


BMJ Open How to discriminate non-small cell lung cancer (NSCLC) cases from an Italian administrative database? A retrospective, secondary data use study for evaluating a novel algorithm performance

William Balzi,¹ Andrea Roncadori,¹ Valentina Danesi ¹, Ilaria Massa,¹ Silvia Manunta,¹ Nicola Gentili,¹ Angelo Delmonte,² Lucio Crinò,² Mattia Altini¹

To cite: Balzi W, Roncadori A, Danesi V, *et al*. How to discriminate non-small cell lung cancer (NSCLC) cases from an Italian administrative database? A retrospective, secondary data use study for evaluating a novel algorithm performance. *BMJ Open* 2021;**11**:e048188. doi:10.1136/bmjopen-2020-048188

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2020-048188>).

WB and AR contributed equally.

Received 29 December 2020
Accepted 09 September 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Outcome Research, Healthcare Administration, IRCCS Istituto Romagnolo per lo Studio dei Tumori (IRST) "Dino Amadori", Meldola, Emilia-Romagna, Italy
²Department of Medical Oncology, IRCCS Istituto Romagnolo per lo Studio dei Tumori (IRST) "Dino Amadori", Meldola, Emilia-Romagna, Italy

Correspondence to

Ing Valentina Danesi;
valentina.danesi@irst.emr.it

ABSTRACT

Objectives To evaluate an algorithm developed for identifying non-small cell lung cancer (NSCLC) candidates among patients with lung cancer with a diagnosis International Classification of Diseases: ninth revision (ICD-9) 162.x code in administrative databases. Algorithm could then be applied for identifying the NSCLC population in order to assess the appropriateness and quality of care of the NSCLC care pathway.

Design Algorithm discrimination capacity to select both NSCLC or non-NSCLC was carried out on a sample for which electronic health record (EHR) diagnosis was available. A bivariate frequency distribution and other measures were used to evaluate algorithm's performances. Associations between possible factors potentially affecting algorithm accuracy were investigated.

Setting Administrative databases used in a specific geographical area of Emilia-Romagna region, Italy.

Participants Algorithm was carried out on patients aged >18 years, with a lung cancer diagnosis from January to December 2017 and resident in Emilia-Romagna region who have been hospitalised at IRST or in one of the hospitals placed in the Forlì-Cesena area and for which EHR diagnosis data were available.

Outcome measures Overall accuracy, positive (PPV) and negative (NPV) predictive values, sensitivity and specificity, positive and negative likelihood ratios and diagnostic OR were calculated.

Results A total of 430 patients were identified as lung cancer cases based on ICD-9 diagnosis. Focusing on the total incident cases (n=314), the algorithm had an overall accuracy of 82.8% with a sensitivity of 88.8%. The analysis confirmed a high level of PPV (90.2%), but lower specificity (53.7%) and NPV (50%). Higher length of stay seemed to be associated with a correct classification. Hospitalisation regimen and a supply of antineoplastic therapy seemed to increase the level of PPV.

Conclusion The algorithm demonstrated a strong validity for identifying NSCLC among patients with lung cancer in hospital administrative databases and can be used to investigate the quality of cancer care for this population.

Trial registration number NCT04676321.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ The algorithm covers a medical need referring to a specific category within lung cancer: the non-small cell lung cancer histotype.
- ⇒ Algorithm discrimination capacity is assessed at an individual level verifying diagnosis on electronic health record (EHR) and is based on rigorously statistical procedures.
- ⇒ Incident and prevalent conditions are assumed as correctly identified, without a countercheck verification in EHR.
- ⇒ Generalisation of algorithm is limited to the variability of healthcare delivery settings among different international contexts and different local laws, regulations or customs.

