odds ratio is calculated as $e^{(1.35\times\sqrt{0.012})}$ and the 95th/5th percentile odds ratio is calculated as $e^{(3.29\times\sqrt{0.012})}$.

early stage indicators suitable for pay-for-performance use are not feasible.

There are no previously published evaluations of the bias or reliability of indicators of cancer stage at diagnosis. Many studies have evaluated the reliability of other performance indicators in healthcare for physicians [7,9], hospitals [23,24], and general practices [8,21] – including for several diagnostic activity indicators reported in the Cancer Services Public Health Profiles [19]. Bias due to missing data is also a common problem for measures based on routinely-collected data, and multiple imputation in particular is commonly used to correct this in cancer registry data [4,25,26].

The key strength of our study is that we use the same English cancer registry data as the early stage indicator, ensuring our results are directly relevant to the current public reporting and pay-for-performance schemes in England. The main weakness is the lack of an objective gold standard for assessing bias in the indicator. Our estimates of bias under different specifications of the indicator are based on comparisons with complete data produced using multiple imputation, as by definition we do not know the stage of tumours with no recorded stage. This approach could itself be biased if the 'missing at random' assumption does not hold, but this is mitigated by the inclusion of important auxiliary information in the imputation process [15,16,25].

As we had no data on successive years, we only estimated true misclassification rates against the 60% early stage target, but as we have shown, CCGs may be additionally misclassified when considering the 4% early stage improvement criterion. The degree of misclassification we report represents an under-estimate.

Among the 10 cancer sites included in the current indicators, some have higher than average proportion of late stage disease (e.g. lung cancer) whereas the opposite is true for other sites (e.g. breast cancer). The indicator does not take into account between-CCG variation in site-specific incidence or in patient demographics, and this may reduce the validity of the current indicator for comparing CCG performance [27,28]. Adjusting for case-mix factors would be expected to reduce variation between organisations, and so a potential case-mix adjusted indicator might be more valid but less reliable. Future studies should establish the degree by which case-mix drives apparent organisational attainment and potential implications for public reporting conventions.

Continuing improvements in stage completeness in English cancer registry data will reduce the size and the variation of bias in the missing-is-late approach. However, bias due to missing stage information under this approach will remain a major problem until all CCGs have very similar stage completeness rates. In our study year the alternative complete-case approach has less bias than the current missing-is-late approach even for CCGs with very high stage completeness, and so would be expected to remain the best option as stage completeness continues to improve.

Aggregating 3 years of data will produce a reliable early stage indicator, suitable for use in public reporting, and we endorse this approach. Pay-for-performance schemes for Clinical Commissioning

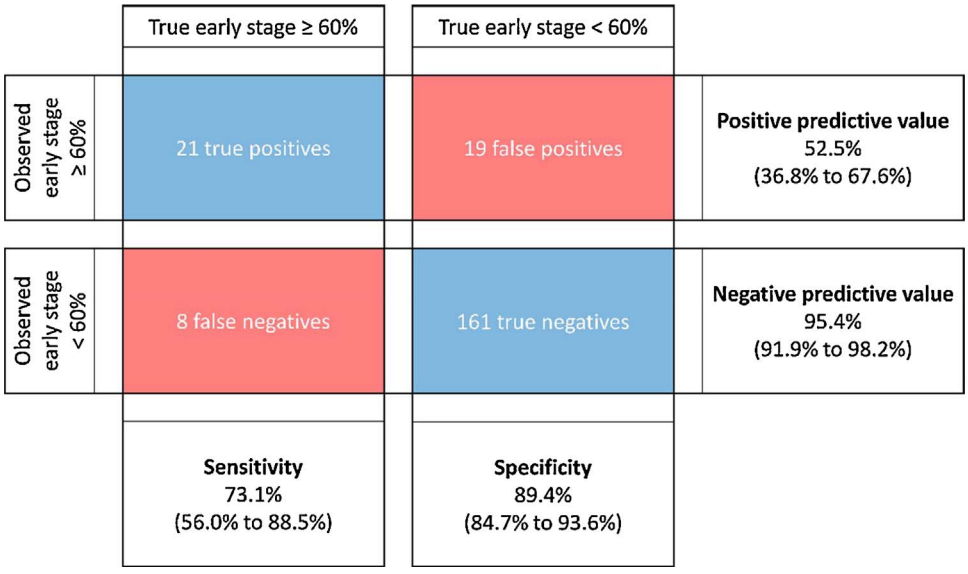

Fig. 3. Estimated number of true positives, false positives, true negatives and false negatives, with associated sensitivity, specificity, positive and negative predictive values (95% confidence intervals), for the 60% early stage target given performance similar to 2013 and tumours counts as in 2013.

