https://doi.org/10.1590/1980-549720210044

ORIGINAL ARTICLE / *ARTIGO ORIGINAL*

Agreement and validity of asbestos-related cancers in the hospital information system of the Brazilian Unified Health System

Concordância e validade dos diagnósticos de cânceres associados ao asbesto no sistema de informação hospitalar do Sistema Único de Saúde

Diego Rodrigues Mendonça e Silva[I,II], Carolina Terra de Moraes Luizaga[III], Tatiana Natasha Toporcov[IV], Eduardo Algranti[V]

ABSTRACT: ***Objective*****:** To estimate the degree of agreement and validity of diagnoses of asbestos-related malignant neoplasms registered in the Hospital Information System of the Brazilian Unified Health System (SIH/SUS), in comparison to the Hospital Cancer Registries of the State of São Paulo (HCR/SP). ***Methods*****:** Deaths with records of malignant neoplasms associated with asbestos were identified and extracted from SIH/SUS between 2007 and 2014. Deaths in cases registered in the HCR/SP were extracted for the same period. The databases were linked using software *Link Plus*. A single ICD-10-coded diagnosis selected from each system was analyzed. The proportion of agreement, and the sensitivity, specificity and predictive values were estimated. ***Results*****:** 19,458 pairs were found with records in both bases. The proportion of agreement was high, ranging from 92.4% for the unknown primary site, to 99.7% for cancer of the pleura. The Kappa Index ranged from 0.05 (95%CI 0.04 – 0.07) for cancer of the pleura to 0.85 (95%CI 0.84 – 0.87) for lung cancer. Sensitivity varied from 0.08 (95%CI 0.01 – 0.25) for cancer of the pleura, to 0.90 (95%CI 0.90 – 0.91) for lung cancer. ***Conclusion*****:** Diagnosis of asbestos-related malignancies reached higher levels of agreement and validity when common. Rare diagnoses showed low accuracy in SIH/SUS.

Keywords**:** Health information interoperability. Asbestos. Neoplasms. Information storage and retrieval.

[I]Centro Internacional de Pesquisa, A. C. Camargo Cancer Center – São Paulo (SP), Brazil.
[II]Postgraduate Program in Epidemiology, Public Health School, Universidade de São Paulo – São Paulo (SP), Brazil.
[III]Board of Information and Epidemiology, Fundação Oncocentro de São Paulo – São Paulo (SP), Brazil.
[IV]Department of Epidemiology, Public Health School, Universidade de São Paulo – São Paulo (SP), Brazil.
[V]Board of Applied Research, Fundação Jorge Duprat Figueiredo, Ministry of Economy – São Paulo (SP), Brazil.

Corresponding author: Diego Rodrigues Mendonça e Silva. Centro Internacional de Pesquisa, A. C. Camargo Cancer Center. Rua Taguá, 440, Liberdade, CEP: 01508-010, São Paulo, SP, Brazil. E-mail: diego.rmesilva@gmail.com

Conflict of interests: nothing to declare – **Financing source:** Ministério Público do Trabalho da 15ª Região, Escritório de Campinas, São Paulo (CODIN No. 45672/2014).

RESUMO: ***Objetivo*:** Estimar o grau de concordância e validade dos diagnósticos de neoplasias malignas relacionadas à exposição ao asbesto registrados no Sistema de Informação Hospitalar do Sistema Único de Saúde (SIH/SUS), em comparação aos Registros Hospitalares de Câncer do Estado de São Paulo (RHC/SP). ***Métodos*:** Óbitos com registros de neoplasias malignas associadas ao asbesto foram identificados e extraídos do SIH/SUS entre 2007 e 2014. Óbitos nos casos de câncer registrados na base do RHC/SP foram extraídos para o mesmo período. Essas bases foram unidas pelos mesmos indivíduos empregando-se o software *Link Plus*. Um único diagnóstico codificado pela CID-O3 ou CID-10 selecionado de cada sistema foi analisado. A proporção de concordância e a sensibilidade, especificidade e valores preditivos foram estimados. ***Resultados*:** Encontraram-se 19.458 pares com registros nas duas bases. A proporção de concordância foi elevada, variando de 92,4% para a localização primária desconhecida a 99,7% para o câncer de pleura. O índice Kappa variou de 0,05 (IC95% 0,04 – 0,07) para o câncer de pleura a 0,85 (IC95% 0,84 – 0,87) para o câncer de pulmão. A menor sensibilidade foi de 0,08 (IC95% 0,01 – 0,25), para o câncer de pleura, e a maior de 0,90 (IC95% 0,90 – 0,91), para o câncer de pulmão. ***Conclusão*:** Diagnósticos de neoplasias malignas associadas ao asbesto alcançaram maiores níveis de concordância e validade quando comuns. Os diagnósticos mais raros apresentaram baixa acurácia no SIH/SUS.

