ORIGINAL ARTICLE
LUNG CANCER

Apples and pears? A comparison of two sources of national lung cancer audit data in England

Aamir Khakwani[1], Ruth H. Jack[2], Sally Vernon[3], Rosie Dickinson[4], Natasha Wood[3], Susan Harden[5], Paul Beckett[6], Ian Woolhouse[7] and Richard B. Hubbard[1]

Affiliations: [1]University of Nottingham, Division of Epidemiology and Public Health, Nottingham, UK. [2]Public Health England, National Cancer Registration and Analysis Services, Nottingham, UK. [3]Public Health England, National Cancer Registration and Analysis Service, Cambridge, UK. [4]Royal College of Physicians, National Lung Cancer Audit, London, UK. [5]Cambridge University Hospital NHS Foundation Trust, Cambridge, UK. [6]Derby Hospital NHS Foundation Trust, Division of Respiratory Medicine, Derby, UK. [7]University Hospital Birmingham NHS Foundation Trust, Division of Respiratory Medicine, Birmingham, UK.

Correspondence: Aamir Khakwani, University of Nottingham, Division of Epidemiology and Public Health, Clinical Sciences Building, Nottingham City Hospital, Nottingham, NG5 1PB, UK.
E-mail: aamir.khakwani@nottingham.ac.uk

ABSTRACT In 2014, the method of data collection from NHS trusts in England for the National Lung Cancer Audit (NLCA) was changed from a bespoke dataset called LUCADA (Lung Cancer Data). Under the new contract, data are submitted *via* the Cancer Outcome and Service Dataset (COSD) system and linked additional cancer registry datasets. In 2014, trusts were given opportunity to submit LUCADA data as well as registry data. 132 NHS trusts submitted LUCADA data, and all 151 trusts submitted COSD data. This transitional year therefore provided the opportunity to compare both datasets for data completeness and reliability.

We linked the two datasets at the patient level to assess the completeness of key patient and treatment variables. We also assessed the interdata agreement of these variables using Cohen's kappa statistic, κ.

We identified 26 001 patients in both datasets. Overall, the recording of sex, age, performance status and stage had more than 90% agreement between datasets, but there were more patients with missing performance status in the registry dataset. Although levels of agreement for surgery, chemotherapy and external-beam radiotherapy were high between datasets, the new COSD system identified more instances of active treatment.

There seems to be a high agreement of data between the datasets, and the findings suggest that the registry dataset coupled with COSD provides a richer dataset than LUCADA. However, it lagged behind LUCADA in performance status recording, which needs to improve over time.

@ERSpublications
New lung cancer data submission method provides a richer dataset http://ow.ly/zE5r30ceaUU

Cite this article as: Khakwani A, Jack RH, Vernon S, *et al.* Apples and pears? A comparison of two sources of national lung cancer audit data in England. *ERJ Open Res* 2017; 3: 00003-2017 [https://doi.org/10.1183/23120541.00003-2017].

Received: Jan 04 2017 | Accepted after revision: May 14 2017

Conflict of interest: Disclosures can be found alongside this article at openres.ersjournals.com

Introduction

The National Lung Cancer Audit (NLCA) was established in 2004 to measure the process and outcomes of care for people with lung cancer [1]. Until recently, each of the 151 National Health Service (NHS) trusts in England uploaded the dataset, known as LUCADA (Lung Cancer Data), annually, using a bespoke system, to a centralised infrastructure provided by the Health and Social Care Information Centre (HSCIC), while separate arrangements were provided for other UK countries to submit data. Although participation in the audit was not mandatory, studies have validated the LUCADA data and shown it to be reliable and representative of lung cancer patients in England [1], and the ascertainment of cases and data completeness improved considerably from 2008 onwards [2]. LUCADA data have been linked to other databases, such as Hospital Episode Statistics (HES), to identify additional cases of surgery and chemotherapy [3, 4], which were not recoded in LUCADA. In addition, HES data has been used to add a comorbidity index (Charlson Score) based on previous inpatient admissions [4, 5].

In part due to the success of LUCADA, a national cancer dataset (Cancer Outcome and Service Dataset (COSD)) with generic and site-specific data items has been developed over the past 10 years by the National Cancer Intelligence Network (NCIN) and the English National Cancer Registration and Analysis Service (NCRAS). Since January 2013, this dataset has become the national standard for reporting all cancer activity in the NHS in England. In 2014, the contract for delivery of the National Lung Cancer Audit was transferred to the Royal College of Physicians (RCP) and, as part of the new contractual arrangements, a new methodology of data collection has been introduced that utilises the COSD, together with other sources of cancer registration data including HES (referred to as the "registry dataset").