INTRODUCTION

Lung cancer is a highly complex disease that causes a heavy burden on the healthcare system both for its frequency, complexity in clinical management and poor prognosis. In recent years, management of patients with non-small cell lung cancer (NSCLC) has been rapidly changing thanks to the availability of new drugs that challenges the sustainability of National Health Care Systems. In such a context, the availability of tools that allow to measure quality of care within the care pathway is crucial to understand the value of provided care, in terms of health outcomes achieved at the population level per amount of expenditure. With this in mind, it is necessary to build indicators to be applied to the healthcare pathway in order to monitor it continuously, and the first step in this way is to identify the correct population to be measured within the pathway. Using hospital administrative databases, which were designed primarily for accounting and management purposes can be helpful. These databases

include a combination of information such as hospital discharge cards, medical and diagnostics procedures, drug prescriptions and laboratory data. They can easily be used to identify diagnoses, treatments and outcomes, as they provide timely and easy access to an inexpensive and large source of knowledge regarding subjects in a defined geographical area. This is the reason why administrative data have been widely exploited in different types of epidemiological, postmarketing surveillance and outcome research studies.¹ However, the use of administrative data to unambiguously identify patients characteristics is still challenging, since administrative data are not as rich in clinical details as electronic health records (EHR). The International Classification of Diseases: ninth revision Clinical Modification (ICD-9-CM) codes in hospital discharge abstracts (HDA) is used to identify subjects with lung cancer. However, administrative databases are not suitable to identify NSCLC cases as histological classification is not deducible from ICD-9-CM codes.² Therefore, an optimal and precise NSCLC case identification is still challenging. And validated algorithms are needed in order to detect patients from administrative databases.^{3,4}

Currently, to our knowledge, there are only two Italian published studies which developed and validated an algorithm for identifying incident lung cancer cases without selection of different types of lung cancer.^{5,6} At European level, an incidence study to estimate the cost of NSCLC treatments has been completed in France, Germany and UK.⁷ Unfortunately, the selection of patients with NSCLC was performed without a procedure validation. Conversely, in the USA, several studies aimed to detect NSCLC cases from medical claims databases. Ramsey's study examined the sensitivity of different administrative claims data estimating an accuracy from 51.1% to 99.4% in identifying NSCLC incident cases.⁸ Unfortunately, specificity could not be estimated as any false positives in their data source (insurer data) was not included. Duh *et al* developed an algorithm in order to identify small cell lung cancer (SCLC) subjects; however, this study considered only stage IV cancer cases.⁹ In order to identify NSCLC, a modified version reversing the inclusion and exclusion criteria (only patients with metastatic cancer aged >65 years) was suggested.¹⁰ An algorithm developed by Turner *et al* had an accuracy of 92.1%, a sensitivity of 94.8% and a specificity of 81.1% were reported. Despite the excellent performance of the algorithm, the authors acknowledge the poor external validity of their results, mainly attributable to the commercial nature of the health insurance plans.¹¹ In general, the inclusion criteria of these algorithms contain procedures and chemotherapies recommended for patients with NSCLC whereas the exclusion criteria consist of treatment regimens applied to patients with SCLC.^{7,11,12}

As part of the 'KIND NSCLC study: Key Performance Indicators for the assessment of diagnostic and therapeutic pathway of NSCLC patients: a multicenter study' (Protocol Code: IRST162.13), selection algorithm of

patients with NSCLC was highly recommended. The KIND study (ClinicalTrials.gov NCT04676321) aimed to assess the appropriateness and quality of care in patients with NSCLC identified through administrative health data of three sites (hospitals of Modena, Reggio-Emilia and Forlì-Cesena provinces) placed in Emilia-Romagna region. Using the EHR as the gold standard, the primary aim of our study was to evaluate an algorithm to identify NSCLC incident cases from among a pool of patients with a primary or secondary diagnosis of ICD-9-CM 162.x code reported in the discharge cards. In addition, we also tried to identify possible factors (ie, not directly implied by the algorithm) potentially affecting algorithm accuracy.

METHODS

NSCLC KIND study population

The study population of NSCLC KIND study consisted of adult patients (aged ≥18 years) residing in Emilia-Romagna region, with a newly diagnosis of NSCLC between January and December 2017 identified in the hospital discharge cards (HDC) and who has been discharged in any one of the participating sites (hospitals of Modena, Reggio-Emilia and Forlì-Cesena provinces) (figure 1).

Setting and data source

Administrative databases of three specific geographical areas of Emilia-Romagna region were queried by an algorithm for identifying eligible study subject NSCLC KIND. The assignment of an anonymised patient identification code to all residents independently from type of admission (inpatient or outpatient) allows deterministic individual cross linked among different databases.