***Palavras-chave*:** Interoperabilidade da informação em saúde. Asbestos. Neoplasias. Armazenamento e recuperação da informação.

INTRODUCTION

The monitoring of epidemiological indicators of relevant diseases for public health is one of the pillars of health surveillance, allowing to estimate the extension of morbidity and mortality, the burden, the temporal variations and spatial distributions, among others, thus contributing with responses of control, prevention and improvement of health of the population. In Brazil, several health information systems (HIS) have been improved and consolidated since the creation of the Unified Health System (SUS)[1-3], made available through DATASUS.

The Hospital Information System (SIH/SUS) is a HIS that registers data from admissions in hospitalization services and is addressed to administrative purposes, especially to the reimbursement of expenses in public, philanthropic and private hospitals, when associated to or hired by SUS. Since it includes 60 to 70% of the total number of hospitalizations in the country[4], SIH/SUS is also interesting for health surveillance by presenting records of codified diagnoses, according to the International Statistical Classification of Diseases and Related Health Problems (ICD)[5].

In a study conducted as part of the project of Occupational Exposure to Asbestos and Its Effects on Health in Brazil, Santana et al.[6] looked for cases of mesothelioma and/or cancer of the pleura deaths in SIH/SUS and found data that allowed to estimate diagnostic agreements in 70% of the pairs. The linkage between SIH/SUS and the Mortality Information System (SIM) increased the records of mesothelioma and/or cancer of the pleura in 32.2% between 2002 and 2012 in the state of São Paulo. These records were recovered from SIH/SUS, and not identified in SIM[6].

The Hospital Cancer Registry (HCR) and the Population Based Cancer Registry (PBCR) are the main bases that compose the Cancer Information System (SISCAN), of the National Cancer Institute (INCA/SUS), addressed to the promotion of cancer prevention, care and research in the country[3,7]. The records of cancer are confirmed through anatomopathological examinations in more than 90% of the cases[7-9].

It is estimated that exposure to carcinogens in the work environment is responsible for 5 to 8% of the cancer burden in men, especially lung cancer. Among the occupational carcinogens, asbestos is responsible for more than 50% of occupational lung cancers[10]. Asbestos, in all of its varieties, is classified by the International Agency for Research on Cancer (IARC) in Group 1, as a causal and/or concausal agent in mesothelioma, lung, larynx and ovarian cancer[11], and possibly associated with other head and neck, pharynx, esophagus, stomach and colorectal cancers[12]. Regarding mesothelioma, the transition from ICD-9 to ICD-10 expressively increased the records of the underlying cause of death by mesothelioma, due to the existence of a specific code for the disease[13]. The use of cancer records was a valid source of mesothelioma diagnoses, when compared to clinical, imaging and anatomopathological records[14].

Due to its coverage, SIH/SUS is a source of search for disease records, and, therefore, of possible underreported cases of asbestos-related diseases. Its records of diagnosis and clinical procedures, however, may present problems, which requires evaluations of accuracy, especially in epidemiological analyses, both for surveillance and/or specific studies[1-2,4,15-18]. Aiming at using SIH/SUS as a data source of malignant neoplasms associated with asbestos, the objective of this study is to estimate the agreement and validity of cancer diagnoses related to asbestos exposure by adopting, as gold standard, the records from the HCR in the state of São Paulo.

