BMJ Open

Accuracy of lung cancer ICD-9-CM codes in Umbria, Napoli 3 Sud and Friuli Venezia Giulia administrative healthcare databases: a diagnostic accuracy study

Alessandro Montedori,[1] Ettore Bidoli,[2] Diego Serraino,[2] Mario Fusco,[3] Gianni Giovannini,[1] Paola Casucci,[4] David Franchini,[4] Annalisa Granata,[3] Valerio Ciullo,[3] Maria Francesca Vitale,[3] Michele Gobbato,[5] Rita Chiari,[6] Francesco Cozzolino,[1] Massimiliano Orso,[1] Walter Orlandi,[7] Iosief Abraha,[1,8] for the D.I.V.O. Group

To cite: Montedori A, Bidoli E, Serraino D, *et al.* Accuracy of lung cancer ICD-9-CM codes in Umbria, Napoli 3 Sud and Friuli Venezia Giulia administrative healthcare databases: a diagnostic accuracy study. *BMJ Open* 2018;**8**:e020628. doi:10.1136/bmjopen-2017-020628

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2017-020628).

Received 14 November 2017
Revised 20 February 2018
Accepted 13 April 2018

For numbered affiliations see end of article.

Correspondence to
Dr Ettore Bidoli; bidolie@cro.it

ABSTRACT

Objectives To assess the accuracy of International Classification of Diseases 9th Revision–Clinical Modification (ICD-9-CM) codes in identifying subjects with lung cancer.

Design A cross-sectional diagnostic accuracy study comparing ICD-9-CM 162.x code (index test) in primary position with medical chart (reference standard). Case ascertainment was based on the presence of a primary nodular lesion in the lung and cytological or histological documentation of cancer from a primary or metastatic site.

Setting Three operative units: administrative databases from Umbria Region (890 000 residents), ASL Napoli 3 Sud (NA) (1 170 000 residents) and Friuli Venezia Giulia (FVG) Region (1 227 000 residents).

Participants Incident subjects with lung cancer (n=386) diagnosed in primary position between 2012 and 2014 and a population of non-cases (n=280).

Outcome measures Sensitivity, specificity and positive predictive value (PPV) for 162.x code.

Results 130 cases and 94 non-cases were randomly selected from each database and the corresponding medical charts were reviewed. Most of the diagnoses for lung cancer were performed in medical departments. True positive rates were high for all the three units. Sensitivity was 99% (95% CI 95% to 100%) for Umbria, 97% (95% CI 91% to 100%) for NA, and 99% (95% CI 95% to 100%) for FVG. The false positive rates were 24%, 37% and 23% for Umbria, NA and FVG, respectively. PPVs were 79% (73% to 83%)%) for Umbria, 58% (53% to 63%)%) for NA and 79% (73% to 84%)%) for FVG.

Conclusions Case ascertainment for lung cancer based on imaging or endoscopy associated with histological examination yielded an excellent sensitivity in all the three administrative databases. PPV was moderate for Umbria and FVG but lower for NA.

Strengths and limitations of this study

- This study is the first to have validated International Classification of Diseases 9th Revision–Clinical Modification (ICD-9-CM) codes for lung cancer in three large administrative databases in Italy using the same case definition.
- Medical chart review was used as reference standard to ascertain cases of lung cancer.
- Case ascertainment was based on the presence of a primary nodular lesion in the lung documented by imaging and cytological or histological documentation of cancer from a primary or metastatic site.
- Validation studies of administrative data are related to the context and are not generalisable to other settings.
- We were not able to determine cancer staging and the accuracy of lung cancer ICD-9-CM codes in secondary position.

INTRODUCTION

There is increasing interest in the use of administrative healthcare databases in clinical and health services research as they provide timely and easy access to a large source of information regarding subjects in a defined geographical area.[1–4] This information may include a combination of hospital discharge data, emergency department visit information, physician prescription data or laboratory data.[5] Administrative databases provide easy and cheap access to large numbers of patients over wide geographic regions.[1] Generally, the diagnoses of the disease are stored in administrative databases using specific codes from the International Classification of Diseases, 9th Revision (ICD-9) or 10th Revision (ICD-10) edition.[6]

The use of administrative databases for research is based on an assumption that they avoid recall bias and that these databases convey plausibly accurate data for healthcare utilisation as well as outcome research.[7] However, the most critical elements that need to be considered when using healthcare databases are completeness and validity of the data. Regarding an event or outcome, a database is complete when the proportion of these events observed in the population are identical with those detected in the database and bias can be introduced in the presence of missing data.[7] On the other hand, validity expresses the proportion of 'true' events (disease or exposure) that are verified within the population covered in the database. To avoid biased results based on the use of inaccurate data, an adequate validation of administrative healthcare databases is mandatory.[8] In other words, since validity of registered diagnoses and procedures is variable,[7 9] the accuracy of the source of information (administrative database) needs to be determined by verifying the corresponding clinical information within the reference source of information (eg, medical charts).[4 10–12]

