Abstract: Objective: To develop a linkage algorithm to match anonymous death records of cancer of the larynx (ICD-10 C32X), retrieved from the Mortality Information System (SIM) and the Hospital Information System of the Brazilian Unified National Health System (SIH-SUS) in Brazil. Methodology: Death records containing ICD-10 C32X codes were retrieved from SIM and SIH-SUS, limited to individuals aged 30 years and over, between 2002 and 2012, in the state of São Paulo. The databases were linked using a unique key identifier developed with sociodemographic data shared by both systems. Linkage performance was ascertained by applying the same procedure to similar non-anonymous databases. True pairs were those having the same identification variables. Results: A total of 14,311 eligible death records were found. Most records, 10,674 (74.6%), were exclusive to SIM. Only 1,853 (12.9%) deaths were registered in both systems, representing true pairs. A total of 1,784 (12.5%) cases of laryngeal cancer in the SIH-SUS database were tracked in SIM with different causes of death. The linkage failed to match 167 (9.4%) records due to inconsistencies in the key identifier. Conclusion: The authors found that linking anonymous data from mortality and hospital records is a feasible measure to track missing records and may improve cancer statistics.

Keywords: Health information systems. Linkage. Medical records. Cancer of the larynx.
INTRODUCTION

Cancer of the larynx (CL) accounts for 1% of all new cases and deaths due to cancer worldwide. In 2018, 177,422 new cases and 94,771 deaths caused by CL were reported. In Brazil, this is the 8th most common type of cancer among men, with 6,390 new cases in 2018. Although smoking and alcohol consumption are the most important risk factors, work-related carcinogens, such as polycyclic aromatic hydrocarbons, inorganic acids, or asbestos, have been linked to an excess of CL in exposed groups. Increased mortality from CL has been reported among miners, tailors, blacksmith, toolmakers and painters.

In high income countries, work-related cancers are commonly-reported occupational diseases, but they remain underreported, particularly in low and middle income countries. An important step in gathering improved number of records is to recover registered cases from distinct sources. Considering that larynx malignancies usually require hospital treatment, administrative information systems could be taken as an additional data source to improve case assessment.

In Brazil, in the last 50 years, asbestos has been extensively used but little is known about its impact on CL. Mapping health information systems (HIS) to recover all registered cases of asbestos-related diseases is one of the aims of the “Asbestos and Health Effects – Brazil” Project. However, privacy and professional or research ethics limit the access to non-anonymous health databases. Data linkage may be used for improving the reporting of a given disease by capturing cases and/or correcting the completeness of a database or a surveillance system. Linking anonymous databases is feasible using computer-based routines based on common shared data. In a previous publication, records on deaths from
mesothelioma and cancer of the pleura found in the Mortality Information System (SIM) and in anonymous databases of the Hospital Information System of the Brazilian Unified National Health System (SIH-SUS) were successfully combined using a unique key algorithm based on date of birth, sex, municipality of death occurrence, and date of death\(^2\), disclosing 32.2% additional records from SIH-SUS with a small linkage failure (1.7%).

Based on death certificates, SIM follows the recommendations of the World Health Organization for underlying and contributing causes of death. SIH-SUS is used for administrative purposes and does not cover hospital admissions in the private system, but registers all comorbidities requiring specific clinical protocols up to nine distinct diagnoses. In addition, considering the conditional relative survival of less than 95% of CL cases for 25 years or more\(^1\), patients may die of other causes, contributing for underreporting. Therefore, SIH-SUS can be used as an alternative source for capturing non-registered cases in SIM.

Up to date, no studies on underreporting of CL or on recovering CL in death records from multiple databases were found. In contrast with a rare tumor, such as mesothelioma, for which anonymous linkage proved feasible, CL is a more prevalent cancer. As part of an effort to track cases of typical or associated asbestos-related diseases to obtain more accurate estimates of its burden in Brazil, this study aims at assessing the feasibility and performance of anonymous linkage of CL records from two health information systems: SIM and SIH-SUS.

**METHODS**

All death records having an ICD-10 code C32X (cancer of the larynx, any subsites) of adults aged 30 years or older were investigated in the period from January 1\(^{st}\), 2002 to December 31, 2012 in the state of São Paulo, Brazil.