METHODS

In this study, records of deaths with diagnosis of asbestos-related cancers in the state of São Paulo coming from SIH/SUS were compared to all cases of cancer with information of death in the HCR/SP database. The period of analysis was from 2007 to 2014, in individuals of both genders, aged 30 years of more.

DATA SOURCES

In SIH/SUS, a single patient may have several hospitalizations and diagnoses. In the present study, only hospitalizations with a death outcome and with a mention of a cancer of interest in one of the diagnoses (ICD-10[19]) were considered.

In the database of HCR/SP (consolidated by Fundação Oncocentro de São Paulo – FOSP)[9], we selected all cases of cancer coded by the International Classification of Diseases for Oncology, third edition (ICD-O3)[20], with death information, regardless of the cause, excluding non-melanoma skin cancer.

Both databases were identified.

DEFINITION OF DIAGNOSES OF INTEREST

The following malignant neoplasms were considered as asbestos-related: mesothelioma (C45), lung cancer (C34), ovarian cancer (C56) and laryngeal cancer (C32). The esophagus (C15), stomach (C16), colon and rectal (C18-20) neoplasms were considered as suspected relation. Cancer of the pleura (C38.4) was also selected for being a possible diagnosis that could cover up mesothelioma. Likewise, we selected cases of head and neck cancer (C00-C14, C30-C31) for being differential diagnosis with laryngeal cancer, as well as cancers of unknown primary site (C80)[10,11].

Among the studied oncologic groups, only mesothelioma presents a difference in identification between ICD-10 codes, classified as C45, and ICD-O3 codes, identified by morphological codes 9050/3, 9051/3, 9052/3 and 9053/3. The other groups of interest present equivalence between the three-digit code of ICD-10 and ICD-O3. Therefore, the three-digit codes of ICD-O3 and ICD-10 were used for asbestos-related cancers that were previously mentioned, except for cancer of the pleura (C38.4); for the latter, we used the four-digit classification, since, in that case, the last digit changes the tumor site[20].

DATABASES LINKAGE

After standardizing the variables in the SIH/SUS and HCR/SP databases, the probabilistic linkage was carried out through the *Link Plus* software, working on standardization, blocking and matching of variables. *Link Plus* is developed by the Centers for Disease Control and Prevention (CDC/USA), and is a record linkage tool for cancer registries, and to link a cancer registry file with external files using common identifiers[21-23].

Gender was selected as the blocking variable and allowed that the databases were logically divided in blocks, according to the link variable. Therefore, the comparison and calculation of scores are limited to the records that belong to the same block.

For the matching variables, the following were selected: full name of the individual and his/her mother as generic string and date of birth. *Link Plus* considers the phonetic differences, as well as the absence of parts or of the complete last name at the time of linkage for computing the classification of matching. Thus, the software computes record linkage scores for each established potential match. Pairs of records in which the matching variables coincide get the maximum score, which is, full name of the individual, mother's name and date of birth. The lower the combination of records between the three identifiers, the lower the score. We selected the "best match" option so that the software could present the best choice in case of multiple match records. *Link Plus* recommends the cut-off value to be between 7 and 10. After trials carried out with the database using 7, a high number of false-positive pairs was observed; then, the choice was for the cut-off value of 10, selecting as potential matches and presenting for manual review those whose score was equal to or higher than 10[21].

Finally, a manual review for the classification as a true match was conducted, in which we not only compared the matching variables, but also the zip code and date of death.

ANALYSIS OF AGREEMENT

For the analysis of agreement, the diagnosis of HCR/SP versus the main diagnosis of SIH/SUS of true pairs was considered, obtained from the probabilistic linkage in the previous stage. When the main diagnosis of SIH/SUS was not oncologic, secondary diagnoses and others were used. When there were more than two oncologic diagnoses in the same record, the determinant was the main diagnosis.

