BMJ Open

Validity of breast, lung and colorectal cancer diagnoses in administrative databases: a systematic review protocol

Iosief Abraha,[1] Gianni Giovannini,[1] Diego Serraino,[2] Mario Fusco,[3] Alessandro Montedori[1]

To cite: 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. doi:10.1136/bmjopen-2015-010409

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2015-010409).

Received 30 October 2015
Revised 1 February 2016
Accepted 10 February 2016

[1]Health Planning Service, Regional Health Authority of Umbria, Perugia, Italy
[2]Epidemiology and Biostatistic Unit, IRCCS Centro di Riferimento Oncologico Aviano, Aviano, Italy
[3]Registro Tumori Regione Campania, ASL NA3 Sud, Brusciano (Na), Italy

Correspondence to
Dr Iosief Abraha;
iosief_a@yahoo.it

ABSTRACT

Introduction: Breast, lung and colorectal cancers constitute the most common cancers worldwide and their epidemiology, related health outcomes and quality indicators can be studied using administrative healthcare databases. To constitute a reliable source for research, administrative healthcare databases need to be validated. The aim of this protocol is to perform the first systematic review of studies reporting the validation of International Classification of Diseases 9th and 10th revision codes to identify breast, lung and colorectal cancer diagnoses in administrative healthcare databases.

Methods and analysis: This review protocol has been developed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocol (PRISMA-P) 2015 statement. We will search the following databases: MEDLINE, EMBASE, Web of Science and the Cochrane Library, using appropriate search strategies. We will include validation studies that used administrative data to identify breast, lung and colorectal cancer diagnoses or studies that evaluated the validity of breast, lung and colorectal cancer codes in administrative data. The following inclusion criteria will be used: (1) the presence of a reference standard case definition for the disease of interest; (2) the presence of at least one test measure (eg, sensitivity, positive predictive values, etc) and (3) the use of data source from an administrative database. Pairs of reviewers will independently abstract data using standardised forms and will assess quality using a checklist based on the Standards for Reporting of Diagnostic accuracy (STARD) criteria.

Ethics and dissemination: Ethics approval is not required. We will submit results of this study to a peer-reviewed journal for publication. The results will serve as a guide to identify appropriate case definitions and algorithms of breast, lung and colorectal cancers for researchers involved in validating administrative healthcare databases as well as for outcome research on these conditions that used administrative healthcare databases.

Trial registration number: CRD42015026881.

Strengths and limitations of this study

- Validation of International Classification of Diseases 9th and 10th revision (ICD-9 and ICD-10) diagnosis codes for breast, lung and colorectal cancers, using administrative healthcare databases, can contribute to health outcome research.
- This review will be the first to systematically identify and evaluate primary studies that validated the accuracy of administrative healthcare databases with ICD-9 and ICD-10 codes related to breast, lung and colorectal cancers.
- The results will serve as a guide to identify appropriate case definitions and algorithms of breast, lung and colorectal cancers for researchers involved in validating administrative healthcare databases.

INTRODUCTION

The burden of cancer is increasingly growing among populations and is associated with major economic expenditure, especially in developed countries.[1] While breast cancer is the most common cancer and the leading cause of cancer death in women, lung cancer is the leading cause of cancer deaths in men and the second leading cause of cancer deaths in women.[2] On the contrary, colorectal cancer is the third most common cancer in men and the second in women, worldwide. Overall, breast, lung and colorectal cancers account for 34% of all neoplasms around the globe.[3]

Administrative healthcare databases are increasingly being used for epidemiological evaluation in oncology,[4] population outcome research,[5] drug utilisation reviews,[6–8] evaluation of health service delivery and quality,[9 10] as well as for health policy development.[11–13] Generally, these databases gather longitudinal information concerning health resource utilisation regarding hospitalisations, outpatient care and often, drug prescriptions and vital statistics.[14] The use of these databases allows for more efficient analyses and unlike randomised trials, their representativeness of routine clinical practice in large populations can provide more generalisable findings.[15] By definition, administrative healthcare databases are those in which data are routinely and

passively collected without an a priori research question, as they are usually established for billing or in general, administrative purposes and not for research uses. Hence the diagnostic codes used to identify, for example, cancers, must be validated according to an accepted 'reference standard' reference diagnosis.[16]