During the transition phase to the new contract (covering patients diagnosed in 2014), it became apparent that NHS trusts might not be as far advanced with monthly COSD submissions as had been hoped, and the trusts were offered a one-off opportunity to submit a LUCADA data file. 132 NHS trusts took up this offer. This dual data entry in 2014 presents a one-time opportunity to assess the data completeness of the COSD (submitted monthly), validate key patient data fields, and assess the interdata agreement of treatment data in comparison to the established and validated LUCADA dataset.

Methods

Comparison of the two datasets

LUCADA 2004–2014

LUCADA is a longitudinal validated database [1], which, in its final iteration, contained 103 key data items on people with lung cancer. In this study, we have used the LUCADA data files submitted in June 2015 by 132 of the 151 English NHS trusts. These files include data on people diagnosed with lung cancer in 2014.

COSD and registration data 2014 onwards

The COSD dataset contains data on core cancer and also lung-specific data items. NHS hospitals in England submit a COSD file to the NCRAS each month for all new cancer cases. The data in the file contribute to the registration process for each new cancer, together with data from pathology reports, patient administration systems and treatment datasets. At the time of analysis of the 2014 data, the detailed radiotherapy and chemotherapy treatment datasets were not available for linkage.

Data linkage and data management

We linked the people with lung cancer in the two datasets using anonymised NHS numbers, and conducted a cross-sectional analysis on all patients in the LUCADA and registry datasets who were first diagnosed in England between January 1, 2014 and December 31, 2014. We excluded people with mesothelioma from this study.

Both LUCADA and the registry dataset include a number of similar data items, including age, sex, lung cancer stage, performance status, pathological diagnosis and anticancer treatment (with dates), in addition to lung cancer pathway data (*e.g.* whether a patient was assessed by a lung cancer nurse specialist and the date the patient was discussed at a lung cancer multidisciplinary team (MDT)). Lung cancer morphology was defined using the recorded Systematised Nomenclature of Medicine (SNOMED) codes where available, and patients with unrecorded SNOMED code were "presumed" to have nonsmall cell lung cancer (NSCLC). Performance status in both datasets is classified according to the World Health Organization definition, and lung cancer stage is defined using the Union for International Cancer Control definitions.

We considered a patient to have received active anticancer treatment (surgery, chemotherapy, external-beam radiotherapy or brachytherapy) if they had a date of treatment in the LUCADA. In the registry dataset, we identified Office of Population Census and Survey Classification of Intervention

version 4 (OPCS-4) codes that correspond to curative surgical treatment with a valid date, which were defined as 1 month before to 6 months after diagnosis date. Chemotherapy, external-beam radiotherapy and brachytherapy records were also identified with the earliest date of treatment in the registry data during the same period.

Statistical analysis

All analyses were performed using STATA 12. Initially, we assessed the proportion of patients who were matched and present in both datasets. We then compared proportions of key patient variables, including age, sex, performance status, lung cancer stage and anticancer treatment and the differences in key dates in both datasets, including date of diagnosis and date discussed at a lung cancer MDT. We assessed the interdata agreement between LUCADA and the registry dataset for these variables using Cohen's kappa statistic, κ. For ordered categorical variables such as performance status and stage, we used weighted kappa to put more emphasis on variation across categories. Finally, we assessed the differences in patients who were matched in the registry dataset and LUCADA and those patients who belonged to the 132 represented in the 2014 LUCADA audit but were not matched.

Results

The registry and LUCADA datasets used in our analyses consisted of 35 518 patients from 151 trusts and 27 995 patients from 132 trusts, respectively. 26 001 patients were present in both datasets (figure 1). 1994 (7%) of the people recorded in LUCADA did not have an entry in the registry dataset, while 9517 (26.7%) patients in the registry dataset did not have an entry in the LUCADA database. Of these 9517 patients, 3482 (37%) patients belonged to one of the 19 trusts who were not included in the 2014 LUCADA audit report and were excluded from further analysis, while 6035 (63%) remained unmatched.

Patient variable agreement

Table 1 presents the difference in recording dates and agreement between the two datasets. Of the 26 001 patients present in both datasets, 14 680 (56%) had a different date of diagnosis in the LUCADA and registry datasets, but this difference was often small, with 19 218 (74%) diagnosis dates within 7 days and 22 358 (86%) within 14 days of each other. For the date of discussion at an MDT in the two datasets, 7397 (28%) had different dates but 18 293 (70%) were within 7 days and 19 124 (74%) were within 14 days of each other.