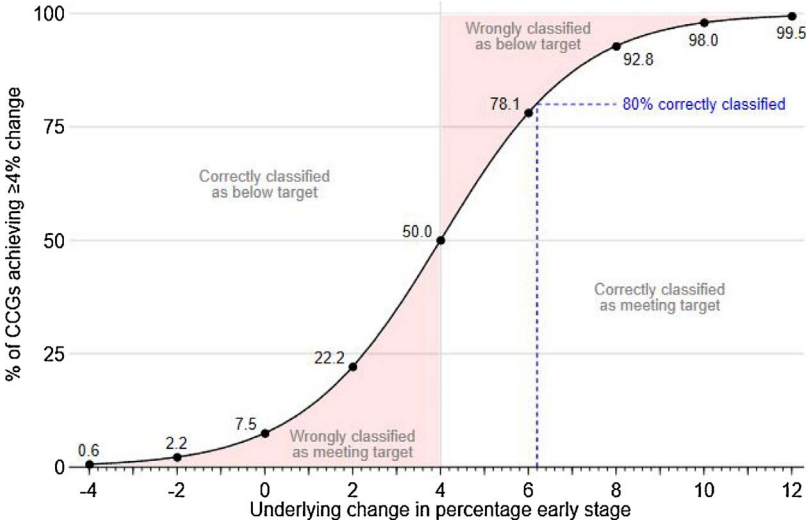

Fig. 4. Expected percentage of CCGs with observed increases in the early stage percentage of 4 percentage points or more, given uniform national changes of between −4 and +12 percentage points. For example, for a typical CCG to have an 80% chance of being classified as achieving a 4%-point increase (blue dashed line), it would need to have an underlying increase of 6.2%-points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Groups should not use the early stage indicator, as sufficiently reliable indicators require more than eight aggregate years of data which greatly limits potential uses. The resulting high levels of misclassification on the indicator when based on a single year mean that many CCGs will receive financial rewards despite their underlying performance being below the pay-for-performance threshold. The opposite is also true, i.e. some CCGs should be rewarded but will not be.

Appropriate process indicators could give more accurate, reliable, and timely information about local diagnostic performance for cancer [29,30], where there are clear links between processes and improved stage at diagnosis, survival, or quality of life. Screening coverage, for example, is a useful measure for breast, colorectal and cervical cancers [31,32]. Other examples might include organisational measures of use of endoscopies or urgent referrals for suspected cancer (otherwise known as 'two-week-wait' referrals), as they are associated with clinical outcomes [33,34]. More generally, there is a need for research to identify diagnostic process indicators which are truly linked to better outcomes for cancer patients, and to identify the organisations best-placed to improve local and national performance.

The development of indicators of cancer diagnosis must involve the evaluation and correction of issues of bias and low reliability. The methods we have highlighted here allow for investigation of these problems, and should form part of the process for the development of such indicators before their introduction into practice. Organisations should not be ranked on severely biased quality measures, and financial incentives should only be linked to highly reliable indicators. Cancer stage indicators should not form part of pay-for-performance schemes for CCGs, and public reporting of the early stage indicator should use three-year reporting periods and be calculated as the percentage of staged tumours diagnosed at an early stage.

Authorship contribution statement

GL and GAA conceived the study. GAA and MB designed the study. MB and DG analysed data. All authors contributed to decisions about data analysis interpretation and drafted the article. All authors approved the final version for submission.

Conflicts of interest

None.

Acknowledgements

GL is funded by a Cancer Research UK Advanced Clinician Scientist Fellowship award (grant number C18081/A18180). We thank Lucy Elliss-Brookes, Sean McPhail, and Sam Johnson for helpful discussions about the design of early stage indicators. Data used in this study were collated, maintained and quality assured by the National Cancer Registration and Analysis Service, which is part of Public Health England (PHE).

Appendix A. Details of multiple imputation of stage for patients with tumours with no recorded stage information

Stage data were 82% complete overall, with at least 70% completeness for each cancer site. However, stage completeness and the distribution of stage at diagnosis where known varied substantially by site (Fig. A1), and stage completeness also varied substantially by CCG (Fig. A2).

Multiple imputation is a recommended method for handling missing stage information in cancer registry data (Table A1). We created a binary stage variable being 'early' (TNM stages 1 or 2) or 'late' (TNM stages 3 or 4) stage. Imputation was performed separately for each cancer site, splitting colorectal cancer into colon and rectal cancer.

We used logistic regression to impute the binary indicator of early stage at diagnosis on:

- CCG of patient at diagnosis
- Region of residence of patient at diagnosis
- Sex of patient
- Interaction between sex and region
- Age group of patient at diagnosis (30–39, then five-year age groups, then 90–99, except for prostate and bladder cancer where the youngest age group was 30–44 due to smaller numbers in this age range)
- Interaction between age group and region
- Deprivation group, fifth of the income domain of IMD 2010
- Interaction between deprivation group and region
- Ethnicity of patient (white or non-white)