Data were retrieved from:

1. HDC for case selection and algorithm classification based on ICD-9-CM code. The HDC summarises information from clinical charts regarding type of discharge, primary diagnosis, up to five secondary diagnoses and eleven surgical, diagnostic or therapeutic procedures, codified according to the ICD-9-CM.
2. Pharmaceutical data (FED and AFT—direct and territorial distribution), such as Anatomical Therapeutic Chemical (ATC) drug classification and supply date, for determining histological diagnosis different from NSCLC based on treatment received.

Algorithm specification to identify eligible patients for NSCLC KIND study

Candidate patients with NSCLC were identified using the ICD-9-CM of malignant tumours of trachea, bronchi and lung (codes 162.x), as primary or secondary diagnosis from HDC (figure 1). For patients with more than one ICD-9-CM 162.x code, the patient index date was defined as the earliest date of year 2017 in which 162.x code appeared. Specific criteria on patient's cancer history and chemotherapy regimens were applied to both

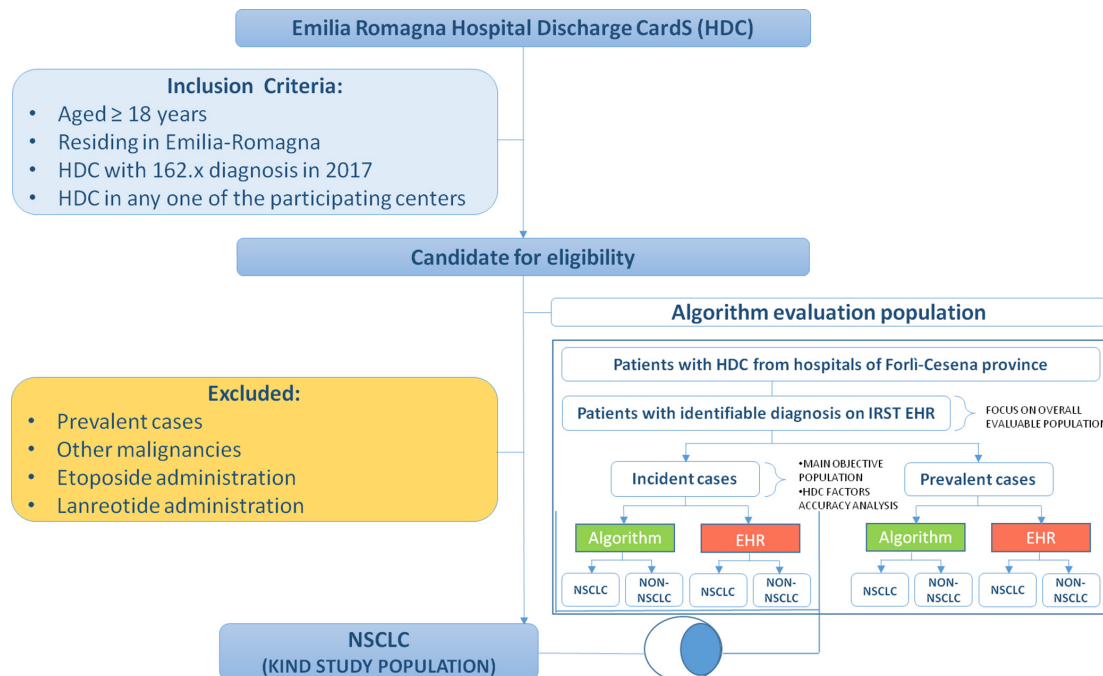


Figure 1 Flow chart for identification of eligible patients for non-small cell lung cancer (NSCLC) KIND study using International Classification of Diseases: ninth revision Clinical Modification (ICD-9-CM) 162.x diagnosis code from administrative databases. Algorithm discrimination capacity was evaluated on a sample for which electronic health records (EHR) was available. The patients hospitalised at IRST or in one of the two hospitals placed in our district between January and December 2017 were identified as lung cancer cases from administrative databases. The unique code reported in the hospital discharge cards (HDC) for each patient was matched to the individual tax code (present in EHR) in a deterministic way to identify patients (1:1 deterministic relation). The matching between the HDC code and the tax ID one was possible for patients, who had at least one access (even in an outpatient setting) in IRST Institute. The classification of incident and prevalent cases were identified by algorithm (no countercheck with the EHR).