For key patient variables, the proportion distribution of sex and age were similar in the LUCADA and registry datasets. 23 patients had different recorded sex and 838 patients had different age recorded (99% of which were within 1 year of each other). There was a high level of agreement for sex and age (agreement=99% and κ=0.99). The registry dataset had a larger proportion of patients with missing performance status values compared with the LUCADA (27% *versus* 11%). Excluding missing values of performance status, we assessed the agreement of performance status in the two databases, which was found to be 97% (κ=0.91). We also looked at the agreement of stage recording in the two datasets, excluding missing stage data, and this was 94% (κ=0.81).

Treatment data agreement

Table 2 presents the agreement of treatment data in the LUCADA and registry datasets. In general, we identified more patients receiving surgery (4127 *versus* 3657), chemotherapy (8775 *versus* 7918), radiotherapy (7739 *versus* 7417) and brachytherapy (34 *versus* 27) in the registry dataset compared with

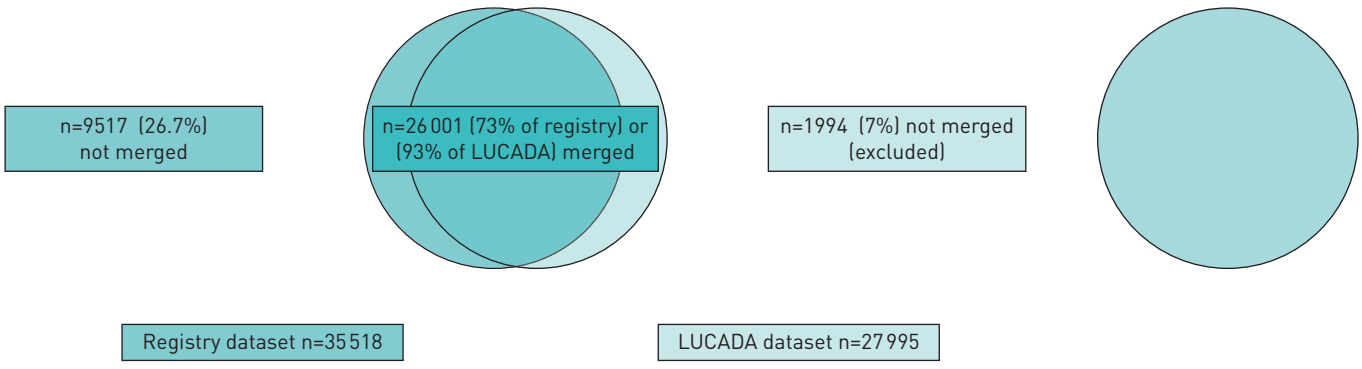

FIGURE 1 Data merging of registry dataset and LUCADA (Lung Cancer Data).

TABLE 1 Key patient features across the registry dataset and LUCADA (Lung Cancer Data) for the 132 National Health Service trusts and level of agreement (n=26 001)

Key variables	Registry database	LUCADA database	Agreement	κ
Sex				
Female	11 990 (46%)	11 987 (46%)		
Male	14 011 (54%)	14 014 (54%)	99%	0.99
Age years				
<65	6032 (23%)	6025 (23%)		
65–80	14 249 (55%)	14 242 (55%)		
>80	5720 (22%)	5734 (22%)	99%	0.99
Pathology confirmed				
No	6341 (24%)	7664 (29)		
Yes	19 660 (76%)	18 337 (71%)	89%	0.73
Lung cancer type				
Small cell	2975 (11%)	2958 (11%)		
Carcinoid	189 (1%)	234 (1%)		
Nonsmall cell	22 837 (88%)	22 809 (88%)	97%	0.87
Performance status				
0	3816 (15%)	4278 (16%)		
1	6550 (25%)	7869 (30%)		
2	4025 (15%)	5109 (20%)		
3	3553 (14%)	4527 (17%)		
4	1152 (4%)	1424 (5%)	97%	0.91[#,¶]
Missing	6905 (27%)	2794 (11%)	83%	0.57[#]
Stage				
IA	2976 (8%)	2226 (9%)		
IB	2296 (6%)	1714 (7%)		
IIA	1405 (4%)	1087 (4%)		
IIB	1236 (4%)	1025 (4%)		
IIIA	3981 (11%)	3330 (13%)		
IIIB	2844 (8%)	2470 (9%)		
IV	16 758 (47%)	12 258 (47%)	96%	0.90[#,¶]
Missing	4022 (11%)	1891 (7%)	94%	0.81[#]