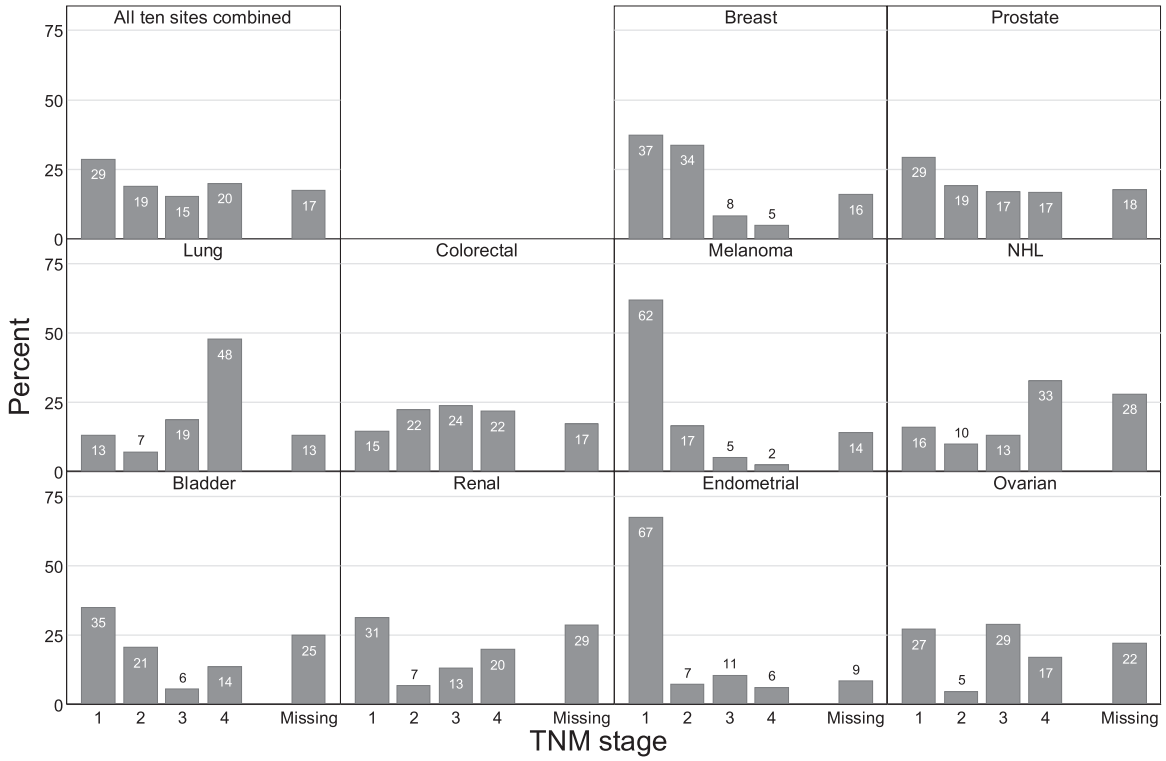

Fig. A1. Percentage of tumours by stage at diagnosis, England 2013.

- Interaction between ethnicity and region
- Nelson-Aalen estimate of cumulative hazard, censored at 365 days after diagnosis
- Indicator of death within 365 days after diagnosis
- Indicator of death within 30 days after diagnosis (not included in imputation of stage for endometrial cancer or melanoma)
- Basis of diagnosis (non-microscopic/microscopic, not included in imputation of stage for endometrial cancer or melanoma)
- Screening detection status (for breast, colon and rectal cancer only)
- Tumour grade (1/2/3/4, not considered for melanoma)

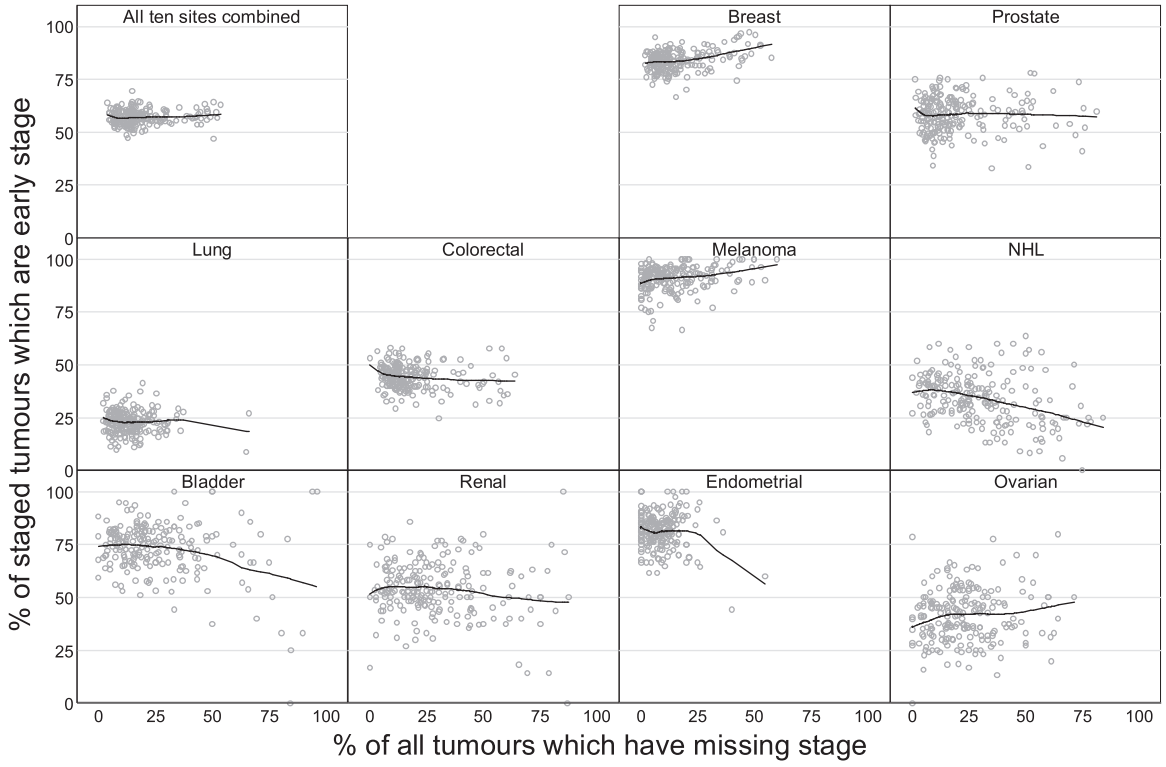

Fig. A2. Percentage of staged tumours which were stage 1 or 2 against percentage of all tumours which were staged, by CCG, England 2013, with LOESS line.

Table A1
Studies evaluating bias introduced by missing data in cancer registry data.