discriminate incident cases from prevalent and to identify other malignancies (non-NSCLC) (figure 1):

- ▶ Patients who had the same ICD-9-CM diagnosis code 162.x recorded in the 3 years before the year under study.
- ▶ Patients who had other malignancies ICD-9-CM diagnosis (ICD-9-CM 140.x-161.x, 163.x-195.x, 200.x-208.x or V.10.xx except V10.11 and V10.12) recorded in the 3 years before the year under study.
- ▶ Patients with at least one therapy administration of Etoposide (code L01CB01 according to ATC classification system) in the following 180 days from the index date were classified as non-NSCLC.
- ▶ Patients with at least one therapy administration of Lanreotide (ATC code H01CB03) and/or Octreotide (ATC code H01CB02) in the following 180 days from the index date were also identified as non-NSCLC.

Algorithm evaluation population

Algorithm discrimination capacity was evaluated only on a restricted sample of patients for which EHR was available. Evaluation of the algorithm was carried out only on patients hospitalised (at least once time) at IRST or in one of the hospitals placed in the Forlì-Cesena area (M. Bufalini of Cesena, G.B. Morgagni—L. Pierantoni of Forlì) with a verifiable diagnosis through EHR (figure 1). The evaluation of the algorithm was assessed at an

individual level by linkage of cases identified by the algorithm (both NSCLC or non-NSCLC) to cases in the EHR.

Statistical analysis

Bivariate frequency distribution of algorithm classified cases (test) and EHR verified cases (gold standard) results were produced for each analysis set and presented as 2×2 tables reporting the number of true positive (TP), false positive (FP), false negative (FN) and true negative (TN). The analysis estimated sensitivity and specificity, with their corresponding 95% CI. Positive predicting values (PPV) and negative predicting value (NPV) were also determined, along with their 95% CIs. The overall accuracy, expressed as the proportion of correctly classified subjects (TP +TN) among all subjects was therefore established. ROC curves were drawn and area under the curve (AUC) were also estimated to estimate accuracy. Moreover, both positive (LR+) and negative (LR-) likelihood ratios, defined as the ratio of the probability of an expected test result in subjects with the disease to the probability in the subjects without the disease, were calculated. Lastly, we determined the ratio of the odds of positivity in subjects with the disease to the odds in subjects without the disease known as the diagnostic OR (DOR). In case of doubtful diagnosis in the EHR, we adopted a conservative approach considering cases as non-NSCLC. Univariate multinomial logistic regression models were

developed to further explore the associations between HDC information and algorithm classification correctness; each record has been classified in four categories: correctly or wrongly detected as NSCLC and correctly or wrongly identified as non-NSCLC. ORs with their 95% CIs were calculated separately for false NSCLC, false non-NSCLC (other) and true other and compared with true NSCLC (reference group). A stepwise approach was used for regression selection in multivariate analyses. Among factors included in the final multivariable model, both correlation and variance inflation factor (VIF) have been calculated to assess the presence of multicollinearity issues. Analysis of data was performed using R statistical software (www.r-project.org) V.3.6.3.

Patient and public involvement

No patient involved.

RESULTS

Data from 430 patients with an HDC in which ICD-9-CM diagnosis code 162.x were collected and for which a verification of diagnosis data on IRST's EHR was available, were included in the sample population. Among the overall sample, 314 patients were identified as incident cases during 2017 (no previous malignancies diagnosis during the previous 3 years) and considered for the main evaluation. However, a secondary analysis on broader populations (eg, including further 116 prevalent cases) was also performed. Focusing on the total incident cases (N=314) as shown in [table 1](#), among the 256 cases classified by the algorithm as NSCLC: 231 were confirmed to have NSCLC (TP), whereas 25 had different diagnoses (FP) resulting in a considerable 90.2% PPV (95% CI 86.6% to 93.9%). Looking at the cases classified by the algorithm