[#]: weighted κ; [¶]: excluding missing data.

the LUCADA dataset. Out of the patients who had surgery or chemotherapy in both datasets, around 90% of the patients had the same treatment date (2961 patients out of 3414 for surgery and 6865 patients out of 7650 patients for chemotherapy), with an agreement of more than 90% (κ=0.85 and κ=0.88,

TABLE 2 Treatment data agreement across registry dataset and LUCADA (Lung Cancer Data) for matched patients (n=26 001)

Registry dataset	LUCADA		Patients with the same date in two datasets	Patients with ±1 day in two datasets	Agreement	κ
	No	Yes				
Surgery						
No	21 631	243				
Yes	713	3414	2961 (87%)	3301 (97%)	96%	0.86
Chemotherapy						
No	16 958	268				
Yes	1125	7650	6865 (90%)	6928 (91%)	95%	0.88
External-beam radiotherapy						
No	17 214	1048				
Yes	1370	6369	6002 (94%)	6073 (95%)	91%	0.77
Brachytherapy						
No	25 951	16				
Yes	23	11	10 (91%)	10 (91%)	99%	0.36

TABLE 3 Patient feature comparison for matched and unmatched patients in the registry dataset (n=32 036)

	LUCADA and registry merged patients from 132 trusts	LUCADA and registry unmerged patients from 132 trusts
Patients	26 001	6035
Sex		
Female	11 990 (46%)	2867 (48%)
Male	14 011 (54%)	3168 (52%)
Age years		
<65	6032 (23%)	1287 (21)
65–80	14 249 (55%)	3023 (50)
>80	5720 (22%)	1725 (29)
Performance status		
0	3816 (15%)	394 (7%)
1	6550 (25%)	639 (11%)
2	4025 (15%)	342 (6%)
3	3553 (14%)	384 (6%)
4	1152 (4%)	131 (2%)
Missing	6905 (27%)	4145 (69%)
Stage		
IA	2190 (8%)	502 (8%)
IB	1836 (7%)	271 (4%)
IIA	1096 (4%)	184 (3%)
IIB	1018 (4%)	131 (2%)
IIIA	3267 (13%)	385 (6%)
IIIB	2353 (9%)	230 (4%)
IV	12 699 (49%)	2443 (41%)
Missing	1542 (6%)	1889 (31%)
Lung cancer type		
Small cell	2975 (11%)	382 (6%)
Carcinoid	189 (1%)	98 (2%)
Nonsmall cell	22 837 (88%)	5555 (92%)
Surgery		
No	21 825 (84%)	5252 (87%)
Yes	4176 (16%)	783 (13%)
Chemotherapy		
No	17 226 (66%)	5031 (83%)
Yes	8775 (34%)	1004 (17%)
External-beam radiotherapy		
No	18 262 (70%)	5198 (86%)
Yes	7739 (30%)	837 (14%)
Brachytherapy		
No	25 967 (99.9%)	6026 (99.9%)
Yes	34 (0.1%)	9 (0.1%)

LUCADA: Lung Cancer Data.

respectively) in the two datasets. 97% of the surgery dates and 91% of the chemotherapy dates were within 1 day of each other. We observed good agreement (κ=0.77) for the recording of radiotherapy, with 94% of the patients having the same date of radiotherapy in the two datasets, but the lowest agreement was seen with brachytherapy (κ=0.36).

Comparison for matched and unmatched patients within COSD

Table 3 shows results of the comparison of key patient variables for patients from the 132 trusts within the registry dataset who were matched (n=26 001) with patients in the LUCADA and who were not matched (n=6032). There was no difference in the proportion distribution for sex between the two groups; however, patients who were not matched had a higher proportion of older patients compared with the matched group (29% *versus* 22%). We also found that the unmatched patients had a higher proportion of missing performance status (69% *versus* 27%) and stage (31% *versus* 6%) compared with the other group. For treatment received, the matched patients had a higher proportion of patients who received surgery (16%