First Author	Year published	Country	Setting	What was imputed	Summary
He Y	2008	US	Regional, California	Indicators of receiving chemotherapy or radiotherapy treatment (colorectal cancer), outcome variables	Correcting under-reporting using internal gold standard.
Krieger N	2008	US	Regional, California	ER-status (breast cancer), outcome variable	Records with missing ER status bias complete-case analysis
Nur U	2010	UK	1 English registry (NWCIS)	Stage (colorectal cancer), covariate	Complete-case analysis is likely to be biased. Indicator methods give spurious precision levels. MI allows inclusion of more information (leading to higher precision than complete-case). MAR assumption probably reasonable, but further research valuable.
Eisemann N	2011	Germany	1 German registry (Schleswig-Holstein)	Simulation (truly MAR), based on real data on breast cancer and melanoma. Stage was imputed by various methods (multinomial logistic; PMM; random forests) with various levels of missing data. Stage was used as outcome (incidence counts) and covariate (survival analysis)	MI is superior to simpler methods for handling missing data. MI using random forests does not perform well (and is associated with model convergence problems).
Howlader N	2012	US	13 SEER registries	ER-status (breast cancer), outcome variable but also for re-use by other researchers	Demonstration with incidence trends. Include the cancer registry in imputation when data from more than one registry are imputed.
Falcaro M	2015	UK	4 English registries	Simulation, based on real data. Stage was imputed by various methods under various (MAR) missingness mechanisms.	Ordinal logistic model is inadequate. Multinomial logistic model works well. Use of Nelson-Aalen estimate of cumulative hazard is recommended.
Andridge R	2016	US	13 SEER registries	ER-status (breast cancer), as outcome variable, using PMM under MAR and various MNAR assumptions	In SEER 1992–2012 breast cancer data, MAR and MNAR approaches give broadly similar results.
Falcaro M	2017	UK	4 English registries	Simulation, based on real data. Stage was imputed by various methods under various (MAR) missingness mechanisms, and the bias in different approaches to imputation was compared.	Can use imputation with non-congenial analysis methods (in this case, Pohar-Perme net survival estimation) to avoid bias associated with “missing indicator” approaches.

ER: Estrogen Receptor.
MAR: Missing At Random.
MNAR: Missing Not At Random.
PMM: Predictive Mean Matching.

We only included patients aged 30–99 at diagnosis. We felt that predictors of stage at diagnosis for patients outside this age range may not reflect those of more typical patients. There were few patients either aged 29 and under (1591 of 208,141, 0.8%) or 100 and older (104 of 208,141, 0.05%), so separate imputation was not feasible.

Screening detection status was applicable for breast, colon and rectal cancers. For melanoma and endometrial cancer, early mortality and non-microscopic diagnosis were both extremely rare and the inclusion of such indicators led to problems with model convergence. For melanoma, tumour grade is both less clinically relevant and had low completeness..

All variables used in imputation models were complete, except for tumour grade. For cancer sites other than melanoma, we used predictive mean matching to impute tumour grade based on the (possibly imputed) binary indicator of early stage at diagnosis and on the other variables and interactions used in imputing stage.

Thus for melanoma we used multiple imputation by logistic regression, while for other sites we used multiple imputation by chained equations. We used ten iterations of the chain as burn-in, having previously checked graphically that doing so led to convergence.

Appendix B. Organisation-level reliability for binary indicators

The statistical reliability of a measure generally indicates its reproducibility (consistency) in repeated measurement and its robustness to random measurement error. Here we are concerned with organisation-level reliability, also termed unit-level reliability where units could be commissioners, providers, or geographical areas. In the context of our study, organisation-level (or Spearman-Brown) reliability represents the extent to which measured percentages of cancer patients diagnosed in early stage reflect true differences between organisations, as opposed to random (i.e. chance) variation. Alternatively, the Spearman-Brown reliability is the proportion of the observed organisational variation not due to chance.

Poor reliability often arises when the typical number of cases per organisation (in a given reporting period) is small. The problem is further exacerbated when small sample sizes are combined with limited variation between organisations. Reliable indicators can help to classify organisational performance and thus enable accurate targeting of improvement efforts and rewards. Conversely, using unreliable indicators can lead to harm through wasting of scarce improvement resources and related opportunity costs. Further, misclassified 'poorly performing' organisations may sustain unfair reputational or financial loss [6,9].

Reliability takes a value between 0 and 1, with higher values denoting more reliable indicators. A reliability of 0.5 indicates that half of the observed variance is due to chance. A reliability of 0.7 is often required for public reporting of indicators, while a reliability of 0.9 may be required for pay-for-performance use [6,20,21]. Organisation-level reliability λ_i for organisation i is defined as

$$\lambda_i = \frac{\text{between-organisation variance}}{\text{between-organisation variance} + \frac{\text{within-organisation variance}}{n_i}}$$

where n_i = achieved sample size for organisation i.

For continuous indicators, this calculation is straightforward [6]. For binary indicators, the within-organisation variance will depend directly on the level of achievement at each individual organisation, according to the binomial distribution [18,20]. It is important to note that as reliability depends on both the organisational sample size and organisational achievement it is specific to each organisation rather than to the indicator as a whole.

We used mixed effects logistic regression models to estimate the organisation-level variance on the log-odds scale ($\widehat{\sigma^2}$). Reliability is then given by

$$\lambda_i = \frac{\widehat{\sigma^2}}{\widehat{\sigma^2} + \frac{1}{\widehat{\pi_i} \times (1 - \widehat{\pi_i}) \times n_i}}$$

where $\widehat{\pi_i}$ is the observed performance of organisation i on the indicator as a proportion [18]. From this formula it can be seen that higher reliability can be achieved by increasing the between-unit variation or by increasing sample sizes. Additionally, for binary indicators, higher reliability is achieved with performance closer to 50%.

Appendix C. Reliability of early stage indicators for the composite indicator for CCGs, local authorities and general practices, with years of data required for indicators suitable for public reporting and pay-for-performance use and associated expected misclassification rates

Table C1
Number of organisations, staged tumours per organisation, odds ratios over estimated underlying distribution of organisational performance, quartiles of the reliability of the complete-case early stage indicator, and the number of tumours and associated aggregated years of data for 50%, 70%, 90% and 100% of organisations to have reliability of 0.7 or higher or of 0.9 or higher, for CCGs, local authorities and general practices.