as non-NSCLC: 29 cases (TN) were correctly classified by the algorithm in non-NSCLC group leading to a 50.0% NPV (95% CI 37.1% to 62.9%). Since NSCLC represents the vast majority of lung cancers, and being both PPV and NPV strongly dependent on the disease prevalence in the study population, the high PPV strongly contributed to obtain a remarkable overall accuracy of 82.8%. The algorithm reached a very high level of sensitivity, 88.8% (95% CI 85.0% to 92.7%), on the contrary, a lower specificity was observed (53.7%; 95% CI 40.4% to 67.0%), leading to an AUC estimate of 71.3% (95% CI 64.3% to 78.3%). In this context, the likelihood ratios is interesting: LR+ is 1.92 (95% CI 1.43 to 2.57), while LR- equals 0.21 (95% CI 0.14 to 0.32) resulting in a DOR of 9.23 (95% CI 4.53 to 18.87).

With the purpose of identifying factors affecting algorithm accuracy, we developed a univariate multinomial logistic regression model for available variables collected in the HDC forms ([table 2](#)). Among EHR confirmed NSCLC (TP and FN), higher length of stay seemed to be associated with correct classification (OR 0.538; 95% CI 0.288 to 1.005) even if only a slight statistical significance was observed ($p=0.052$), while patients older than 75 years at hospital admission showed a greater risk of misclassification (OR 2.285; 95% CI 1.045 to 4.993; $p=0.038$). Among HDC forms the algorithm classified as NSCLC, at least 1 day waiting for admission (OR 0.385; 95% CI 0.167 to 0.887; $p=0.025$), ordinary regime hospitalizations—vs day hospital—(OR 0.354; 95% CI 0.142 to 0.879; $p=0.025$), DRG lung disease-specific or oncological treatment related (OR 0.384; 95% CI 0.159 to 0.928; $p=0.034$) and oncological treatment (ie, ATC L0) during 2017 (OR 0.124; 95% CI 0.036 to 0.426; $p=0.001$) resulted in protective factors for misclassification (ie, to be false NSCLC, FP). Lastly, looking at the correctly assigned records (TP and TN), patients discharged at home are more likely to be true NSCLC, while patients with at least one prescribed antineoplastic therapy during the year are 'at higher risk' to be correctly identified as non-NSCLC. We, moreover, developed a multivariable model to try to understand which are the most important predictors of algorithm reliability ([table 3](#)), length of stay, ordinary regime hospitalisations, discharges at home and oncological treatment during 2017 were the variables selected in the model adopting a stepwise approach (based on AIC), and no multicollinearity issues were found (see online supplemental tables 1 and 2). Lastly, to account for different proportion of surgical DRG in the day hospital and in the inpatient regime, we conducted a multinomial regression analysis on the ordinary regime only (inpatient) which showed a slight statistical significance (HR=0.509; $p=0.056$) for the variable LoS on increasing the sensitivity of the algorithm (see online supplemental tables 3 and 4). Based on this results, we looked at the case-identification correctness among patients hospitalised for at least 5 days within an ordinary regime (we did not consider neither the discharge type which was associated with the reduced risk of true negativity to be

Table 1 Bivariate frequency distribution of algorithm classified cases and electronic health records (EHR) verified cases

		EHR verified		
		NSCLC	Other	Total
(A)				
Algorithm classification	NSCLC	231	25	256
	Other	29	29	58
	Total	260	54	314
(B)				
Algorithm classification	NSCLC	305	40	45
	Others	46	39	85
	Total	351	79	430

Analysis reported in panel A was performed on the sample identified by the algorithm as incident cases (N=314). Conversely, a secondary analysis on broader populations (N=430) which included both incidents and further 116 prevalent cases was reported in panel B.
NSCLC, non-small cell lung cancer.

Table 2 Univariate multinomial logistic regression model for available variables collected in the HDC forms to detect factors which influenced accuracy of the algorithm