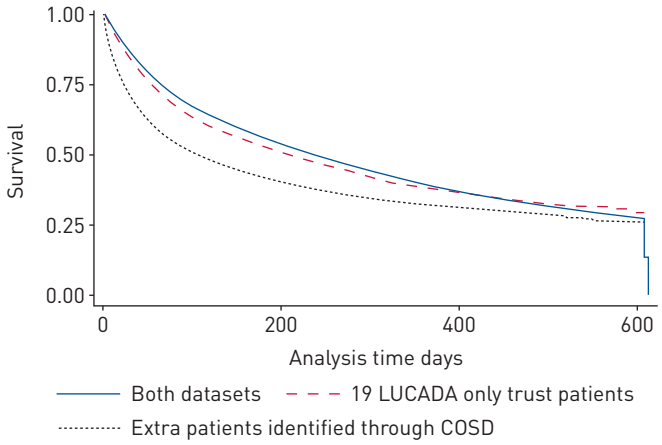

FIGURE 2 Kaplan–Meier curve of survival (in days) by data source. LUCADA: Lung Cancer Data; COSD: Cancer Outcome and Service Dataset.

versus 13%), chemotherapy (34% *versus* 17%) and external-beam radiotherapy (30% *versus* 14%) compared with the patients who were not matched.

Post hoc *validation*

As a result of our initial analyses, we conducted two *post hoc* investigation of patients who were present in one database but not the other. For patients who were identified by the registry and not LUCADA (n=6035), we investigated patients first diagnosed in 7 NHS trusts (n=583 (10%)) and summarised the reason for their exclusion from the LUCADA dataset. We also investigated 70 random patients out of 1994 patients (3.5%) who were in LUCADA but not in the registry dataset by sending the anonymised patient identification to Public Health England to assess their reason for exclusion from the registry dataset.

Our review of patients identified by the cancer registry but not LUCADA for 7 out of 132 NHS trusts (n=583) revealed that 33% (n=193) had an error in data entry or submission to the system, 8% (n=45) did not have a primary lung cancer diagnosis, 10% (n=57) had either a recurrent diagnosis of lung cancer, had no documentation of cancer in the system, or were not from an English trust, 4% (n=22) had a diagnosis year outside the analysis period, and 46% (n=266) had incorrect recording of the trust where the patient was first seen by a physician. Our review of the sample of 70 patients in the LUCADA but not in the registry dataset found that the majority of the patients had a diagnosis date outside the analysis period (58%) while others were subsequently entered in the COSD system (after the raw data were provided to us for the analysis as the data entry is an open process) (33%). This could be because the old LUCADA-based audit relied on "date first seen by a lung cancer specialist", whereas the new registry system relies on the date of diagnosis of lung cancer, which can lead to patients first seen late in a year in the LUCADA system being allocated in the registry data for the subsequent year.

Overall survival

We compared the median survival in days for patients who were present in both datasets (n=26 001), those belonging to the 19 trusts who only submitted LUCADA data (n=3482) and the extra patients identified through the registry system (n=6035). Figure 2 presents a Kaplan–Meier curve showing that patients who were present in both datasets had a better median day survival (234 days, interquartile range (IQR) of 64–608 days) compared with the 19 trusts who submitted LUCADA data (202 days) and the extra patients (105 days). This difference was accounted for when we adjusted for patient features including age, sex, performance status, stage and lung cancer type (figure 3).

Discussion

Our results demonstrate that that the new system for collecting lung cancer audit data using the cancer registration process identifies a greater number of patients with lung cancer compared to the legacy LUCADA submission process. For patients that were represented in both datasets, there was over 95% agreement for key patient features including age, sex, lung cancer type and stage of disease. There was a higher proportion of missing data on performance status (16% higher) in the registry dataset compared with LUCADA. The new registry dataset identified more patients who received treatment with surgery, chemotherapy and external-beam radiotherapy compared with LUCADA. We also observed that the patients who were identified in the registration dataset but not LUCADA tended to be older, with higher rates of missing performance status and stage information and a lower proportion of recorded active

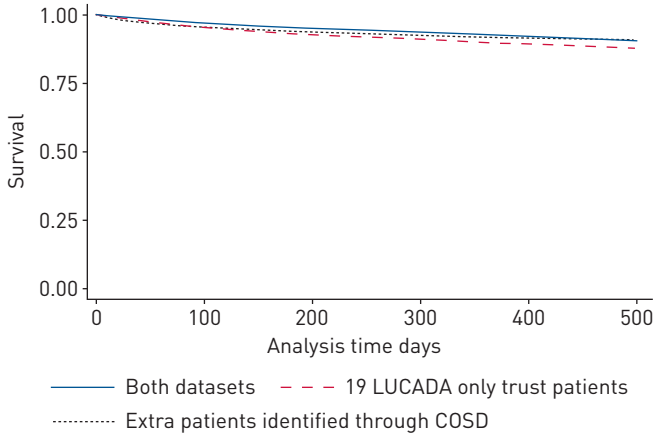

FIGURE 3 Kaplan–Meier curve of survival (in days) by data source adjusted for age, sex, performance status, cancer stage and lung cancer type. LUCADA: Lung Cancer Data; COSD: Cancer Outcome and Service Dataset.