**DATA SOURCES**

Death records were retrieved from SIM, a universal vital statistics database, and from SIH-SUS, an administrative hospital information system of the Brazilian Unified National Health System, which only covers state-owned or publicly funded hospitals. Both anonymous databases are freely available.

To assess the linkage performance, corresponding non-anonymous SIM and SIH-SUS datasets were obtained. Each database has multiple ICD codes: in SIM, there is underlying cause of death and multiple contributing causes; SIH-SUS has ICD codes for one principal and a maximum of eight secondary diagnoses, including the death-related cause when applicable. CL deaths consisted in records having at least one assigned C32X, from any coded subtype. In case of multiple C32X in the same individual, the most specific one was used in the analysis.
LINKAGE PROCEDURES

SIM and SIH-SUS databases were checked for duplicates, i.e., when more than one record shared
the same unique key identifier and the same hospital admission form (from Portuguese, Autorização
de Internação Hospitalar — AIH) or death certificate (DC), which were eliminated in both anonymous
and non-anonymous versions prior to linkage. Only SIH-SUS presented duplicates, and records were
manually verified, maintaining records with most of the remaining columns filled. After linkage
performed with anonymous records using the unique key identifier, matched cases were checked
for correction using the non-anonymous corresponding databases, which enabled the authors to
verify full names and mother’s names of the patients. Records with missing data related to variables
required for linkage and its performance assessment were excluded. Records with missing names in
the non-anonymous database were also excluded. Similar to the linkage strategy formerly used12, a
unique identification key variable corresponding to a sequence of coded data (sex, municipality of
death occurrence, date of birth, and date of death) was created. Then, the key was used to merge
both databases, allowing for the identification of paired records — the same case was recorded in
both databases and the algorithm successfully matched the records; unpaired cases only recorded
in SIM; and unpaired cases only recorded in SIH-SUS.

To assess linkage accuracy, the same procedures were applied to the corresponding
non-anonymous SIM and SIH-SUS databases. Both the deceased’s and the mother’s names
were checked for similarity in each matched pair. For a final cross-check, unmatched cases
from SIH-SUS were searched in the complete non-anonymous SIM database to find their
pair, regardless of diagnosis.

DATA MANAGEMENT AND ANALYSIS

A relational database management system (RDBMS), MS SQL Server, was chosen to write
each step of the linkage algorithm and repeat the process when necessary. Features, such as data
transformation tools and index creation, were used to optimize the linkage procedures. Excel
and SAS 9.4 (SAS Institute Inc., Cary, NC, USA, version 9.4) were used for quantitative analyses.

The study protocol was registered at the Brazilian National Research Ethics Committee
(CONEP) and approved by the Ethics Committee through CAAE 36547514900005030, reports No. 962145 and 1761856.

RESULTS

STUDY POPULATION

In the state of São Paulo, from 2002 to 2012, there were 12,530 CL death records in the
SIM anonymous database. In the SIH-SUS database, a total of 4,020 records were found.
Records with missing data that prevented the unique key variable creation were excluded,
specifically three from SIM, totaling 12,527 death certificates, and 383 from SIH-SUS, resulting in 3,637 hospital records for analysis (Figure 1).

**LINKAGE PERFORMANCE**

There were 1,853 (12.9%) paired cases, 10,674 (74.6%) unpaired cases from SIM, and 1,784 (12.5%) unpaired cases from SIH-SUS (Figure 1).

Results from the linkage performance are summarized in Table 1. All paired cases were correctly matched by the linkage strategy. Of the unpaired SIH-SUS records (n = 1,784), 167 (9.4%) had typing errors in the variables used to compose the unique key, which precluded matching. The remaining unpaired records were tracked in the SIM complete database, which contains CL and non-CL cases. Fifty-seven (3.2%) hospital C32X death records were not found in SIM and the remaining 1,560 could be identified with a non-CL ICD code as the underlying cause of death.

Table 2 shows the specific four-digit codes of C32X according to paired status and HIS. Cancer of the larynx, unspecified (C32.9) was the most common regardless of pairing status or HIS, followed by overlapping lesions of larynx (C32.8), which prevails in SIH-SUS among paired (n = 499; 26.9%) or unpaired groups (n = 245; 13.7%). The distribution of sex and age groups according to pairing status (Table 3) shows that most CL records occurred among men (90%) in the paired group, 89.2% only in SIM. Age distribution of men’s and women’s C32X cases differ, with men’s deaths occurring earlier in life compared to women’s.