Lung cancer is the most commonly diagnosed neoplasm worldwide and it is the leading cause of cancer-related mortality.[13 14] Consequently, lung cancer raises particular interest within the research community[15 16] and the government as it has enormous implication targets in terms of public health, quality of cancer care,[17] economic burden[18] as well as industry in terms of the development of new innovative drugs.[19] Administrative databases can play an important role in the evaluation of the quality of cancer care,[20] variation in the epidemiology and outcome of the lung cancer,[21 22] survival and other benefits of treatment[23 24] as well as healthcare utilisation and costs.[25]

Several assessments of the validity of oncological codes have been made[26–31] using different case definitions or algorithms as well as multiple sources, including inpatient and physician office records, and the accuracy estimates differed depending on the cancer site and the case definition.[30 31] In Italy, the validity of ICD-9 codes related to lung cancer in administrative databases is limited.[2 32] A systematic review identified only one study that assessed the accuracy of ICD-9 codes related to lung cancer disease.[31] To exploit the productivity of Italian administrative databases in terms of research, evaluation of quality of care and drug utilisation and review, three groups of researchers proposed a research proposal—within a call—to determine the accuracy of ICD-9 codes of relevant cancer diseases in their respective administrative databases.[3 33] The aim of this study was to assess the validity of ICD-9 codes related to lung cancer based on a simple case definition ascertained using the medical chart across three large healthcare databases from Umbria, (NA) and Friuli Venezia Giulia (FVG).

METHODS

Setting and data source

Administrative databases

The target administrative databases for the present study were those of the Umbria Region (890 000 residents), the NA (1 170 000 residents), and the FVG Region (1 227 000 residents). For each database, the corresponding unit (Regional Health Authority of Umbria for the Umbria Region, Registro Tumori Regione Campania for the Local Health Unit 3 of Napoli and Centro di Riferimento Oncologico Aviano for the FVG Region) conducted the same validation process.

In Italy, administrative databases initiated collecting healthcare information regarding their residents starting from the early 90s. These databases gather diagnostic discharge data from public and private hospitals, vital statistics, hospital admission and discharge dates, the admitting hospital department, the principal diagnosis and a maximum of five secondary discharge diagnoses and the principal, and up to five secondary, surgical or pharmacological treatments and diagnostic procedures as well as all drug prescriptions listed in the National Drug Formulary together with the basic characteristics of patients' physicians. The various types of information can be linked within the database and all residents' data can be traced as each resident has a unique, lifetime national identification code. In Italy, healthcare is covered almost entirely by the Italian National Health System; therefore, most residents' significant healthcare information can be found within the healthcare databases.

Every resident has a unique code within the entire national/regional database. For every medical chart, a Hospital Discharge Register is generated and this has a unique code which is generated in a chronologically progressive way throughout the year and is independent from the type of admission (hospital or day hospital, week surgery, etc). The code comprises a root of numbers that are a combination of the regional code, the hospital code and the department code that helps avoid any duplicate even at the national level. Other controls to avoid duplication of the medical charts identity include control of duplicates of rows and potential duplication based on the admission and/or discharge dates of the same subjects independent from the department in which the patient has been admitted.

Source population

The source population was represented by permanent residents aged 18 or above in the Umbria Region, the Local Health Unit 3 of Napoli and the FVG Region. Any resident that has been discharged from hospital with a diagnosis of lung cancer was considered. Residents that have been hospitalised outside the regional territory of competence were excluded from analysis due to the difficulty in obtaining the medical charts.

Patient and public involvement

This was a retrospective study based on consultation of medical charts. Patients were not directly involved.

Case selection and sampling method

In each administrative database, the following process was followed to identify new cases with lung cancer: (1)

records of patients with occurrence of diagnosis of lung cancer between 1 January 2012 and 31 December 2014 were identified using the ICD-9-CM codes 162.x located in primary position of the hospital discharge; (2) records subsequent to the index date were deleted; (3) prevalent cases, that is, those with the same diagnosis (ICD-9-CM codes 162.x in any position) in the 5 years (2007–2011) before the period of interest, were excluded.

This cohort represented our target population from which a sample of cases was obtained using a simple random method.

For controls (non-cases), the following process was followed: (1) subjects aged 18 or older with a diagnosis of cancer disease (ie, patients having a diagnosis of cancer in primary position (ICD-9 140–239)) were identified; (2) from this cohort, subjects with lung cancer (ICD-9-CM codes 162.x in primary position) were excluded; (3) prevalent cases, that is, those with the same diagnosis (ICD-9 140–239 codes in any position) in the 5 years (2007–2011) before the period of interest, were excluded.