		CCG	LA	GP
Units with staged tumours		209	326	8075
Staged tumours per unit	Minimum	125	12	1
	25th percentile	479	311	9
	Median	691	427	17
	75th percentile	943	634	30
	Maximum	3575	2992	150
Odds ratio over unit distribution*	75th/25th percentiles	1.16	1.18	1.29
	95th/5th percentiles	1.43	1.49	1.85

(continued on next page)

Table C1 (*continued*)

		CCG	LA	GP
Reliability	Minimum	0.26	0.04	0.01
	25th percentile	0.58	0.53	0.06
	Median	0.66	0.61	0.12
	75th percentile	0.73	0.70	0.20
	Maximum	0.91	0.92	0.56
Tumours required for reliability 0.7	50% of units	803	641	280
	70% of units	812	652	302
	90% of units	833	668	358
	All units	926	784	1546
Data years required for reliability 0.7	50% of units	1.2	1.5	17.1
	70% of units	1.5	2.0	30.2
	90% of units	2.3	2.7	89.5
	All units	6.6	53.7	269.0
Tumours required for reliability 0.9	50% of units	3095	2470	1078
	70% of units	3132	2514	1165
	90% of units	3210	2575	1380
	All units	3570	3022	5963
Data years required for reliability 0.9	50% of units	4.5	5.8	65.8
	70% of units	5.6	7.5	116.4
	90% of units	8.7	10.5	345.0
	All units	25.3	206.8	1,035.0

* $p < 0.0001$ across CCGs, LAs and GPs.

Table C2
National number of diagnoses and median reliability of complete-case composite and site-specific early stage indicators for general practices, CCGs and local authorities, with number of years of data at current completeness levels required for reliable indicators for 70% of organisations.

Cancer site	Tumours			Median reliability			Years of data required for reliable indicators ($\lambda \geq 0.7$) for 70% of organisations		
	Total	Staged	Stage 1–2	GP	CCG	LA	GP	CCG	LA
All ten sites combined	208,141	172,001	98,780	0.12	0.66	0.61	30.2	1.5	2.0
Breast	44,558	37,465	31,635	0.08	0.59	0.42	28.7	2.3	4.8
Prostate	39,934	32,859	19,422	0.05	0.71	0.62	75.3	1.3	1.9
Lung	35,972	31,234	7307	0.02	0.44	0.32	142.0	4.0	7.0
Colorectal	33,477	27,719	12,398	0.04	0.26	0.15	92.0	8.7	17.3
Melanoma	12,245	10,520	9591	0.10	0.26	0.19	34.0	9.7	19.2
NHL	11,222	8080	2916		0.33	0.24		7.0	10.8
Endometrial	7232	6615	5405		0.10	0.06		30.6	50.1
Bladder	8669	6505	4835		0.25	0.14		10.3	22.7
Renal	8368	5970	3202		0.28	0.21		8.1	13.4
Ovarian	6464	5034	2069		0.14	0.14		20.7	21.6

Table C3
Estimated number of true positives, false positives, true negatives and false negatives, with associated sensitivity, specificity, positive predictive value and negative predictive values (95% confidence intervals), for the 60% early stage target given performance similar to 2013 and tumours counts as in 2013 for reporting periods of 1, 2.5 and 9 years.

Reporting period	1 year		2.5 years		9 years	
	Expected value	(95% CI)	Expected value	(95% CI)	Expected value	(95% CI)
True positives	21	(13, 50)	23	(15, 32)	25	(16, 35)
False positives	19	(11, 28)	11	(6, 18)	5	(1, 10)
True negatives	161	(149, 172)	169	(157, 179)	175	(164, 185)
False negatives	8	(3, 14)	6	(2, 11)	4	(1, 8)
Sensitivity	0.73	(0.56, 0.89)	0.80	(0.64, 0.93)	0.88	(0.73, 0.97)
Specificity	0.89	(0.85, 0.94)	0.94	(0.90, 0.97)	0.97	(0.94, 0.99)
Positive predictive value	0.52	(0.37, 0.68)	0.67	(0.50, 0.82)	0.83	(0.68, 0.95)
Negative predictive value	0.95	(0.92, 0.98)	0.97	(0.94, 0.99)	0.68	(0.96, 0.99)