treatment including surgery and chemotherapy. In summary, the new system seems to perform better than its predecessor in all areas except completeness of performance status data.

The additional patients identified in the registry dataset but not LUCADA are probably a result of the multiple sources of data that are used by cancer registration officers to register a case of lung cancer, with increased opportunities for patients to be identified. In addition to the registry dataset, which largely replicates LUCADA, the NCRAS also has direct access to pathology reports and death certificates, which are less reliant on hospital informatics systems for submission. The fact that these extra patients tended to be older, with less complete performance status data and less likely to receive active treatment, raises the possibility that there is a cohort of lung cancer patients that hospital lung cancer MDTs are not aware of and were not previously included in the audit. We also suspect that these patients missed out on being discussed by the MDT because of their short survival.

The higher level of agreement for key data items in patients represented in both datasets is encouraging. Although there were discrepancies between the two datasets for the dates of an event (*e.g.* date of diagnosis and date of surgery), most of these dates lay very close to each other and so would not significantly affect analysis of treatment times or survival. Similarly, although more patients were identified as having received active treatment in the registration dataset, the treatment rates expressed as a proportion of the whole population were similar to those reported *via* LUCADA in the 2013 report [6] (*e.g.* surgery: 15.5% registry *versus* 15.1% LUCADA). Accurate recording of performance status is important to allow robust risk adjustment of patient outcomes, so it was disappointing to find that this field was incomplete in 27% of patients in the registry dataset. However, it is anticipated that completeness for performance status will improve as lung cancer teams become more familiar with the new process for data submission.

One of the limitations of the registry dataset, which also existed for the LUCADA dataset, is the lack of detailed information collected on chemotherapy and radiotherapy, in particular the dose and number of cycles or fractions of treatment [3, 7]. However, the Systemic Anticancer Therapy (SACT) and Radiotherapy Datasets (RTDS) will become linked with the COSD system and form part of an even complete registry dataset in due course.

Conclusion

In general, the registry dataset with the COSD system is a more effective dataset than the LUCADA dataset, as it has better recording of pathology and treatment information, including surgery and chemotherapy, while other patient features have a similar level of completeness. Although the missing information of performance status is high in the 2014 registry data, it is anticipated that trusts will strive to improve this in future years. The results from the lung cancer audit shows that this is an excellent model to assess nationwide cancer practices and can lead to standardising cancer pathways and care for other cancers.

References

1 Rich AL, Tata LJ, Stanley RA, *et al.* Lung cancer in England: information from the National Lung Cancer Audit (LUCADA). *Lung Cancer* 2011; 72: 16–22.
2 Health and Social Care Information Centre. National Lung Cancer Audit 2010. http://content.digital.nhs.uk/catalogue/PUB02684/clin-audi-supp-prog-lung-canc-nlca-2010-rep1.pdf
3 Khakwani A, Rich AL, Tata LJ, *et al.* Small-cell lung cancer in England: trends in survival and chemotherapy using the National Lung Cancer Audit. *PLoS ONE* 2014; 9: e89426.

4 Powell HA, Tata LJ, Baldwin DR, *et al.* Early mortality after surgical resection for lung cancer: an analysis of the English National Lung Cancer Audit. *Thorax* 2013; 68: 826–834.
5 Rich AL, Khakwani A, Free CM, *et al.* Non-small cell lung cancer in young adults: presentation and survival in the English National Lung Cancer Audit. *QJM* 2015; 108: 891–897.
6 Health and Quality Improvement Partnership, National Lung Cancer Audit Report 2014: Report for audit period 2013. http://content.digital.nhs.uk/catalogue/PUB16019/clin-audi-supp-prog-lung-nlca-2014-rep.pdf
7 Powell HA, Tata LJ, Baldwin DR, *et al.* Treatment decisions and survival for people with small-cell lung cancer. *Br J Cancer* 2014; 110: 908–915.