This cohort represented our target population from which a sample of non-cases (controls) were obtained using a simple random method.

Chart abstraction and case ascertainment

Medical charts of the randomly selected samples of cases and non-cases were obtained from hospitals for case ascertainment. From each medical chart, the following data were collected: clinical chart number, hospital and ward of admission, date of birth, sex, dates of hospital admission and discharge, signs and symptoms, any diagnostic procedures that contributed to the diagnosis of the cancer, any pharmacological or surgical interventions that were provided for the treatment of the cancer.

Within each unit, two medical doctors (MDs) acting as reviewers received training on data abstraction evaluating the same (n=20) medical charts independently. The inter-rater agreement among the pairs of reviewers within each unit was near perfect ($\kappa > 0.9$). Following the consensus review, data abstraction was completed independently. To ensure consistency among all the reviewers, cases with uncertainty were discussed and resolved through a third party involvement (IA, RC).

We considered the ICD-9-CM codes 162.x valid, when there is evidence of a pulmonary nodule documented with (1) imaging (eg, CT scan) or endoscopy and (2) a cytological or histological diagnosis from a primary or metastatic site positive for either small cell lung cancer or non-small cell lung cancer. Cases and non-cases were validated by pairs of MDs, one of whom was an oncologist.

Statistical analysis

We calculated that a sample of 130 charts of cases was necessary to obtain an expected sensitivity of 80% with a precision of 10% and a power of 80%. For specificity, we calculated that a sample of 94 charts of non-cases was necessary to obtain an expected specificity of 90% with a precision of 10% and a power of 80%,[3] according to binomial exact calculation.[34] The 2×2 tables were developed to calculate sensitivity and specificity with their corresponding 95% CI. Accuracy data were calculated separately for each administrative database.

In case of missing medical charts, we performed a formal sensitivity analysis based on a worst case scenario in which the missing cases were considered as false positives and missing controls were considered false negatives.

RESULTS

The exclusion of prevalent cases of lung cancer in primary position allowed the identification of a cohort of 1690 new cases from Umbria, 1655 from NA and 2013 from FVG. Subsequently, each unit randomly selected 130 cases of which the corresponding medical charts were requested for evaluation. These random samples represented 7.7% of the original new cases for Umbria, 7.6% for NA and 6.5% for FVG. Four (3%) medical charts were not available from NA. Figure 1 displays the identification of cases from the three operative units. For the non-cases, each unit randomly selected 94 medical charts. Two medical charts of non-cases from Umbria were missing (see online supplemental table A).

The most common ICD-9-CM subgroup was the code 162.9 (ie, bronchus and lung, unspecified) accounting for 51% of cases in Umbria, 58% in NA and 35% in FVG, followed by the code 162.3 (ie, upper lobe, bronchus or lung) accounting for 25% in Umbria, 14% in NA and 24% in FVG. The mean age of the patients was 70 years in Umbria, 68 years in NA and 72 years in FVG. Most of the diagnoses (range 66% to 87%) of lung cancer were performed in medical departments. The instrumental tools for diagnosis included CT scan, bronchoscope, chest X-ray and positron emission tomography/CT. The surgical interventions were limited to only 12%–26% of patients and included lobectomy, pneumonectomy and other surgical interventions. Table 1 displays the basic characteristics of lung cancer cases in each unit.

True positive rates resulted very high for all the three units. The sensitivity was 99% (95% CI 95% to 100%) for Umbria, 97% (95% CI 91% to 100%) for NA and 99% (95% CI 95% to 100%) for FVG. The false positive rates were 24%, 37% and 23% for Umbria, NA and FVG, respectively. PPVs were 79% (73% to 83%)%) for Umbria, 58% (53% to 63%)%) for NA and 79% (73% to 84%)%) for FVG.

Table 2 provides cross tabulation of the ICD-9-CM code results from the results of the medical charts, whereas figure 2 displays sensitivities and specificities across the three operative units.

Misclassification of cases and non-cases is described in table 3. Most of false positives cases (89%) were due to missing histological documentation (28 in Umbria, 39 in NA and 23 in FVG), whereas in 11 (11%) cases overall, the histological documentation resulted negative for lung cancer. Overall, only four false negatives were identified and the reasons were due to unclear or