Appendix D. Stata code for estimating expected misclassification rates on the 60% early stage and 4% increase in early stage criteria

```
/*******************************************************************************
* Calculate expected misclassification rates for the CCG Quality Premium
* pay-for-performance criteria
*******************************************************************************/
log using CCGQP_misclassification.smcl, replace

/*******************************************************************************
* Inputs
*******************************************************************************/

/* Seed */
set seed 4782

/* Number of repetitions */
local reps4    10000

/* Temporary files */
tempfile results data inc_4pp_results

/*******************************************************************************


* Input tumour counts and save in tempfile
*******************************************************************************/
clear

qui {
        input cases1

        /* Input CCG-specific staged tumour counts here

                These data are freely available online
                for example, from the NHS Digital indicator portal
                CCG Outcomes Indicator Set indicator 1.17
                "Record of stage at diagnosis"
                Available at: https://indicators.hscic.gov.uk/webview/

                Direct link:
                https://indicators.hscic.gov.uk/webview/velocity?v=2&mode=documentation&submode
                =ddi&study=http%3A%2F%2F192.168.229.22%3A80%2Fobj%2FfStudy%2FP01816

                The numbers from the indicator portal will differ slightly from the
                ones we used in this paper due to late registrations and so forth.

        */

        end
}

/* Produce ID var */
gen id = _n
order id cases1

/* Add number of cases in 2.5 and 9 years */
gen cases25 = ceil(cases1*2.5)
gen cases9  = cases1*9

/* Save */
save `data'


/*******************************************************************************
* Estimate expected misclassification on 60% criterion
*******************************************************************************/
/* set up results file */
clear
set obs 1
gen run = 0

save `results'

/* loop over `reps4' times */
forval i = 1/`reps4' {

        /* Load in data */
        use `data', clear

        /* Assign 'true' proportion early stage
                (Numbers taken from model results, not shown) */
        gen pr = invlogit( .2871914706672214 + rnormal(0,.1088848496729132) )

        /* Loop over each number of years */
        foreach y in 1 25 9 {

                /* Simulate observed number of early stage, proportion of cases */
                gen early`y'  = rbinomial(cases`y',pr)
                gen obs_pr`y' = early`y'/cases`y'

                /* Calculate whether observed values agree with 'true' values */
```

```
                gen byte hit`y'     = pr >= 0.6
                gen byte hit_rec`y' = obs_pr`y' >= 0.6

                /* Calculate PPV, NPV */
                gen byte ppv`y'     =  hit`y' &  hit_rec`y' if  hit_rec`y'
                gen byte npv`y'     = !hit`y' & !hit_rec`y' if !hit_rec`y'

                /* Calculate misclassification rates */
                gen byte misclass`y'    = (obs_pr`y' >= 0.6 & pr <  0.6) ///
                                                | (obs_pr`y' <  0.6 & pr >= 0.6)
                gen byte misclass_up`y' = (obs_pr`y' >= 0.6 & pr <  0.6)
                gen byte misclass_do`y' = (obs_pr`y' <  0.6 & pr >= 0.6)

                /* Calculate sensitivity, specificity */
                gen byte sens`y' = (obs_pr`y' >= 0.6 & pr >= 0.6) if  hit`y'
                gen byte spec`y' = (obs_pr`y' <  0.6 & pr <  0.6) if !hit`y'

                /* Calculate number of true positives, false positives etc */
                gen byte tpv`y' = obs_pr`y' >= 0.6 & pr >= 0.6
                gen byte fpv`y' = obs_pr`y' >= 0.6 & pr  < 0.6
                gen byte tnv`y' = obs_pr`y' <  0.6 & pr  < 0.6
                gen byte fnv`y' = obs_pr`y' <  0.6 & pr >= 0.6

        }

        /* Collapse to get "summary" statistics for this run */
        collapse (sum) misclass* hit* tpv* fpv* tnv* fnv* (mean) ppv* npv* sens* spec*
        gen run = `i'
        order run misclass* misclass_up* misclass_do* hit* ppv* npv* spec*

        /* Add to results file */
        append using `results'
        save `results', replace

        /* Amuse the user */
        if mod(`i',1000) == 0 {
                nois di "`i' simulations"
        }
}

/* Load in results file */
use `results', clear
order run misclass* hit*

/* Remove empty row used to pre-generate results file */
drop if run == 0

/* Reshape to tidier format */
reshape long misclass misclass_up misclass_do hit hit_rec tpv fpv tnv fnv ppv npv sens spec,
i(run) j(years)
replace years = 2.5 if years == 25

sort years run

/* Display summary statistics for each run */
nois display as input _newline "{ul:Likely misclassification counts:}"
nois by years: centile misclass   , c(2.5 50 97.5)

nois display as input _newline "{ul:Likely misclassification counts, up:}"
nois by years: centile misclass_up, c(2.5 50 97.5)

nois display as input _newline "{ul:Likely misclassification counts, down:}"
nois by years: centile misclass_do, c(2.5 50 97.5)

nois display as input _newline "{ul:Hit target:}"
nois by years: centile hit, c(2.5 50 97.5)
```

```
nois display as input _newline "{ul:Recorded as hit target:}"
nois by years: centile hit_rec, c(2.5 50 97.5)

nois display as input _newline "{ul:PPV:}"
nois by years: centile ppv, c(2.5 50 97.5)

nois display as input _newline "{ul:NPV:}"
nois by years: centile npv, c(2.5 50 97.5)

nois display as input _newline "{ul:SENS:}"
nois by years: centile sens, c(2.5 50 97.5)

nois display as input _newline "{ul:SPEC:}"
nois by years: centile spec, c(2.5 50 97.5)

nois display as input _newline "{ul:T+ve / F+ve / T-ve / F-ve}"
nois by years: centile tpv fpv tnv fnv, c(2.5 50 97.5)

/*******************************************************************************
* Estimate expected misclassification on 4% increase criterion
*******************************************************************************/
/* Load in data */
use `data', clear

/* Make `reps4' times bigger
       This allows us to avoid looping 161*9999 times */
expand `reps4'
order id cases1 cases25 cases9
sort  id cases1 cases25 cases9
by id: gen replic = _n

/* Assign 'true' Y0 early stage probabilities */
gen pr = invlogit( .2871914706672214 + rnormal(0,.1088848496729132) )

/* Simulate observed Y0 early stage probabilities for each reporting period */
foreach y in 1 25 9 {
       gen early`y' = rbinomial(cases`y',pr)
       gen obs_pr`y' = early`y'/cases`y'
}

/* Simulate true increases in 0.1% steps */
qui {
       forvalues i=1/161 {

              /* Assign changed Y1 early stage probability */
              gen pr_`i' = pr+((`i'-41)/1000)

              foreach y in 1 25 9 {

                     /* Simulate observed Y1 early stage probability */
                     gen early`y'_`i' = rbinomial(cases`y',pr_`i')
                     gen obs_pr`y'_`i' = early`y'_`i'/cases`y'

                     /* Calculate difference between observed Y0 and Y1 early stage
                            probability */
                     gen dif`y'_`i' = obs_pr`y'_`i'-obs_pr`y'

                     /* Check whether observed difference and true difference are on
                            same side of 4pp increase criterion */
                     gen ch`y'_`i'=dif`y'_`i'>.04
                     su ch`y'_`i'

                     /* Calculate overall mean proportion agreeing */
                     egen p`y'_`i'=mean(ch`y'_`i')

                     /* Calculate mean proportion agreeing for each replication
                             in order to assign Monte Carlo CIs */
```

```
                        egen b`y' = mean(ch`y'_`i'), by(replic)
                        replace b`y' = . if id != 1
                        centile b`y', c(2.5 97.5)

                        gen lb`y'_`i' = r(c_1)
                        gen ub`y'_`i' = r(c_2)
                        drop b`y'

                        drop early`y'_`i' obs_pr`y'_`i' dif`y'_`i' ch`y'_`i'

                }
                drop pr_`i'

                /* Amuse the user */
                nois di ". " _cont
        }
}

/* Keep important results */
keep p1_* p25_* p9_* lb* ub*

duplicates drop
gen i=1

/* Reshape to tider format */
reshape long p1_ p25_ p9_ lb1_ lb25_ lb9_ ub1_ ub25_ ub9_ ,i(i) j(ch)

/* Replace as %tages */
foreach var in p1_ p25_ p9_ lb1_ lb25_ lb9_ ub1_ ub25_ ub9_ {
        replace `var' = `var'*100
}

/* Map 'change' variable to percentage difference */
replace ch=(ch-41)/10
save `inc_4pp_results', replace

/*******************************/
/* Plot the 4pp-change results */
/*******************************/
use `inc_4pp_results', clear

gen success = ch >= 4
gen fail    = ch <= 4
gen one = 100
gen zero = 0

qui summ p1_ if ch == -4
local prop_m4 = r(mean)
local prop_m4 = string(`prop_m4', "%9.1f")

qui summ p1_ if ch == -2
local prop_m2 = r(mean)
local prop_m2 = string(`prop_m2', "%9.1f")

forval i = 0/6 {
        local j = 2*`i'
        qui summ p1_ if ch == `j'
        local prop_`j' = r(mean)
        local prop_`j' = string(`prop_`j'', "%9.1f")
}

gen plab = string(p1_,"%9.1f")
gen byte mlabp = 10 if  fail
replace mlabp  =  4 if !fail
replace mlabp  = 12 if ch == -4 | ch == -2
replace mlabp  =  6 if ch == 12 | ch == 10
replace mlabp  = 10 if ch == 6 // to get it out of the way
```

```
/* Draw the plot */
twoway          (rarea p1_ one  ch if  success, fcolor(red*.2) fintens(100) lcolor(gs16) )  ///
                (rarea p1_ zero ch if  fail   , fcolor(red*.2) fintens(100) lcolor(gs16) )  ///
                (function y= 25                                , lstyle(grid) range(-4 12) ) ///
                (function y= 50                                , lstyle(grid) range(-4 12) ) ///
                (function y= 75                                , lstyle(grid) range(-4 12) ) ///
                (function y=100                                , lstyle(grid) range(-4 12) ) ///
                (function y=4                          ///
                       , lstyle(grid) range(0 100) horizontal ) ///
                (function y=80                         ///
                       , lcolor(blue) lpattern(shortdash) range(6.2 8)  ) ///
                (function y=6.2                                ///
                       , lcolor(blue) lpattern(shortdash) range(0 80) horizontal  ) ///
                (line p1_  ch, lcolor(gs0 ) lpattern(l) ) ///
                (scatter p1_ ch if mod(ch,2) == 0 ///
                       , msymb(o) mcolor(gs0) mlab(plab) mlabvp(mlabp)) ///
                ,      xtitle("Underlying change in percentage early stage") ///
                       ytitle("% of CCGs achieving {&ge}4% change"  ) ///
                       xlabel(-4(2)12 ) ///
                       xtick(-3(2)11 ) ///
                       xmtick(-4(.2)12 ) ///
                       ylabel(0(25)100, angle(h) grid tlstyle(grid) ) ///
                       xline(4, lstyle(grid) ) ///
                       ysc(noline) ytick(0, tlstyle(tick) ) ///
                       legend(off) plotregion( margin(b=0) lstyle(none) ) ///
                       text(   62.5  0 "Correctly classified" "as below target" ///
                                 37.5  9 "Correctly classified" "as meeting target" ///
                                 , size(small) color(gs8) ) ///
                       text(    5    2 "Wrongly classified" "as meeting target" ///
                                 95    6 "Wrongly classified" "as below target" ///
                                 , size(small) color(gs8) ) ///
                       text(   80    10 "80% correctly classified" ///
                                 , size(small) color(blue) ) ///
                       ysize(4) xsize(6)
graph export fig4_change_mc_revised_v1.png, replace

log close
```