The current International Classification of Diseases, 9th revision (ICD-9) codes are 233.0 and 174.0–174.9 for breast cancer, 162.0–162.9 for lung cancer, and 153.0–153.9 and 154.8 for colorectal cancer. The ICD-10 codes are D05 and C50 for breast cancer, C34 for lung cancer and C18-C20 for colorectal cancer. Generally, the present diagnostic criteria for breast, lung and colorectal cancers rely on histological examinations and radiological analyses can contribute to staging. A number of different claim-based algorithms have been proposed for case identification of breast, lung and colorectal cancers, such as a combination of healthcare claims data,[17] the use of chemotherapy[18] and the number of medical claims on separate dates.[11] In addition, since patients with metastatic cancer have different prognoses and typically different treatment patterns to those with earlier-stage malignancies, researchers suggest using algorithms to identify patients with metastatic cancer.[11 19]

To the best of our knowledge, data on the validity of breast, lung and colorectal cancer diagnosis codes have not been synthesised in the medical literature. With the present protocol, we express our aim to systematically evaluate validation studies of administrative data algorithms identifying these cancers in administrative databases. Relevant studies will be those that coded in a sample population with breast, lung or colorectal cancer, using a medical chart as a reference standard and evaluated the accuracy of the validated ICD-9 or ICD-10 codes related to the cancer diseases.

Research question

The primary research question is 'What is the accuracy of administrative data algorithms related to breast, lung and colorectal cancers in administrative databases for correctly identifying the respective diseases?'. The target populations are patients with primary diagnosis of breast, lung or colorectal cancer; the index test will be represented by administrative data algorithms related to breast, lung and colorectal cancers and the reference standard will be represented by medical charts, validated electronic health records (EHRs) or cancer registries. Our primary outcome is the accuracy (expressed in terms of sensitivity, specificity and positive and negative predictive values) of administrative data algorithms in discriminating cases of breast, lung and colorectal cancer diseases.

METHODS

Literature search

Comprehensive searches of MEDLINE, EMBASE, Web of Science and the Cochrane Library, from their inception, will be performed to identify published peer-reviewed literature. We will employ a search strategy that we developed based on the combination of: (1) keywords and MeSH terms to identify records concerning breast, lung and colorectal cancers; (2) terms to identify studies likely to contain validity or accuracy measures and (3) a search strategy designed to capture studies that use healthcare administrative databases based on the combination of terms used by Benchimol *et al*[20] and the Mini-Sentinel's program.[21 22] The developed search strategy is available as online supplementary appendix 1. To retrieve additional articles, the authors will hand search relevant reference lists of key articles. We will also use the 'Cited-By' tools in PubMed and Google Scholar to find relevant articles that cited the article of interest, identified through the aforementioned search strategy. Titles and abstracts will be screened for eligibility by two independent reviewers. Discrepancies will be solved by discussion.

This review protocol is prepared according to the Preferred Reporting Items for Systematic reviews and Meta-Analysis Protocols (PRISMA-P) 2015 statement[23] and the results will be presented following the PRISMA flow diagram (figure 1).[24] This protocol has also been published in the PROSPERO International Prospective Register of systematic reviews with registration number CRD42015026881 (http://www.crd.york.ac.uk/PROSPERO).

Inclusion criteria

Full texts of eligible peer-reviewed articles that used administrative data for the ICD-9 or ICD-10 codes related to breast, lung and colorectal cancer diagnoses, without publication date restriction, published in English, will be obtained. For each study, the following inclusion criteria will be applied: (1) the presence of a reference standard case definition for the cancer of interest; (2) the presence of at least one test measure (eg, sensitivity, positive predictive values (PPVs), etc); (3) the data source was from an administrative database (ie, a database in which data is routinely and passively collected without a priori research question) and (4) the study database was from a representative sample of the general population.