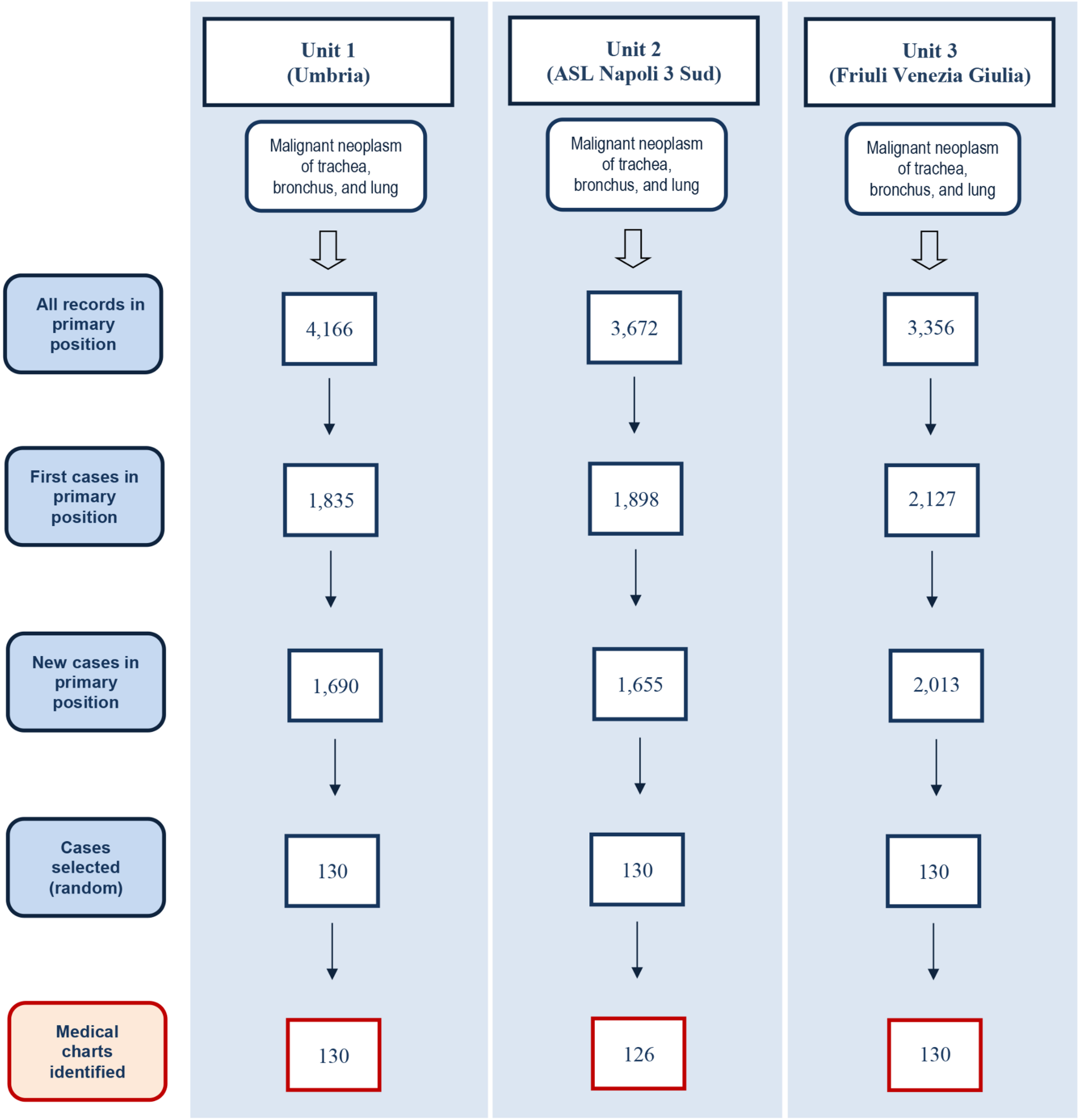
Unit 1
(Umbria)
Unit 2
(ASL Napoli 3 Sud)
Unit 3
(Friuli Venezia Giulia)
Malignant neoplasm of trachea, bronchus, and lung
Malignant neoplasm of trachea, bronchus, and lung
Malignant neoplasm of trachea, bronchus, and lung
All records in primary position
4,166
3,672
3,356
First cases in primary position
1,835
1,898
2,127
New cases in primary position
1,690
1,655
2,013
Cases selected (random)
130
130
130
Medical charts identified
130
126
130

Figure 1 Flow-chart of incident cases identification using the administrative databases and the corresponding charts (final cell) identified and examined.

possible lung cancer histology. No coding errors were identified.

Missing data for cases and non-cases did not affect the estimates of sensitivity and specificity.

A subgroup analysis based on age showed that false positive rates were higher in the age group ≥65 than in the age group <65 years influencing specificity in the Umbria and FVG databases (see online supplemental table B).