![Figure 1. Results from anonymous linkage of cancer of the larynx death records from death certificates (SIM) and hospital records (Hospital Information System of the Brazilian Unified National Health System — SIH-SUS), São Paulo, 2002–2012.](image-url)
DISCUSSION

The findings of this study support the feasibility of anonymous databases linkage based on death records containing an annotated CL code. All matched records were confirmed as true pairs after checking the deceased’s and their mothers’ names. Most death records

Table 1. Anonymous linkage performance according to pairing status, São Paulo, 2002–2012.

| Linkage status       | Description                                                                 | Performance          |
|----------------------|----------------------------------------------------------------------------|----------------------|
| Paired               | All 1,853 pairs were truly matched.                                         | 100% success         |
| Only in SIH–SUS records | To every death occurring at a hospital, a death certificate must be issued, as it is a legally required document. Of the 1,784 deaths reported in SIH-SUS, 57 could not be found in SIM (3.2%), which may have been caused by a typing mistake or misclassification. In addition, 167 (9.4%) had pairs in SIM but were not matched, representing a linkage failure. | 90.6% success 9.4% failure |
| Only in SIM records  | After being tracked down, 10,674 cases from SIM did not have a pair in SIH-SUS. Not all cases are expected to die while under hospital care. | 100% success         |
| Total                | Not all cancer of the larynx deaths registered in SIH-SUS are true death cases reported in SIM, even considering any other diagnoses. | Overall, the linkage strategy performed well, with 9.4% failure in matching true pairs. All failures were caused by inconsistencies in sociodemographic data. |

SIH–SUS: Hospital Information System of the Brazilian Unified National Health System; SIM: Mortality Information System.

Table 2. Distribution of specific ICD–10 codes of cancer of the larynx deaths by pairing status after linkage.

| ICD-10 Codes                | Paired | Unpaired |
|-----------------------------|--------|----------|
|                             | SIM    | SIH–SUS  | SIM    | SIH–SUS  |
|                             | N = 1,853 | %    | N = 1,853 | %    | N = 10,674 | %    | N = 1,784 | %    |
| Cancer of the larynx        |        |        |        |        |        |        |        |        |
| C32.0 Glottis               | 37     | 2.0    | 102    | 5.5    | 233     | 2.2    | 99      | 5.5    |
| C32.1 Supraglottis          | 35     | 1.9    | 58     | 3.1    | 185     | 1.7    | 45      | 2.5    |
| C32.2 Subglottis            | 1      | 0.1    | 4      | 0.2    | 18      | 0.2    | 8       | 0.4    |
| C32.3 Laryngeal cartilage   | 0      | 0.0    | 23     | 1.2    | 56      | 0.5    | 8       | 0.4    |
| C32.8 Overlapping lesions of larynx | 17     | 0.9    | 499    | 26.9   | 88      | 0.8    | 245     | 13.7   |
| C32.9 Cancer of the larynx, unspecified | 1,763 | 95.1   | 1,167  | 63.0   | 10,094 | 94.6   | 1,379   | 77.3   |

ICD–10: International Classification of Diseases; SIH–SUS: Hospital Information System of the Brazilian Unified National Health System; SIM: Mortality Information System.
Table 3. Distribution of cancer of the larynx records according to linkage status by sex and age groups, São Paulo, 2002–2012.