The agreement was assessed according to the ICD-10 code and the anatomical location of malignant neoplasms among those related to asbestos exposure, according to the Cohen's Kappa statistics measure[24], an indicator of reliability that quantifies the agreement between observers, adjusted by those that occur by chance; therefore, it is adequate for the analysis in this study. Besides the Kappa, prevalence-adjusted and bias-adjusted Kappa (PABAK), the prevalence index (PI) and the bias index (BI) were calculated. Low PI and BI values suggest that Kappa values are less subjected to bias and the effect of prevalence, whereas higher PI and BI values tend to reduce the Kappa value[24,25]. The analyses were carried out with software R, version 4.0 (R Development Core Team), epiR package.

VALIDITY ANALYSIS

For the validity analysis, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated, with their respective 95% confidence intervals (95%CI), considering SIH/SUS in relation to HCR/SP. The results obtained in the linkage of death records in HCR/SP (gold standard) and SIH/SUS of asbestos-related tumors (lung, head and neck, larynx, mesothelioma, pleura, esophagus, stomach, colon and rectum, ovary and cancer of unknown primary site) were considered. The analyses were conducted in the R software, version 4.0 (R Development Core Team), using the epiR package.

ETHICAL ASPECTS

The study was registered in the National Commission for Research Ethics (CONEP), and approved with the Certificate of Presentation for Ethical Appreciation (CAAE) 36547514 9 0000 5030, Addendum no. 962 145 and 1 761 856, Instituto de Saúde Coletiva at Universidade Federal da Bahia.

RESULTS

The SIH/SUS database displaying records of diseases and malignant neoplasms associated with asbestos exposure contained 360,443 multiple records of hospitalizations from 2007 to 2014, in the state of São Paulo. After identifying the single records with the death outcome, the oncologic diagnoses were selected; therefore, the total obtained was 56,623 records in SIH/SUS. In the same period, a total of 151,766 deaths were identified in HCR/SP, except for cases of non-melanoma skin cancer.

AGREEMENT

Both in SIH/SUS and in HCR/SP, the male gender was prevalent: 61.3 and 56.8%, respectively. The population aged more than 50 years corresponded to 87.0% in SIH/SUS and 84.8% in HCR/SP (Table 1). After the probabilistic linkage, 64.4% of the paired cases resulted in the male gender, and 56.2% above the age of 60 years (Supplementary Table 1).

Table 1. Characteristics of cancer deaths, according to the data sources, Hospital Cancer Registries in the State of São Paulo (HCR/SP) and the Hospital Information System (SIH/SUS), state of São Paulo, Brazil, 2007-2014.

Variables	HCR/SP		SIH/SUS	
	n	%	n	%
Sex				
Male	86,137	56.8	34,682	61.3
Female	65,629	43.2	21,941	38.7
Age group (years)				
30–49	23,128	15.2	7,420	13.1
50–69	78,290	51.6	30,329	53.6
70+	50,348	33.2	18,874	33.4
Year				
2007	16,874	11.1	6,824	12.1
2008	18,989	12.5	5,653	10.0
2009	19,883	13.1	6,410	11.3
2010	19,983	13.2	7,025	12.4
2011	20,202	13.3	7,125	12.6
2012	19,173	12.6	7,560	13.4
2013	18,553	12.2	7,786	13.8
2014	18,109	11.9	8,240	14.6
State				
São Paulo	143,708	94.7	55,254	97.6
Other states	8,058	5.3	1,369	2.4
Total	151,766	100.0	56,623	100.0

The probabilistic linkage of the databases resulted in 26,196 potential matches with scores between 10.0 and 28.3. After the manual review with divergent variables and/or scores below 20.0, 19,458 were considered as true pairs.