Factors	TP versus FN			TP versus FP			TP versus TN		
	OR	95% CI	P value	OR	95% CI	P value	OR	95% CI	P value
Sex	1.235	0.549 to 2.776	0.610	0.827	0.360 to 1.902	0.655	0.800	0.367 to 1.742	0.570
Age >75	2.285	1.045 to 4.993	0.038	1.152	0.474 to 2.797	0.755	1.101	0.477 to 2.542	0.821
waiting for admission (at least 1 day)	1.311	0.535 to 3.211	0.554	0.385	0.167 to 0.887	0.025	1.095	0.462 to 2.593	0.837
Length of stay (each 5 days)	0.538	0.288 to 1.005	0.052	1.116	0.872 to 1.428	0.384	1.060	0.825 to 1.362	0.650
Hospitalisation regime (ordinary)	1.119	0.515 to 2.431	0.777	0.354	0.142 to 0.879	0.025	1.489	0.673 to 3.289	0.326
Discharged at home	1.638	0.367 to 7.306	0.518	0.384	0.140 to 1.052	0.063	0.319	0.128 to 0.795	0.014
Diagnosis Related Groups (DRG) type (medical vs surgical)	1.545	0.711 to 3.355	0.272	1.529	0.667 to 3.499	0.316	1.345	0.617 to 2.930	0.456
DRG lung disease-specific or oncologic treatment related	1.036	0.373 to 2.876	0.946	0.384	0.159 to 0.928	0.034	1.349	0.445 to 4.086	0.597
Oncological treatment during 2017 (ATC L0)	0.848	0.392 to 1.838	0.677	0.124	0.036 to 0.426	0.001	2.857	1.175 to 6.950	0.021

ATC, Anatomical Therapeutic Chemical; FN, false negative; FP, false positive; HDC, hospital discharge cards; TN, true negative; TP, true positive.

considered in the context of correct selection, nor the presence of prescription of antitubercular therapy which was a discordant factor resulting protective for false positivity, and simultaneously risky for true negativity): a 93.2% accuracy was achieved in 44 cases, which rose to 100% by eliminating additional 30 patients with hospitalisations of exactly 5 days.

Since the algorithm can be considered as a composite process of multiple steps, we also assess its performance

when skipping the prevalent cases removal phase (table 1—panel B): less than 3% of overall accuracy was lost (80.0%) including further 116 prevalent cases (AUC: 68.1%—95% CI 62.3% to 74.0%). The most relevant aspect is perhaps the poor ability of the algorithm to correctly identify the other lung tumours when the prevalent cases are also considered in the analysis (specificity: 49.4%—95% CI 38.3% to 60.4% and npv: 45.9%—95% CI 35.3% to 56.5%).

Table 3 Multivariable multinomial logistic regression model for available variables collected in the HDC forms to detect which are the most important predictors of algorithm reliability

Factors	TP versus FN			TP versus FP			TP versus TN		
	OR	95% CI	P value	OR	95% CI	P value	OR	95% CI	P value
Length of stay (each 5 days)	0.444	0.214 to 0.922	0.029	0.885	0.629 to 1.243	0.480	1.148	0.858 to 1.538	0.353
Hospitalisation regime (ordinary)	0.598	0.254 to 1.409	0.240	0.366	0.132 to 1.017	0.054	2.126	0.765 to 5.904	0.148
Discharged at home	1.283	0.279 to 5.895	0.749	0.577	0.184 to 1.810	0.345	0.227	0.083 to 0.623	0.004
Oncological treatment during 2017 (ATC L0)	0.784	0.358 to 1.719	0.544	0.138	0.040 to 0.479	0.002	3.536	1.370 to 9.129	0.009

ATC, Anatomical Therapeutic Chemical; FN, false negative; FP, false positive; HDC, hospital discharge cards; TN, true negative; TP, true positive.