| Sex/age groups | Paired SIM/SIH-SUS | Present only in SIM | Present only in SIH-SUS | All |
|----------------|---------------------|---------------------|------------------------|-----|
|                | n       | %     | n   | %    | n   | %    | n   | %    |
| Men            |         |       |     |      |     |      |     |      |
| Age groups (years) |     |       |     |      |     |      |     |      |
| 30–58          | 697     | 41.8  | 3,599 | 37.8 | 481 | 38.6 | 4,777 | 38.4 |
| 59–70          | 613     | 36.8  | 3,367 | 35.4 | 443 | 35.5 | 4,423 | 35.6 |
| Older than 70  | 358     | 21.4  | 2,557 | 26.8 | 323 | 25.9 | 3,238 | 26.0 |
| Subtotal       | 1,668   | 90.0  | 9,523 | 89.2 | 1,247 | 69.9 | 12,438 | 86.9 |
| Women          |         |       |     |      |     |      |     |      |
| Age groups (years) |     |       |     |      |     |      |     |      |
| 30–58          | 57      | 30.8  | 319  | 27.7 | 173 | 32.2 | 549  | 29.3 |
| 59–70          | 63      | 34.1  | 363  | 31.5 | 158 | 29.4 | 584  | 31.2 |
| Older than 70  | 65      | 35.1  | 469  | 40.8 | 206 | 38.4 | 740  | 39.5 |
| Subtotal       | 185     | 10.0  | 1,151 | 10.8 | 537 | 30.1 | 1,873 | 13.1 |

SIH-SUS: Hospital Information System of the Brazilian Unified National Health System; SIM: Mortality Information System.

only registered in SIH-SUS could be tracked in SIM in which the underlying and contributing causes of death had been registered with another ICD-10 code. Misclassification of death records in SIH-SUS represented 3.2% of unpaired cases in this database. The use of hospital data enabled to recover 12.5% CL records unreported in death certificates from SIM. Therefore, hospital records can disclose a significant number of deaths displaying a CL coding, which is a quite common disease.

The most commonly recorded ICD-10 codes were cancer of the larynx, unspecified (ICD-10 C32.9) and overlapping lesions of larynx (ICD-10 C32.8). Cases between men prevailed, regardless of HIS or pairing status. Deaths in men occurred at an earlier age compared to women.

The feasibility of anonymous database linkage for CL is compelling, considering the simple strategy required and the availability of public health information systems, either of vital statistics, surveillance or administrative data. This strategy has been widely recognized as a useful tool to generate knowledge necessary to outline public health policies and programs for disease prevention and health promotion14. It can be of particular importance for epidemiological surveillance, concerning Workers’ Health, considering the existence of large demographic databases, specially of labor forces and employment. Linkage is also crucial to develop complex study designs and long-term follow-ups of large populations using secondary data. Similar procedures have been successfully used for deaths caused by mesothelioma and cancer of the pleura, both considered rare diseases12. The present results show that using combined sociodemographic data as a unique key identifier can accurately match records of a quite common malignant neoplasia, the cancer of the larynx.
Inconsistencies in demographic data were the only reason for linkage failure. Such aspect highlights the importance in ensuring the quality of simple data because they can be used for other purposes such as sociodemographic descriptors. The poor quality of records may not only bias the estimates, but also compromise data management as the linkage itself. The potential use of multiple data sources in epidemiologic research or surveillance could be a strong reason to develop programs focused on data quality improvement in HIS. Most HIS are based on electronic forms and computational solutions may be introduced to block inconsistent data entry.

Inconsistencies were also flagged in death records from SIH-SUS, as a small proportion of reported cases could not be found in SIM. At present, SIM coverage is evaluated as “good” or “poor,” being considered good in Brazil, particularly in the South and Southeast regions of the country\textsuperscript{15}, where access to and the quality of healthcare and other basic social services are better compared with other Brazilian regions.

The proportion of 12.9\% matched pairs for CL from SIM and SIH-SUS was higher compared with 5.7\% obtained from the same linkage procedure for mesothelioma and cancer of the pleura\textsuperscript{12}. This was expected, as CL is more common, more easily recognized, and has a longer survival time. Consequently, patients have an increased chance of dying from competing causes.

The SIH-SUS database only covers hospitalizations in publicly funded hospitals. In the Southeast region, where the state of São Paulo is located, the public system accounted for 63.7\% of these events\textsuperscript{16}. In 2012, 42.6\% of the population of São Paulo had private health insurance. The SIH-SUS database is fed by hospital admission forms (AIH). These forms are administrative documents whose main function is the reimbursement of hospital expenses that, in turn, are tied to the requested procedures\textsuperscript{17}. There is no employee-specific training to include ICD codes in the hospital admission forms, unlike what happens in SIM\textsuperscript{18,19}. In the 2008-2010 period, 92.1\% of hospitalizations in the public system were notified, which demonstrates a good coverage\textsuperscript{20}. However, the nosological information contained in the AIH reflects the reason for hospitalization, and may omit other diseases or comorbidities of importance. In contrast to the rules for filling out death certificates, the field of secondary diagnoses is not used in most SIH-SUS records, being a factor of loss of information\textsuperscript{20}.