In the agreement analysis, results are presented for the ICD-10 groups of cancers related, or possibly related, to asbestos exposure. The proportion of agreement for cancer of the pleura (99.7%), mesothelioma (99.6%) and ovarian cancer (98.7%) stood out, whereas a lower proportion was estimated for cancer of unknown primary site (92.4%) and lung cancer (94.7%). However, Kappa values, differently, ranged from 0.05, for cancer of the pleura, to 0.85 for lung cancer, suggesting an inverse performance in relation to findings of proportions of agreement. These contrasts appear due to the variation in the number of cases for specific diagnoses: 5,179 for lung cancer and only 99 for mesothelioma, and 71 for cancer of the pleura. Supplementary Table 2 lists the observations of each cancer of interest in the two databases. It is possible to observe that, among the 70 paired cases of mesothelioma registered in SIH/SUS, only 20 (28.6%) presented the same diagnosis in HCR, and, among the 73 registered as cancer of the pleura, 6 (8.2%) presented the diagnosis of mesothelioma or cancer of the pleura in HCR (Supplementary Table 3). The Kappa of the general table was 0.74.

For lung cancer, Kappa was 0.85 (95%CI 0.84 – 0.87), and 0.84 for stomach cancer (95%CI 0.83 – 0.86). For rare tumors, the agreement was 0.33 (95%CI 0.32 – 0.35) for mesothelioma, and 0.05 (95%CI 0.04 – 0.07) for cancer of the pleura, whereas for cancer of unknown primary site it was 0.33 (95%CI 0.32 – 0.35). For prevalence and bias indexes, the values were 0.52 to 0.99, and < 0.01 to 0.02, respectively (Table 2).

Table 2. Agreement of asbestos-related cancer deaths linkage between in the Hospital Cancer Registries in the State of São Paulo (HCR/SP) and the Hospital Information System (SIH/SUS), São Paulo, Brazil, 2007–2014.

ICD-10	n	Agreement (%)	Kappa (95%CI)	PI	BI	PABAK
Respiratory system						
Lung (C34)	5,179	94.7	0.85 (0.84 – 0.87)	0.52	< 0.01	0.89
Larynx (C32)	1,150	96.7	0.61 (0.60 – 0.63)	0.91	< 0.01	0.93
Mesothelioma (C45)	99	99.5	0.33 (0.32 – 0.35)	0.99	< 0.01	0.99
Pleura (C38.4)	71	99.6	0.05 (0.04 – 0.07)	0.99	< 0.01	0.99
Digestive system						
Esophagus (C15)	2,407	96.6	0.82 (0.81 – 0.83)	0.79	< 0.01	0.93
Stomach (C16)	3,694	95.7	0.84 (0.83 – 0.86)	0.66	< 0.01	0.91
Colon and rectum (C18-C20)	2,711	96.6	0.84 (0.83 – 0.86)	0.76	< 0.01	0.93
Others						
Head and neck (C00-C14)	2,956	95.0	0.77 (0.76 – 0.79)	0.75	< 0.01	0.90
Ovary (C56)	777	98.7	0.79 (0.78 – 0.81)	0.93	< 0.01	0.97
Unknown primary site (C80)	1,930	92.4	0.33 (0.32 – 0.35)	0.88	0.02	0.85

ICD-10: International Statistical Classification of Diseases and Related Health Problems; n: connected pairs; 95%CI: 95% confidence interval; PI: prevalence index; BI: bias index; PABAK: prevalence-adjusted and bias-adjusted kappa.

VALIDITY

In the analysis of validity, we identified higher sensitivity among lung (0.90, 95%CI 0.90 – 0.91) and stomach cancer deaths (0.86, 95%CI 0.85 – 0.87), whereas sensitivity was lower for rare tumors, such as mesothelioma (0.41, 95%CI 0.27 – 0.56) and cancer of the pleura (0.08, 95%CI 0.01 – 0.25), and those classified as cancers of unknown primary site (0.45, 95%CI 0.42 – 0.49). The positive predictive value was also lower for those registered as cancers of unknown primary site (0.32, 95%CI 0.29 – 0.34), mesothelioma (0.29, 95%CI 0.18 – 0.41) and cancer of the pleura (0.04; 95%CI 0.01 – 0.16). However, the specificity was high for all of the analyzed tumors, as well as the negative predictive value (Table 3). The absolute values are presented in Supplementary Table 2.