DISCUSSION

This study evaluated the ability of three administrative databases (Umbria, NA and FVG) to identify incident lung cancer cases. According to our case definition, that is, the requirement of a clinical or instrumental documentation of a lesion together with the presence of histological documentation within the same medical chart, we determined that ICD-9 codes have an excellent sensitivity across the three databases but a moderate specificity and PPVs in Umbria and FVG, while NA yielding a lower value of specificity (63%) and PPV (58%). The rate of false positives influenced the results of specificity and PPVs and this was predominantly due to missing histological documentation that resulted not present during the evaluation of the first medical chart of the cases. Part of the rate of false positives could be explained by the unavailability of the histological documentation within the first medical chart of admission. If we have used a broader criteria, such as the evaluation of a subsequent medical chart,[28] the addition of surgical procedures[35] or a combination of both,[28] it may have led to higher PPVs. However, despite the PPV estimate resulted similar to that of another Italian study[31] that compared the accuracy of lung cancer ICD-9codes of a regional administrative database versus a cancer

Table 1 Characteristics of patients with lung cancer who were identified in the three administrative healthcare databases

Characteristics	Unit 1 (Umbria)	Unit 2 (ASL Napoli 3 Sud)	Unit 3 (Friuli Venezia Giulia)
Incident cases (N medical chart reviewed)	130	126	130
International Classification of Diseases, 9th Revision code			
162.0 Trachea	1 (1)	0	2 (2)
162.2 main bronchus	3 (2)	16 (13)	7 (5)
162.3 upper lobe, bronchus or lung	33 (25)	18 (14)	31 (24)
162.4 middle lobe, bronchus or lung	3 (2)	3 (2)	5 (4)
162.5 lower lobe, bronchus or lung	19 (15)	10 (8)	19 (15)
162.8 other parts of the bronchus or lung	4 (3)	6 (5)	20 (15)
162.9 bronchus and lung, unspecified	67 (51)	73 (58)	46 (35)
Admission to department			
Medical	86 (66)	109 (87)	105 (81)
Surgical	44 (34)	17 (14)	25 (19)
Sex			
Male	78 (60)	97 (77)	83 (64)
Age, N (%)			
<40	1 (1)	3 (2)	–
40–59	21 (16)	26 (21)	15 (12)
≥60	108 (83)	97 (77)	115 (88)
Instrumental diagnosis			
CT scan (lung)	73	95	57
Bronchoscopy	40	55	46
Chest X-ray	53	27	48
Positron emission tomography/CT (including lung)	20	23	7
Brain CT scan or MRI	7	11	2
Surgical procedures			
Lobectomy	21 (16)	4 (3)	19 (15)
Pneumonectomy	3 (2)	–	3 (2)
Other surgical interventions	10 (8)	11 (9)	4 (3)
Histological/cytological documentation			
Bronchoalveolar lavage (BAL)	37	–	5
Pleural fluid	7	3	5
Biopsy	73	34	38
Resection specimens (after surgical intervention)	30	4	28

Table 2 Cross tabulation of the index test (ICD-9-CM code) results by the results of the reference standard (medical chart)

Operative unit	TP	FP	TN	FN
Unit 1 (Umbria)	102	28	91	1
Unit 2 (ASL Napoli 3 Sud)	73	53	92	2
Unit 3 (Friuli Venezia Giulia)	103	27	93	1

registry, this study obtained a similar PPV of 78.7%—a result similar to that of Umbria and FVG, despite the fact that the tested algorithm was based on a combination of ICD-9-CM diagnosis, surgical procedures, chemotherapy and radiotherapy codes.[31]

In our study, biopsies or surgical procedures could not be performed due to the critical clinical conditions of the patients, or for their advanced age that may explain in part the higher rates of false positives from the NA operative unit. Indeed, published medical literature reports that most lung cancer patients present with advanced disease and are diagnosed based mainly on symptoms.[36 37] This condition may also explain why in our assessment the most prevalent ICD-9-CM subgroup code was 162.9, namely 'bronchus and lung, unspecified', in which case, given the metastatic or locally advanced disease, the site of the primitive tumour loses its relevance because a radical surgical approach is not possible.

The validation of our algorithm can be extended and tested in other regional settings as well as at national level especially in the areas that are not covered by cancer registries. By combining the lung cancer ICD-9 codes with prescription databases, mortality databases and other sources, researchers at regional and national level can efficiently identify a cohort with lung cancer and perform pharmacoepidemiological studies or other health services-related research.

Strength and limitation

Strengths of our work include the requirement for validation purposes of the presence of histological or cytological documentation in addition to a radiological or endoscopic presence of a primary lesion. Unlike studies that used cancer registries to validate lung cancer codes, we ascertained the presence of the disease by using clinical charts to confirm the accuracy of cases that were identified in the administrative databases.