The proposed unique key solely based on sociodemographic data proved to be effective for linking rare diagnoses\textsuperscript{12} and for cancer of the larynx. The slight increase in the failure rate can be considered negligible due to the proportion of new records that were recovered. It paves the way for the enhanced use of multiple health information systems to capture unreported cases of both rare and common diseases. However, there is need to test the performance of the proposed linkage for more incident cancers, in such a way there will be more records sharing the same sociodemographic data.

Despite being a successful method for recovering records of interest, the validity of the ICD codes should be pursued by checking records, whenever possible, using other data sources, such as cancer registries, pathology reports, and/or clinical notes, in order to strengthen the linkage procedure and allow estimates of the yield of true disease cases. Confirmation rates were slightly higher than detection rates for cancer of the larynx in an accuracy study of cancer mortality statistics comparing the underlying cause of death with population-based cancer registries in three US states, suggesting a tendency for underreporting\textsuperscript{21}.\normalsize
Therefore, other data sources in addition to SIH-SUS can be used for capturing non-registered cases in SIM.

In conclusion, SIH-SUS is valuable to identify records of diseases of interest unreported in SIM. CL is often a comorbidity when death is the endpoint; therefore, it is precluded of being recorded as the underlying or even contributing cause of death. The use of SIH-SUS for this purpose is still limited due to its partial coverage and data inconsistencies as far as sociodemographic variables are concerned. Efforts to improve the quality of data and to standardize the SIH-SUS database can boost its use in epidemiological and demographic studies and may be implemented as effective tools for management purposes and health research in the future.

ACKNOWLEDGEMENTS

Non-anonymous databases from publicly funded hospitals were made available by Simone Santos and Rosemeire Norye Inamine, Health Department of the State of São Paulo; data linkage to non-anonymous SIM database was performed at the Institute of Collective Health, Universidade Federal da Bahia.

REFERENCES

1. World Health Organization. Larynx [Internet]. Globocan; 2020 [cited on July 7, 2019]. Available at: https://gco.iarc.fr/today/data/factsheets/cancers/14-Larynx-fact-sheet.pdf
2. Instituto Nacional de Câncer José Alencar Gomes da Silva. Coordenação de Prevenção e Vigilância. Estimativa 2018: incidência de câncer no Brasil. Rio de Janeiro: INCA; 2017.
3. Talamini R, Bosetti C, La Vecchia C, Dal Maso L, Levi F, Bidoli E, et al. Combined effect of tobacco and alcohol on laryngeal cancer risk: A case-control study. Cancer Causes Control 2002; 13(10): 957-64. https://doi.org/10.1023/a:1021944123914
4. Wagner M, Bolm-Audorff U, Hegewald J, Fishta A, Schlattmann P, Schmitz J, et al. Occupational polycyclic aromatic hydrocarbon exposure and risk of larynx cancer: A systematic review and meta-analysis. Occup Environ Med 2015; 72(3): 226-33. https://doi.org/10.1136/oemed-2014-102317
5. International Agency for Research on Cancer (IARC). A review of human carcinogens: arsenic, metals, fibres, and dusts. Vol. 100C, IARC Monographs. IARC; 2012.
6. Straif K, Benbrahim-Tallaa L, Baan R, Grosse Y, Secretan B, El Ghissassi F, et al. A review of human carcinogens–Part C: metals, arsenic, dusts, and fibres. Lancet Oncol 2009; 10(5): 453-4. https://doi.org/10.1016/s1470-2045(09)70134-2
7. Bayer O, Cámara R, Zeissig SR, Ressing M, Dietz A, Locati LD, et al. Occupation and cancer of the larynx: a systematic review and meta-analysis. Eur Arch Otorhinolaryngol 2016; 273(1): 9-20. https://doi.org/10.1007/s00405-014-3321-y
8. European Agency for Safety and Health at Work. Exposure to carcinogens and work-related cancer: A review of assessment methods [Internet]. European Agency for Safety and Health at Work; 2014 [cited on Sept. 20, 2019]. Available from: https://osha.europa.eu/en/publications/exposure-carcinogens-and-work-related-cancer-review-assessment-methods
9. Algranti E, Giannasi F, Santana VS. The fight for the asbestos ban in Brazil and the 2nd international seminar "Brazil without asbestos". Epidemiol Prev 2018; 42(5-6): 388-90. https://doi.org/10.19191/ep18.5-6.p388.115
10. World Medical Association. Declaration of Taipei. Ethical considerations regarding health databases and biobanks [Internet]. World Medical Association; 2016 [cited on Dec 9, 2018]. Available at: https://www.wma.net/policies-post/wma-declaration-of-taipie-on-ethical-considerations-regarding-health-databases-and-biobanks/
11. Bernillon P, Lievre L, Pillonel J, Laporte A, Costagliola D. Record-linkage between two anonymous databases for a capture-recapture estimation of underreporting
of AIDS cases: France 1990-1993. The Clinical Epidemiology Group from Centres d’Information et de Soins de l’Immunodéficience Humaine. Int J Epidemiol 2000; 29(1): 168-74. https://doi.org/10.1093/ije/29.1.168