Table 3. Sensitivity and specificity between asbestos-related cancer deaths, linkage between the Hospital Cancer Registries in the State of São Paulo (HCR/SP) and the Hospital Information System (SIH/SUS), state of São Paulo, Brazil, 2007-2014.

ICD-10	n	Sensitivity (95%CI)	Specificity (95%CI)	PPV (95%CI)	NPV (95%CI)
Respiratory System					
Lung (C34)	5,179	0.90 (0.90 – 0.91)	0.96 (0.96 – 0.96)	0.88 (0.87 – 0.89)	0.97 (0.97 – 0.97)
Larynx (C32)	1,150	0.68 (0.64 – 071)	0.98 (0.98 – 0.98)	0.58 (0.55 – 0.62)	0.99 (0.98 – 0.99)
Mesothelioma (C45)	99	0.41 (0.27 – 0.56)	1.00 (1.00 – 1.00)	0.29 (0.18 – 0.41)	1.00 (1.00 – 1.00)
Pleura (C38.4)	71	0.08 (0.01 – 0.25)	1.00 (1.00 – 1.00)	0.04 (0.01 – 0.16)	1.00 (1.00 – 1.00)
Digestive system					
Esophagus (C15)	2,407	0.87 (0.85 – 0.88)	0.98 (0.97 – 0.98)	0.81 (0.80 – 0.83)	0.98 (0.98 – 0.99)
Stomach (C16)	3,694	0.86 (0.85 – 0.87)	0.98 (0.97 – 0.98)	0.88 (0.87 – 0.89)	0.97 (0.97 – 0.97)
Colon and rectum (C18-C20)	2,711	0.86 (0.84 – 0.87)	0.98 (0.98 – 0.98)	0.86 (0.85 – 0.88)	0.98 (0.98 – 0.98)
Others					
Head and neck (C00-C14)	2,956	0.75 (0.74 – 0.77)	0.98 (0.98 – 0.98)	0.86 (0.84 – 0.87)	0.96 (0.96 – 0.97)
Ovary (C56)	777	0.86 (0.83 – 0.89)	0.99 (0.99 – 0.99)	0.74 (0.71 – 0.77)	1.00 (0.99 – 1.00)
Unknown primary site (C80)	1,930	0.45 (0.42 – 0.49)	0.95 (0.95 – 0.95)	0.32 (0.29 – 0.34)	0.97 (0.97 – 0.97)

ICD-10: International Statistical Classification of Diseases and Related Health Problems; n: connected pairs; 95%CI: 95% confidence interval; PPV: positive predictive value; NPV: negative predictive value.

DISCUSSION

The probabilistic linkage between asbestos-related cancer deaths in SIH/SUS and HCR/SP from 2007 to 2014 in the state of São Paulo revealed good agreement for more frequent cancers, such as lung [Kappa = 0.85 (95%CI 0.84 – 0.87)], laryngeal, ovarian and digestive cancers. However, the agreement remained low for rare cancers such as mesothelioma [Kappa = 0.33 (95%CI 0.32 – 0.35)] and cancer of the pleura [Kappa = 0.05 (95%CI 0.04 – 0.07)], both with very low positive predictive value (PPV), 0.29 and 0.04, respectively. General Kappa was 0.74 (Supplementary Table 3), classified as substantial, according to the criteria by Landis and Koch[26]; or good, according to Fleiss[27] and Altman[28]. However, in this indicator it is possible to observe the effect of the low prevalence of tumors such as mesothelioma, cancer of the pleura and cancers of unknown primary site. This effect has led to lower agreement values, if compared to the specific agreement for more prevalent tumors, one of the paradoxes of the Cohen's Kappa coefficient described by Feinstein and Cicchetti[29].