Additionally, our study assessment was based on a prepublished protocol[3] and no deviation from protocol occurred during the study development. We followed recommended guidelines based on the criteria published by the Standard Protocol Items: Recommendations for Interventional Trials initiative for the accurate reporting of investigations of diagnostic studies. Hence, we used

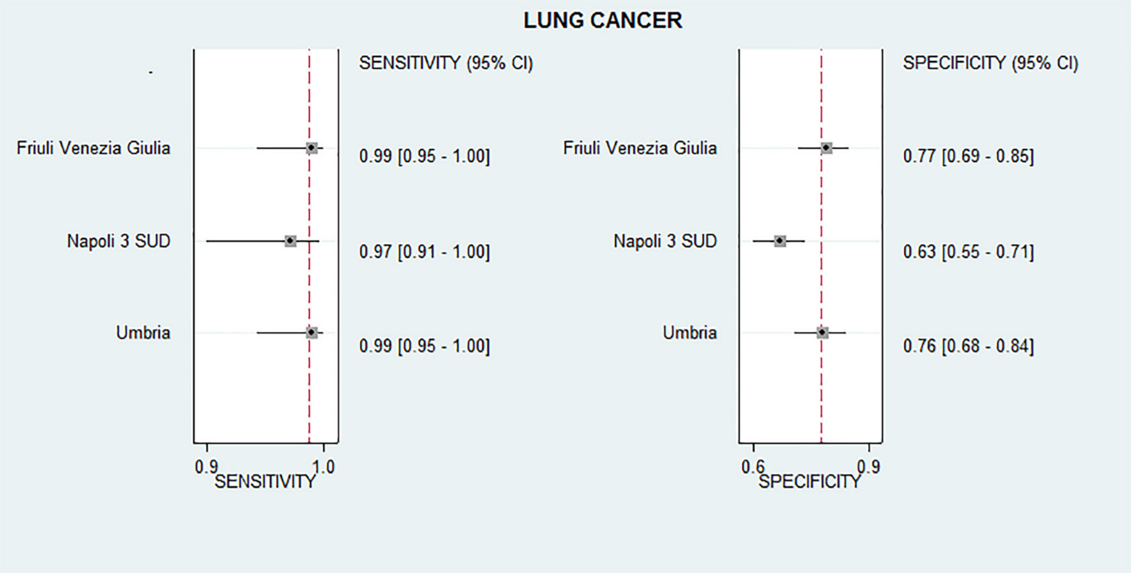

Figure 2 Sensitivity and specificity with 95% CIs for lung cancer International Classification of Diseases 9th Revision–Clinical Modification codes for the three administrative databases.

detailed and explicit eligibility criteria, as well as duplicate and independent processes for medical charts review and data abstraction.[38–40]

We acknowledge some limitations in our study. First, although in our study we considered three Italian regions from three different areas (North, Middle, South) of Italy, the accuracy results of this validation study could not be generalisable to other settings due to the specific characteristics of the patients included in the three regions (such as age, sex, clinical conditions, comorbidities). Second, the stage of the disease could be an important factor that may have influenced the sensitivity, but we could not perform this analysis because the cancer staging is an element that cannot be found in the index test. Third, we did not perform the accuracy of cancer codes in secondary position that may underestimate the incidence of lung cancer disease but further research is necessary to quantify the estimate.

Table 3 Reason for incorrect identification of cases and controls

	Type of misclassification	Umbria	ASL 3 Napoli	Friuli Venezia Giulia
	False positives			
1	Histological examination missing	28	39	23
2	Negative histology	0	7	4
	a) Negative	–	4	1
	b) Squamous metaplasia	–	2	0
	c) Kidney cancer metastases	–	1	0
	d) Lymphoma	–	–	2
	e) Carcinoma in situ	–	–	1
Total		28	46	27
	False negatives			
1	Possible lung cancer	1	2	1
Total		1	2	1

CONCLUSION

We developed a case definition for lung cancer based on imaging or endoscopy associated with histological examination that yielded excellent sensitivity for three population-based healthcare databases, two of which had a moderate PPV. In the NA healthcare database, the PPV resulted lower and future research is needed to address the reason for a higher rate of false positives. The development of this case definition can be extended in other regional and local areas where cancer registries are lacking in Italy. Results from our study support the use of healthcare databases as a valuable tool to investigate several aspects of lung cancer and to conduct population-based longitudinal studies with long-term outcomes.

Author affiliations
[1]Health Planning Service, Regional Health Authority of Umbria, Perugia, Italy
[2]Cancer Epidemiology Unit, Centro di Riferimento Oncologico Aviano, Aviano, Italy
[3]Registro Tumori Regione Campania, ASL Napoli 3 Sud, Brusciano, Italy
[4]Health ICT Service, Regional Health Authority of Umbria, Perugia, Italy
[5]SOC Epidemiologia Oncologica, Centro di Riferimento Oncologico Aviano, Aviano, Italy
[6]Dipartimento di Oncologia, Azienda Ospedaliera Perugia, Perugia, Italy
[7]Direzione salute, Regional Health Authority of Umbria, Perugia, Italy
[8]Centro Regionale Sangue, Azienda Ospedaliera di Perugia, Perugia, Italy

Collaborators Giuliana Alessandrini; Marcello De Giorgi; Roberto Cirocchi; Paolo Collarile; Fabrizio Stracci.