References

[1] NHS England, CCG Outcomes Indicator Set 2014/15—at a glance, 2013. https://www.england.nhs.uk/wp-content/uploads/2013/12/ccg-ois-1415-at-a-glance.pdf. (Accessed 19 September 2016).

[2] Department of Health, Improving Outcomes: A Strategy for Cancer, 2011. https://www.gov.uk/government/publications/the-national-cancer-strategy.

[3] NHS England, Quality Premium: 2016/17 Guidance for CCGs, 2016. https://www.england.nhs.uk/wp-content/uploads/2016/03/qualty-prem-guid-2016-17.pdf. (Accessed 29 June 2016).

[4] Y. He, R. Yucel, A.M. Zaslavsky, Misreporting, missing data, and multiple imputation: improving accuracy of cancer registry databases, Chance (New York, N.Y.) 21 (3) (2008) 55–58.

[5] A. Finkelstein, M. Gentzkow, P. Hull, H. Williams, Adjusting risk adjustment—accounting for variation in diagnostic intensity, N. Engl. J. Med. 376 (7) (2017) 608–610.

[6] G. Lyratzopoulos, M.N. Elliott, J.M. Barbiere, L. Staetsky, C.A. Paddison, J. Campbell, M. Roland, How can health care organizations be reliably compared? Lessons from a national survey of patient experience, Med. Care 49 (8) (2011) 724–733.

[7] J.L. Adams, A. Mehrotra, J.W. Thomas, E.A. McGlynn, Physician cost profiling—reliability and risk of misclassification, N. Engl. J. Med. 362 (11) (2010) 1014–1021.