We aim to focus on primary diagnosis of cancers, hence studies that considered algorithms to identify cancer history, cancer progression or recurrence will not be evaluated.

Studies that used EHRs to validate the disease of our interest will also be included. An EHR consists of a digital file used by healthcare providers for patient care and generally it includes clinical notes, prescription records and radiology and laboratory data.[25] Similar to most administrative databases, EHRs are not established for research purposes.[26] However, studies that used validated EHRs as a reference standard will be considered in our evaluation.

In addition, studies that employ databases that were not truly administrative (eg, cancer registries, epidemiology surveillance systems, etc) will be excluded.

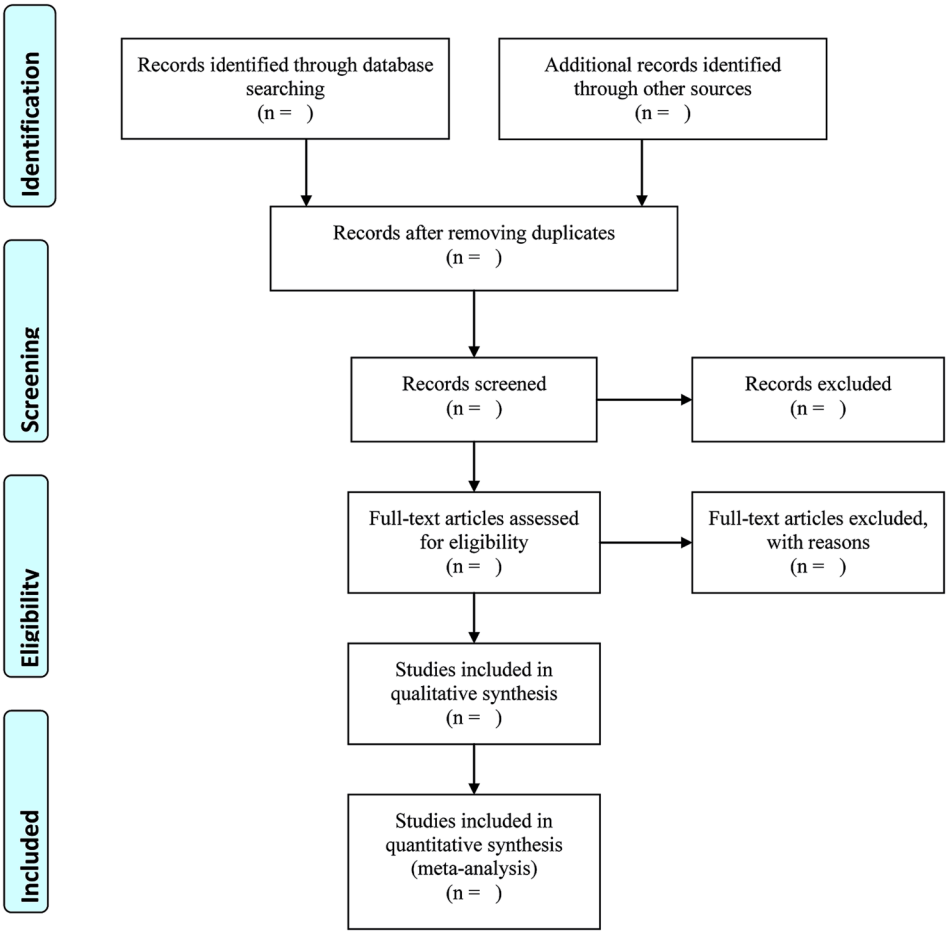

Figure 1 Study screening process.