Contributors AM, IA, MF and DS conceived the original idea of the study. IA, DS, AM, MF, EB, GG, FC, MO and WO designed the study. PC, DF, PC, AG, MFV, VC and

MG identified the cohort using administrative database with the supervision of WO, EB, DS, MF, and AM. IA, FC, MO, AG, PC, VC, MFV and MG undertook the data abstraction with the supervision of AM, GG, WO, FS, MF, EB, PC and DS. IA, RC, AM, DS and MF performed case ascertainment. IA, AM, FC, EB, MF, MG and MO performed the analysis. DS, GG, PC, DF, AG, VC, RC, MFV and WO helped with the interpretation of the results. The initial draft of the manuscript was prepared by IA, AM, EB, DS and MF. DS, GG, PC, DF, AG, VC, MFV, MG, RC, FC, MO and WO revised critically the manuscript for important intellectual content. All the authors read and approved the final manuscript. AM, MF and EB are the guarantors of the data for the respective operative units.

Funding This study was developed within the D.I.V.O. project (Realizzazione di un Database Interregionale Validato per l'Oncologia quale strumento di valutazione di impatto e di appropriatezza delle attività di prevenzione primaria e secondaria in ambito oncologico) supported by funding from the National Centre for Disease Prevention and Control (CCM 2014), Ministry of Health, Italy. The study funder was not involved in the study design or the writing of the protocol.

Competing interests None declared.

Patient consent Not required.