DISCUSSION

In order to perform the KIND NSCLC study, with the objective to assess the diagnostic and therapeutic pathway of patients with NSCLC, we developed an algorithm designated to identify NSCLC from hospital administrative databases and we tested its performance, using EHR review as gold standard. While previous Italian studies focused on patients with lung cancer, this research covers a medical need referring to a specific category within lung cancer: the NSCLC histotype. Results showed a high level of accuracy of the algorithm (82.8%) and moreover sensitivity (88.8%). Findings confirmed that the algorithm reached a high level of PPV for identifying NSCLC (90.2%), but modest specificity (53.7%) and NPV (50%). The main reason of modest specificity is due to a misclassification of clinical diagnoses and coding errors in HDC. Furthermore, it should be noted that AUC is not much above 0.7, which is slightly above the cut-offs usually reported in literature for indicating good algorithm's performance.¹³ Similar considerations must be made for likelihood ratios. Our findings are consistent with previous studies, where sensitivity ranged between 51.1% and 99.4%,^{8 12 14} and PPV value was 95.3%.¹² This slight variability could be explained considering that patients identification is affected by: the data quality of the database used (which is likely to be more complete and accurate in the US payer insurance databases) and by the different criteria used for the algorithm (often based on pharmaceutical claims). Several studies reported evaluations of NSCLC case selection algorithms based only on drug prescription databases, which however, not allowing to identify untreated cases, by providing an unreal picture of the NSCLC population.¹² A fundamental limit that we tried to overcome with the multiple data source method assessed in this study. The study by Turner *et al*, reported an accuracy of 94.8% with a PPV of 95.3%, even if patients generally ineligible to undergo antineoplastic therapy, namely early stages NSCLC and unfit patients, are unlikely to be selected.¹² The choice to select patients starting from discharge cards made us lose accuracy, although well over 80% while maintaining very high levels of both PPV (90.3%) and sensitivity (88.9%), mostly at the expense of specificity and NPV which, however, is counterbalanced by the low prevalence of other malignancies among those of the lung. Moreover, in accordance with Italy and Emilia-Romagna region's regulations, drugs in off-label use or antineoplastic therapies administered to patients participating in clinical trials are not tracked in the administrative data flows, causing a potential loss of many cases. An additional objective of this study was the identification of the factors influencing the accuracy of the algorithm, providing a clinical (and/or administrative) interpretation of these factors. These results can be transposed in other contexts and give an a-priori estimation of the algorithm accuracy applying their setting characteristics. When evaluating the hospitalisation regimen (ordinary vs day hospital), we can deduce its influence on PPV, in particular we observe how

day hospital setting increases the risk of false positivity: we know that in the Emilia-Romagna region, chemotherapy administration were done on an outpatient basis (day service), while day-hospital is used for some invasive diagnostic services (eg, biopsy). In such cases, secondary lesions investigation is not uncommon to assess the presence of metastases, but also to investigate the presence of any other primary malignancy. The associated diagnostic code may therefore have been incorrect. Anyway, the loss to follow-up attributable misclassification is the most frequent: four out of seven wrongly NSCLC classified with DH regimen hospitalisation were patients who decided to be followed at other institutions and for whom the last available and verifiable diagnosis in EHR was not yet certain and, doubtful diagnosis in EHR were considered actual non-NSCLC. The presence of oncological treatment (ie, ATC L0) led to similar results: excluding patients with specific treatments for other neoplasms (eg, neuroendocrine tumour and SCLC) which were classified as non-NSCLC (and for which it is strongly 'risky' for correct classification as true non-NSCLC), the PPV seems to increase for patients with a supply of antineoplastic therapy (during the year). The incoming of new therapies in the treatment landscape of lung cancer could influence the algorithm performance, the introduction of new drugs specific for the treatment of the SCLC or neuroendocrine cancer would allow the algorithm to better discriminate NSCLC from other lung malignancies. Conversely, in case of new drugs with therapeutic indications for the treatment of both NSCLC and non-NSCLC, the performance of the algorithm could be negatively affected. This analysis, moreover, suggests that a higher length of stay (>5 days) may be an important factor associated with correct classification. This is probably linked with the timing of histological diagnosis: longer hospitalisations are more likely to have histological diagnosis available before HDC completion, resulting in less codification errors. Reflection is required by observing the reduced sensitivity effect of the algorithm among patients aged >75 years (univariate logistic regression model): elderly patients may receive less specific treatments, which makes administrative data less precise for the algorithm's purpose. Additionally, elderly are more likely to have more comorbidities than their younger counterparts. Synchronous pathologies may worsen the clinical picture and require additional treatments which must be reported in the HDC (as they often absorb many resources). As a result, HDCs are missing information that would be useful for histology discrimination.

Limitations

The main limitation of the study, as all the other studies conducted on administrative databases (including pharmaceutical claims DBs), is due to the variability of healthcare delivery settings among different international contexts, but also, sometimes, due to different local laws, regulations or customs: in some contexts, the delivery of drugs is allowed only on an outpatient setting (less hospital

discharge cards), as well as some diagnostic procedures, generating heterogeneity in the data sources of the algorithm. Actually, the intent of multinomial logistic models was precisely to give an idea of the levels of accuracy that can be achieved even in application contexts other than ours. Although we only collected data from a few geographical areas of the Emilia-Romagna region, the same administrative data are available nationwide. The algorithm may therefore be used to estimate the national incidence of NSCLC.