[8] S.J. Stocks, E. Kontopantelis, A. Akbarov, S. Rodgers, A.J. Avery, D.M. Ashcroft, Examining variations in prescribing safety in UK general practice: cross sectional study using the Clinical Practice Research Datalink, BMJ 351 (2015).

[9] K. Walker, J. Neuburger, O. Groene, D.A. Cromwell, J. van der Meulen, Public reporting of surgeon outcomes: low numbers of procedures lead to false complacency, Lancet 382 (9905) (2013) 1674–1677.

[10] Public Health England, Public Health Outcomes Framework, 2016. http://www.phoutcomes.info/. (Accessed 8 March 2016).

[11] L. Sobin, M. Gospodarowicz, C. Wittekind, TNM Classification of Malignant Tumours, Wiley-Blackwell, Oxford, UK, 2009.

[12] T.P. Morris, I.R. White, P. Royston, Tuning multiple imputation by predictive mean matching and local residual draws, BMC Med. Res. Methodol. 14 (1) (2014) 1–13.

[13] S. van Buuren, H.C. Boshuizen, D.L. Knook, Multiple imputation of missing blood pressure covariates in survival analysis, Stat. Med. 18 (6) (1999) 681–694.

[14] I.R. White, P. Royston, A.M. Wood, Multiple imputation using chained equations: issues and guidance for practice, Stat. Med. 30 (2011).

[15] 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 26 (3) (2015) 421–428.

[16] U. Nur, L.G. Shack, B. Rachet, J.R. Carpenter, M.P. Coleman, Modelling relative survival in the presence of incomplete data: a tutorial, Int. J. Epidemiol. 39 (1) (2010) 118–128.

[17] D.B. Rubin, Multiple Imputation for Nonresponse in Surveys, John Wiley and Sons, New York, 1987.

[18] E.H. Lawson, C.Y. Ko, J.L. Adams, W.B. Chow, B.L. Hall, Reliability of evaluating hospital quality by colorectal surgical site infection type, Ann. Surg. 258 (6) (2013) 994–1000.

[19] G. Abel, C.L. Saunders, S.C. Mendonca, C. Gildea, S. McPhail, G. Lyratzopoulos, Variation and statistical reliability of publicly reported primary care diagnostic activity indicators for cancer: a cross-sectional ecological study of routine data, BMJ Qual. Saf. (2017), http://dx.doi.org/10.1136/bmjqs-2017-006607 pii: bmjqs-2017-006607. [Epub ahead of print].

[20] J.L. Adams, The Reliability of Provider Profiling: A Tutorial, RAND, Corporation, Santa Monica, CA, 2009.

[21] M. Roland, M. Elliott, G. Lyratzopoulos, J. Barbiere, R.A. Parker, P. Smith, P. Bower, J. Campbell, Reliability of patient responses in pay for performance schemes: analysis of national General Practitioner Patient Survey data in England, BMJ (2009) b3851, http://dx.doi.org/10.1136/bmj.b3851.

[22] StataCorp, Stata Statistical Software: Release 13, StataCorp LP, College Station, TX, 2013.

[23] J.B. Dimick, H. Welch, J.D. Birkmeyer, Surgical mortality as an indicator of hospital quality: the problem with small sample size, JAMA 292 (7) (2004) 847–851.

[24] S. Siregar, R.H.H. Groenwold, E.K. Jansen, M.L. Bots, Y. van der Graaf, L.A. van Herwerden, Limitations of ranking lists based on cardiac surgery mortality rates, Circ. Cardiovasc. Qual. Outcomes 5 (3) (2012) 403–409.

[25] N. Howlader, A.-M. Noone, M. Yu, K.A. Cronin, Use of imputed population-based cancer registry data as a method of accounting for missing information: application to estrogen receptor status for breast cancer, Am. J. Epidemiol. 176 (4) (2012) 347–356.

[26] N. Eisemann, A. Waldmann, A. Katalinic, Imputation of missing values of tumour stage in population-based cancer registration, BMC Med. Res. Methodol. 11 (1) (2011) 129.

[27] B. Jarman, S. Gault, B. Alves, A. Hider, S. Dolan, A. Cook, B. Hurwitz, L.I. Iezzoni, Explaining differences in English hospital death rates using routinely collected data, BMJ 318 (7197) (1999) 1515–1520.

[28] G.A. Abel, C.L. Saunders, G. Lyratzopoulos, Cancer patient experience, hospital performance and case mix: evidence from England, Future Oncol. 10 (9) (2013) 1589–1598.

[29] J. Mant, N. Hicks, Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction, BMJ 311 (7008) (1995) 793–796.

[30] A. Donabedian, Evaluating the quality of medical care, Milbank Memorial Fund Q. 44 (3) (1966) 166–206.

[31] M. Quinn, P. Babb, J. Jones, E. Allen, Effect of screening on incidence of and mortality from cancer of cervix in England: evaluation based on routinely collected statistics, BMJ 318 (7188) (1999) 904.

[32] S.M.E. Geurts, N.J. Massat, S.W. Duffy, Likely effect of adding flexible sigmoidoscopy to the English NHS Bowel Cancer Screening Programme: impact on colorectal cancer cases and deaths, Br. J. Cancer 113 (1) (2015) 142–149.

[33] H. Møller, C. Gildea, D. Meechan, G. Rubin, T. Round, P. Vedsted, Use of the English urgent referral pathway for suspected cancer and mortality in patients with cancer: cohort study, BMJ 351 (2015) h5102, http://dx.doi.org/10.1136/bmj.h5102.

[34] M. Shawihdi, E. Thompson, N. Kapoor, G. Powell, R.P. Sturgess, N. Stern, M. Roughton, M.G. Pearson, K. Bodger, Variation in gastroscopy rate in English general practice and outcome for oesophagogastric cancer: retrospective analysis of Hospital Episode Statistics, Gut 63 (2) (2014) 250–261.