Ethics approval Regional Ethics Committee of Umbria (CEAS), authorisation number: 2656/15 (04/11/2015).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol* 2005;58:323–37.
2. Abraha I, Montedori A, Eusebi P, *et al*. The current state of validation of administrative healthcare databases in italy: a systematic review. *Pharmacoepidemiology and Drug Safety* 2012;21:400–00.
3. Abraha I, Serraino D, Giovannini G, *et al*. Validity of ICD-9-CM codes for breast, lung and colorectal cancers in three Italian administrative healthcare databases: a diagnostic accuracy study protocol. *BMJ Open* 2016;6:e010547.
4. Cozzolino F, Abraha I, Orso M, *et al*. Protocol for validating cardiovascular and cerebrovascular ICD-9-CM codes in healthcare administrative databases: the Umbria Data Value Project. *BMJ Open* 2017;7:e013785.
5. Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. *Annu Rev Public Health* 2011;32:91–108.
6. World Health Organization. *International statistical classification of diseases and health related problems*. 10th revision. Geneva: WHO, 1992.
7. West SL, Ritchey ME, Poole C. Validity of pharmacoepidemiologic drug and diagnosis data. *Pharmacoepidemiology*: Wiley-Blackwell, 2012:757–94.
8. Prins H, Hasman A. Appropriateness of ICD-coded diagnostic inpatient hospital discharge data for medical practice assessment. A systematic review. *Methods Inf Med* 2013;52:3–17.
9. Campbell SE, Campbell MK, Grimshaw JM, *et al*. A systematic review of discharge coding accuracy. *J Public Health Med* 2001;23:205–11.
10. Rawson NSB, Shatin D. Assessing the validity of diagnostic data in large administrative healthcare utilization databases. In: Hartzema A, Tilson H, Chan K, eds. *Pharmacoepidemiology and therapeutic risk management*: Harvey Whitney Books, 2008.
11. Montedori A, Abraha I, Chiatti C, *et al*. Validity of peptic ulcer disease and upper gastrointestinal bleeding diagnoses in administrative databases: a systematic review protocol. *BMJ Open* 2016;6:e011776.
12. Rimland JM, Abraha I, Luchetta ML, *et al*. Validation of chronic obstructive pulmonary disease (COPD) diagnoses in healthcare databases: a systematic review protocol. *BMJ Open* 2016;6:e011777.
13. Ferlay J, Soerjomataram I, Dikshit R, *et al*. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015;136:E359–E386.
14. Cheng TY, Cramb SM, Baade PD, *et al*. The international epidemiology of lung cancer: latest trends, disparities, and tumor characteristics. *J Thorac Oncol* 2016;11:1653–71.
15. Barni S, Maiello E, Di Maio M, *et al*. Adherence to AIOM (Italian Association of Medical Oncology) lung cancer guidelines in Italian clinical practice: Results from the RIGHT-3 (research for the identification of the most effective and highly accepted clinical guidelines for cancer treatment) study. *Lung Cancer* 2015;90:234–42.
16. Dubey AK, Gupta U, Jain S. Epidemiology of lung cancer and approaches for its prediction: a systematic review and analysis. *Chin J Cancer* 2016;35:71.
17. Sano M, Fushimi K. Association of palliative care consultation with reducing inpatient chemotherapy use in elderly patients with cancer in japan: analysis using a nationwide administrative database. *Am J Hosp Palliat Care* 2017;34:685–91.
18. Migliorino MR, Santo A, Romano G, *et al*. Economic burden of patients affected by non-small cell lung cancer (NSCLC): the LIFE study. *J Cancer Res Clin Oncol* 2017;143:783–91.
19. Scagliotti G, Nishio M, Satouchi M, *et al*. A phase 2 randomized study of TAS-102 versus topotecan or amrubicin in patients requiring second-line chemotherapy for small cell lung cancer refractory or sensitive to frontline platinum-based chemotherapy. *Lung Cancer* 2016;100:20–3.
20. Kudjawu YC, Chatellier G, Decool E, *et al*. Timing in initiating lung cancer treatment after bronchoscopy in France: Study from medico-administrative database. *Lung Cancer* 2016;95:44–50.
21. Abdulmalak C, Cottenet J, Beltramo G, *et al*. Haemoptysis in adults: a 5-year study using the French nationwide hospital administrative database. *Eur Respir J* 2015;46:503–11.
22. Busco S, Buzzoni C, Mallone S, *et al*. Italian cancer figures--Report 2015: the burden of rare cancers in Italy. *Epidemiol Prev* 2016;40:1–120.
23. Feliciano J, Gardner L, Hendrick F, *et al*. Assessing functional status and the survival benefit of chemotherapy for advanced non-small cell lung cancer using administrative claims data. *Lung Cancer* 2015;87:59–64.
24. Kunisawa S, Yamashita K, Ikai H, *et al*. Survival analyses of postoperative lung cancer patients: an investigation using Japanese administrative data. *Springerplus* 2014;3:217.
25. Karve SJ, Price GL, Davis KL, *et al*. Comparison of demographics, treatment patterns, health care utilization, and costs among elderly patients with extensive-stage small cell and metastatic non-small cell lung cancers. *BMC Health Serv Res* 2014;14:555.
26. Hassett MJ, Ritzwoller DP, Taback N, *et al*. Validating billing/encounter codes as indicators of lung, colorectal, breast, and prostate cancer recurrence using 2 large contemporary cohorts. *Med Care* 2014;52:e65–e73.
27. Nordstrom BL, Simeone JC, Malley KG, *et al*. Validation of claims algorithms for progression to metastatic cancer in patients with breast, non-small cell lung, and colorectal cancer. *Front Oncol* 2016;6.
28. Ramsey SD, Scoggins JF, Blough DK, *et al*. Sensitivity of administrative claims to identify incident cases of lung cancer: a comparison of 3 health plans. *J Manag Care Pharm* 2009;15:659–68.
29. Abraha I, Giovannini G, Serraino D, *et al*. Validity of breast, lung and colorectal cancer diagnoses in administrative databases: a systematic review protocol. *BMJ Open* 2016;6:e010409.
30. Penberthy L, McClish D, Manning C, *et al*. The added value of claims for cancer surveillance: results of varying case definitions. *Med Care* 2005;43:705–12.
31. Baldi I, Vicari P, Di Cuonzo D, *et al*. A high positive predictive value algorithm using hospital administrative data identified incident cancer cases. *J Clin Epidemiol* 2008;61:373–9.
32. Abraha I, Orso M, Grilli P, *et al*. The current state of validation of administrative healthcare databases in italy: a systematic review. *Int J Stat Med Res* 2014;3:309–20.

33 Orso M, Serraino D, Abraha I, *et al*. D.I.V.O. Group. Validating malignant melanoma ICD-9-CM codes in Umbria, ASL Napoli 3 Sud and Friuli Venezia Giulia administrative healthcare databases: a diagnostic accuracy study. *BMJ Open* 2018;8:e020631.

34 Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* 1927;22:209–12.

35 McClish DK, Penberthy L, Whittemore M, *et al*. Ability of medicare claims data and cancer registries to identify cancer cases and treatment. *Am J Epidemiol* 1997;145:227–33.

36 Koyi H, Johansson L, From J, *et al*. Biopsy testing in an inoperable, non-small cell lung cancer population-a retrospective, real-life study in Sweden. *J Thorac Dis* 2015;7:2226–33.

37 Koyi H, Hillerdal G, Brandén E. A prospective study of a total material of lung cancer from a county in Sweden 1997-1999: gender, symptoms, type, stage, and smoking habits. *Lung Cancer* 2002;36:9–14.

38 Benchimol EI, Manuel DG, To T, *et al*. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* 2011;64:821–9.

39 De Coster C, Quan H, Finlayson A, *et al*. Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium. *BMC Health Serv Res* 2006;6:77.

40 Bossuyt PM, Reitsma JB, Bruns DE, *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med* 2003;138:40–4.