Selection process

At the initial stage, titles and abstracts will be screened for potentially eligible studies. Subsequently, full texts of articles will be obtained and assessed to determine if they meet the inclusion and exclusion criteria. We will conduct data abstraction using standardised data collection forms that will be first tested on a sample of eligible articles. Two review authors working independently and in tandem will be involved in title and abstract screening, full-text screening and data abstraction. Any discrepancies will be resolved by consensus and where necessary, with the involvement of a third review author. Calibration exercises will be performed at each level of the process.

Data extraction

Data extraction will include the following information:

A. The details of the included study (including title, year and journal of publication, country of origin and sources of funding; the first author will be used as the study identification);
B. The disease of interest (breast, lung and colorectal cancer);
C. The target population from which the administrative data were collected;
D. The type of administrative database used (eg, hospitalisation discharge data), outpatient records (eg, physician billing claims), etc;
E. The ICD-9 or ICD-10 code used or the administrative data algorithms tested these may include current procedural terminology; prescription fills, timing of diagnosis, etc);
F. The modality of development of the algorithm (eg, using classification and regression trees, logistic regression, expert opinion, etc);
G. External validation;
H. Use of training and testing cohorts;
I. The reference standard used to determine the validity of the diagnostic code (eg, medical chart review, patient self-reports, cancer registry, etc);
J. The characteristic of the test used to determine the validity of the diagnostic code or algorithm (eg, sensitivity, specificity, PPVs and negative predictive values (NPVs), area under the receiver operating characteristic curve, likelihood ratios and κ statistics);
K. Any funding source and conflict of interest.

Quality assessment

The design and method of the included primary studies will be assessed using a checklist developed by Benchimol *et al*,[20] based on the criteria published by the Standards for Reporting of Diagnostic accuracy initiative for the accurate reporting of studies using diagnostic studies.[27] The checklist is provided in online supplementary appendix 2. The presence of potential

biases within the studies will be reported in a descriptive way. Neither subgroup analysis nor publication bias assessment is anticipated.

Analysis

For each algorithm, we will abstract the performance statistics provided in the included studies. Validation statistics may include sensitivity, specificity, PPV and NPV. Sensitivity measures the degree to which an ICD-9 or ICD-10 code (eg, ICD-9 174) correctly identifies individuals who possess the characteristic of interest (ie, breast cancer) in the source used as a reference standard (eg, medical chart).[16] We will calculate 95% CI when they are not reported in the articles. Where possible, we will calculate PPV and NPV will be calculated if not reported.

PPV is the number of true positives divided by the total number of cases receiving the code and expresses the likelihood that the code corresponds to a true-positive case. NPV is the number of true negatives divided by the total number of cases without the code of interest and expresses the likelihood that the absence of the code corresponds to a true-negative case.

Where possible, validation statistics will be aggregated and stratified by administrative data source (outpatient vs inpatient data), type of ICD code (ICD-9 or ICD-10), stage of disease and country of origin.

ETHICS AND DISSEMINATION

This review protocol will use publicly available data without directly involving human participants, hence approval from an ethics committee is not required. An outline of the protocol is published in the PROSPERO International Prospective Register of Systematic Reviews in 2015, registration number CRD42015026881. The results will summarise the studies' validating diagnostic codes that identify breast, lung and colorectal cancers in administrative data. The results will serve as a guide to identify appropriate case definitions and algorithms of breast, lung and colorectal cancers for researchers involved in validating administrative healthcare databases as well as for outcome research on these conditions that used administrative healthcare databases.

Findings of the review will be presented at relevant scientific conferences and disseminated through publication in a peer-reviewed journal.

Contributors IA, GG, AM, MF and DS conceived the study and were responsible for designing the protocol. IA and AM drafted the protocol manuscript and developed the search strategy. IA, GG, AM, MF and DS critically revised the successive versions of the manuscript and approved the final version.

Funding This systematic review protocol was developed within the D.I.V.O. project (*Realizzazione di un **D**atabase **I**nterregionale **V**alidato per l'**O**ncologia quale strumento di valutazione di impatto e di appropriatezza delle attivitá 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.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The findings of this systematic review will be disseminated via peer-reviewed publications and conference presentations. All the data will be available from IA.