12. Santana VS, Algranti E, Campos F, Cavalcante F, Salvi L, Santos SA, et al. Recovering missing mesothelioma deaths in death certificates using hospital records. Am J Ind Med 2018; 61(7): 547-55. https://doi.org/10.1002/ajim.22846

13. AIRTUM Working Group. Italian cancer figures, report 2014: Prevalence and cure of cancer in Italy. Epidemiol Prev 2014; 38(6 Suppl. 1): 1-122. https://doi.org/10.19191/ep14.6.st.113

14. Jutte D, Roos L, Brownell M. Administrative record linkage as a tool for public health research. Annu Rev Public Health 2011; 32: 91-108. https://doi.org/10.1146/annurev-publhealth-031210-100700

15. Vasconcelos AMN. Qualidade das estatísticas de óbitos no Brasil: uma classificação das unidades da federação. In: Anais do XII Encontro Nacional de Estudos Populacionais [Internet]. [cited on Sept 20, 2019]. Available from: http://www.abep.org.br/publicacoes/index.php/anais/article/view/1001

16. Instituto Brasileiro de Geografia e Estatística. Coordenação de Trabalho e Rendimento. Pesquisa Nacional de Saúde: 2013. Acesso e utilização dos serviços de saúde, acidentes e violências: Brasil, grandes regiões e unidades da federação. Rio de Janeiro: IBGE; 2015.

17. Moreira ML, Novaes HMD. Internações no sistema de serviços hospitalares, SUS e não SUS: Brasil, 2006. Rev Bras Epidemiol 2011; 14(3): 411-22. https://doi.org/10.1590/S1415-790X2011000300006

18. Veras CMT, Martins MS. A confiabilidade dos dados nos formulários de Autorização de Internação Hospitalar (AIH), Rio de Janeiro, Brasil. Cad Saúde Pública 1994; 10(3): 339-55. https://doi.org/10.1590/S0102-311X1994000300014

19. Mathias TAF, Soboll MLMS. Confiabilidade de diagnósticos nos formulários de autorização de internação hospitalar. Rev Saúde Pública 1998; 32(6): 526-32. https://doi.org/10.1590/0034-8910199800600005

20. Machado JP, Martins M, Leite IC. Qualidade das bases de dados hospitalares no Brasil: alguns elementos. Rev Bras Epidemiol 2016; 19(3): 567-81. https://doi.org/10.1590/1980-5497201600030008

21. German RR, Fink AK, Heron M, Stewart SL, Johnson CJ, Finch JL, et al. The accuracy of cancer mortality statistics based on death certificates in the United States. Cancer Epidemiol 2011; 35(2): 126-31. https://doi.org/10.1016/j.canep.2010.09.005

22. Cascão AM, Mello Jorge MHP, Costa AJL, Kale PL. Uso do diagnóstico principal das internações do Sistema Único de Saúde para qualificar a informação sobre causa básica de mortes naturais em idosos. Rev Bras Epidemiol 2016; 19(4): 713-26. https://doi.org/10.1590/1980-54972016000400003