Due to low numbers, to further explore the PPV for mesothelioma and cancer of the pleura, we restricted the total number of observations in the contingency tables for specific diagnoses, which could hide these tumors: lung (C34), heart, thymus and cancer of the pleura (C38), cancer in the connective tissue and other thoracic soft tissues (C49.3), abdominal soft tissues (C49.4), pelvic soft tissues (C49.5), other sites and poorly defined sites (C76.1, C76.2, and C76.3), and unknown primary site (C80). This exercise showed there were no changes in sensitivity, specificity and PPV, since the denominators were high. Specifically, the records of cancer of the pleura (C38.4) in SIH/SUS presented very low sensitivity (Table 3): of the 47 cases, only 2 were confirmed by the HCR (Supplementary Table 4). Possibly, the record of hospitalizations that was not corroborated by the cancer record is owed to the fact that the pleura is a common site of metastasis for other tumors. A study approaching the accuracy of death certificates including cancer showed that when there is a higher percentage of death records caused by a specific type of cancer, in comparison to specific cancer records, the death certificates possibly express an overestimation of the disease[30].

There is a potential to use the data from SIH/SUS for epidemiological purposes. Veras and Martins[16] analyzed 1,934 records of Hospitalization Authorizations (AIH), comparing the diagnoses in medical charts of private hospitals associated with the public network. The authors found good agreement for the sociodemographic and administrative data; however, for the selected diagnoses, there was variation according to aggregation: Kappa = 0.81 (95%CI 0.77 – 0.85) with three digits, and Kappa = 0.72 (95%CI 0.68 – 0.76) with four digits. Mathias and Soboll[17], in another study of agreement between the ICDs that were present in the AIHs of 1,595 hospitalizations, considering the diagnoses in medical records as comparison, found low three-digit aggregation agreement between the diagnoses of neoplasms (Kappa = 0.46 95%CI 0.34 – 0.57), unlike the diseases of the circulatory system (Kappa = 0.91 95%CI 0.88 – 0.94). The authors concluded that the agreement was higher for more common diagnoses. A study on the quality of data in deaths caused by acute myocardial infarction in the city of Rio de Janeiro, comparing the records of SIH/SUS with those of SIM, identified the absence of records of deaths in emergency services, the use of other diagnoses as the primary cause, and the lack of secondary diagnoses, besides the error in the codification of the type of discharge in the AIH[31]. The authors commented that, unlike the reality of SIM, there are no clear rules for the inclusion and revision of ICD codes in the hospitalization forms, and the administrative staff is responsible for filling out the clinical data, without attending any specific training[16,17,31].

The systems of disease classification used in both bases together are different. SIH/SUS uses ICD-10, which represents specific diseases, whereas HCR uses ICD-O3, a biaxial classification that considers the topography and morphology of cancers, therefore being prone to different combinations. Even though they are different, in a study of integration, the ICD-10 and ICD-O3 classifications agreed conceptually, in 88%[32]. In the present analysis, the codes of interest presented direct topography correspondence between classifications, except for mesothelioma, identified in HCR for its morphology.

This study shows the better agreement between the information in SIH/SUS and in HCR/SP for some malignant neoplasms of higher incidence, and low agreement for rare or poorly defined cancers. However, these results were obtained by comparing two databases in which the information is consolidated; it was not possible to access and verify the clinical information in all hospital units of occurrence. On the other hand, the high number of analyzed pairs and the high proportion of diagnoses with histological confirmation in HCR/SP suggest that the findings are reliable. It is important to emphasize that, unlike the scenario for AIHs, the registry of cancer cases in the database of HCR follows national and international protocols, based on trained cancer registers and on the information collected in medical charts[7-9].

By assessing the reliability of underlying causes of death due to cancer between PBCR and SIM in Goiânia, Goiás, Oliveira et al.[33] observed that rare tumors (with low incidence) presented low agreement; however, when the values were adjusted both by prevalence and by bias (PABAK), the agreement increased. The same trend was observed here.