Another limitation of this study may be to assume incident and prevalent cases as correctly identified (no countercheck with the EHR). Nevertheless, we are quite sure about prevalent classification correctness because of an oncological diagnosis in the previous 3 years. This reasonable confidence decreases for incident cases, although no hospitalisation in the previous 3 years for a patient with a malignant disease is highly unlikely.

CONCLUSION

In summary, the results of this study demonstrate that our algorithm may be useful for identifying newly diagnosed patients with NSCLC in hospital administrative databases. Thanks to the widespread use of these databases, the assessment of the performance on NSCLC care pathway, applying to the identified population a set of KPIs, could be feasible everywhere in Italy, not only for a direct measurement of the patient journey, but also for a benchmark process between different hospitals that could help to improve quality of care for patients.

Twitter Nicola Gentili @nicolagentili

Contributors WB, AR, VD, IM, MA: contributed to the design of the algorithm, assisted by AD and LC. WB: developed the algorithm and contributed to collection and assembly of data. AR: led statistical analysis. NG: contributed to analyse the data. WB, AR, VD, IM, AD, LC and MA contributed to interpretation of findings. SM: verified the cases on EHR. AR and VD drafted the manuscript, supervised by IM and all authors critically revised the work and approved the final manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This retrospective study was approved by the Scientific and Medical Committee and the Ethics Committee (EC) of IRST-IRCCS Area Vasta Romagna (CEROM) and of Area Vasta Emilia Nord (AVEN). The approval number/ID was Prot. 1383/2020 1.5/233 for the CEROM and Prot. 2020/0104403 and AOU 0019762/20 for the AVEN.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The anonymized datasets generated and/or analyzed during the current study are available from the corresponding author (valentina.danesi@irst.emr.it) on reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Valentina Danesi <http://orcid.org/0000-0001-9261-6467>

REFERENCES

- Schulman KL, Berenson K, Tina Shih Y-C, *et al*. A checklist for ascertaining study cohorts in oncology health services research using secondary data: report of the ISPOR oncology good outcomes research practices Working group. *Value Health* 2013;16:655–69.
- Whittle J, Steinberg EP, Anderson GF, *et al*. Accuracy of Medicare claims data for estimation of cancer incidence and resection rates among elderly Americans. *Med Care* 1991;29:1226–36.
- West S, Strom BL, Poole C. Validity of pharmacoepidemiology drug and diagnosis data. In: Strom BL, ed. *Pharmacoepidemiology*. 3rd edn. West Sussex, UK: John Wiley & Sons Ltd, 2000: 661–705.
- Burns EM, Rigby E, Mamidanna R, *et al*. Systematic review of discharge coding accuracy. *J Public Health* 2012;34:138–48.
- 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.
- 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.
- McGuire A, Martin M, Lenz C, *et al*. Treatment cost of non-small cell lung cancer in three European countries: comparisons across France, Germany, and England using administrative databases. *J Med Econ* 2015;18:525–32.
- 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.
- Duh MS, Reynolds Weiner J, Lefebvre P, *et al*. Costs associated with intravenous chemotherapy administration in patients with small cell lung cancer: a retrospective claims database analysis. *Curr Med Res Opin* 2008;24:967–74.
- 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.
- Turner RM, Chen Y-W, Fernandes AW. Validation of a case-finding algorithm for identifying patients with non-small cell lung cancer (NSCLC) in administrative claims databases. *Front Pharmacol* 2017;8:1–8.
- Fernandes AW, Wu B, Turner RM. Brain metastases in non-small cell lung cancer patients on epidermal growth factor receptor tyrosine kinase inhibitors: symptom and economic burden. *J Med Econ* 2017;20:1136–47.
- Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;240:1285–93.
- Whyte JL, Engel-Nitz NM, Teitelbaum A, *et al*. An evaluation of algorithms for identifying metastatic breast, lung, or colorectal cancer in administrative claims data. *Med Care* 2015;53:e49–57.