REFERENCES

1. Sullivan R, Peppercorn J, Sikora K, *et al*. Delivering affordable cancer care in high-income countries. *Lancet Oncol* 2011;12:933–80.
2. Jemal A, Center MM, DeSantis C, *et al*. Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiol Biomarkers Prev* 2010;19:1893–907.
3. Ferlay J, Shin HR, Bray F, *et al*. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer* 2010;127:2893–917.
4. Chen HF, Liu MD, Chen P, *et al*. Risks of breast and endometrial cancer in women with diabetes: a population-based cohort study. *PLoS ONE* 2013;8:e67420.
5. Escribà JM, Pareja L, Esteban L, *et al*. Trends in the surgical procedures of women with incident breast cancer in Catalonia, Spain, over a 7-year period (2005–2011). *BMC Res Notes* 2014;7:587.
6. Montero AJ, Eapen S, Gorin B, *et al*. The economic burden of metastatic breast cancer: a U.S. managed care perspective. *Breast Cancer Res Treat* 2012;134:815–22.
7. Earle CC, Venditti LN, Neumann PJ, *et al*. Who gets chemotherapy for metastatic lung cancer? *Chest* 2000;117:1239–46.
8. Song X, Zhao Z, Barber B, *et al*. Treatment patterns and metastasectomy among mCRC patients receiving chemotherapy and biologics. *Curr Med Res Opin* 2011;27:123–30.
9. Vachon B, Désorcy B, Gaboury I, *et al*. Combining administrative data feedback, reflection and action planning to engage primary care professionals in quality improvement: qualitative assessment of short term program outcomes. *BMC Health Serv Res* 2015;15:391.
10. Yuen E, Louis D, Cisbani L, *et al*. Using administrative data to identify and stage breast cancer cases: implications for assessing quality of care. *Tumori* 2011;97:428–35.
11. 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.
12. Bremner KE, Krahn MD, Warren JL, *et al*. An international comparison of costs of end-of-life care for advanced lung cancer patients using health administrative data. *Palliat Med* 2015;29:918–28.
13. Mittmann N, Liu N, Porter J, *et al*. Utilization and costs of home care for patients with colorectal cancer: a population-based study. *CMAJ Open* 2014;2:E11–17.
14. 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.
15. Schulman KL, Berenson K, Tina Shih YC, *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.
16. West SL, Strom BL, Poole C. *Validity of pharmacoepidemiologic drug and diagnosis data, in pharmacoepidemiology*. Wiley, 2007:709–65.
17. Weycker D, Edelsberg J, Kartashov A, *et al*. Risk and healthcare costs of chemotherapy-induced neutropenic complications in women with metastatic breast cancer. *Chemotherapy* 2012;58:8–18.
18. Vera-Llonch M, Weycker D, Glass A, *et al*. Healthcare costs in patients with metastatic lung cancer receiving chemotherapy. *BMC Health Serv Res* 2011;11:305.
19. Nordstrom BL, Whyte JL, Stolar M, *et al*. Identification of metastatic cancer in claims data. *Pharmacoepidemiol Drug Saf* 2012;21(Suppl 2):21–8.

20. 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.
21. Carnahan RM, Moores KG. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned. *Pharmacoepidemiol Drug Saf* 2012;21(Suppl 1):82–9.
22. McPheeters ML, Sathe NA, Jerome RN, *et al*. Methods for systematic reviews of administrative database studies capturing health outcomes of interest. *Vaccine* 2013;31(Suppl 10): K2–6.
23. Liberati A, Altman DG, Tetzlaff J, *et al*. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;339:b2700.
24. Shamseer L, Moher D, Clarke M, *et al*. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 2015;349:g7647.
25. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. *J Clin Epidemiol* 2012;65:343–9.e2.
26. Dean BB, Lam J, Natoli JL, *et al*. Review: use of electronic medical records for health outcomes research: a literature review. *Med Care Res Rev* 2009;66:611–38.
27. 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.