Among the limitations of this study, it is important to mention the analysis restriction from 2007 to 2014, even though SIH/SUS has had computerized data since 1984, and HCR/SP, since 2000. However, due to the absence of a single identifier between databases, the analysis was limited after 2007, when the database of SIH/SUS present better information regarding the variable name of the individual's mother, which is an important distinction for homonyms individuals[34]. Another limitation is the small number of records for rare tumors, showing the limitations of Cohen's Kappa statistics, once this test is extremely sensitive to the distributions of marginal totals and can produce false results. However, the prevalence-adjusted and the bias-adjusted Kappa (PABAK) can be used to understand how sensitive the statistics for the distribution of marginal totals is[35]. Lima et al. verified that the Kappa statistics has been useful in reliability analyses of data in SIH/SUS, SIM, the Notifiable Diseases Information System (SINAN) and the Live Birth Information System (SINASC), and the analysis of validity has also been applied to these databases. From this perspective, therefore, this study considered that the two analyses are complementary in their interpretation[36]. Before the analyses presented here, the agreement between the same databases was tested in a trial with 226 cases in which cancer of the pleura and mesothelioma were identified in deaths (SIH/SUS) in the state of São Paulo, and had no correspondence in SIM[6]. These cases were linked to the database of HCR/SP, and 104 true pairs were found. Likewise, the Kappa analysis, with more balanced cells, showed low agreement, and most cases were lung cancer and/or metastatic disease in the pleura. Finally, there is the limitation related to the temporality of records, since HCR receives information from hospitals of the participating network after the established diagnosis; the event of death can occur later, so it is possible that, in cases of low-lethality cancers, death may not have been related to the neoplasm, and, therefore, this information was not captured by the hospitalization record.

Even though the purposes of the majority of public HIS are mostly administrative, the use of its data in epidemiological studies should be considered. In regard to SIH/SUS, the information should be used carefully, considering specific diagnoses or the group of diagnoses to be analyzed. Because SIH/SUS is national and has good coverage of hospital events, with adjustments in the rules and, mainly, in the training of the administrative agents responsible for filling it out[15,37], it can be a rich source of information in health.

It is important to emphasize, besides raising awareness to the insertion of the proper codes of hospitalization diagnoses, the need of filling out a single key in HIS to provide more accuracy and safety in the identification of the same individuals between databases, thus contributing with the use of historical series of health information systems[1,38].

ACKNOWLEDGMENTS

The non-anonymous hospitalization databases in the public system were kindly made available by Simone Santos and Rosemeire Norye Inamine, Secretariat of Health in the State of São Paulo.

REFERENCES

1. Ali MS, Ichihara MY, Lopes LC, Barbosa GCG, Pita R, Carreiro RP, et al. Administrative Data Linkage in Brazil: Potentials for Health Technology Assessment. Front Pharmacol 2019; 10: 984. http://dx.doi.org/10.3389/fphar.2019.00984
2. Mendes ACG, Silva Junior JB, Medeiros KR, Lyra TM, Melo Filho DA, Sá DA. Avaliação do sistema de informações hospitalares - SIH/SUS como fonte complementar na vigilância e monitoramento de doenças de notificação compulsória. Inf Epidemiol Sus 2000; 9(2): 67-86. http://dx.doi.org/10.5123/S0104-16732000000200002
3. Brasil. Ministério da Saúde. Departamento de Ciência e Tecnologia. Secretaria de Ciência e Tecnologia e Insumos Estratégicos. Integração de informações dos registros de câncer brasileiros. Rev Saúde Pública 2007; 41(5): 865-8.
4. Brasil. Ministério da Saúde. 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.
5. Campos MR, Martins M, Noronha JC, Travassos C. Proposta de integração de dados do Sistema de Informações Hospitalares do Sistema Único de Saúde (SIH/SUS) para pesquisa. Inf Epidemiol Sus 2000(1); 9: 51-8. http://dx.doi.org/10.5123/S0104-16732000000100005

6. 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. http://dx.doi.org/10.1002/ajim.22846
7. Instituto Nacional de Câncer. Registros hospitalares de câncer: rotinas e procedimentos. Rio de Janeiro: INCA; 2000.
8. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer 2009; 45(5): 747-55. http://dx.doi.org/10.1016/j.ejca.2008.11.032
9. Fundação Oncocentro de São Paulo. Registro Hospitalar de Câncer: Conceitos, rotinas e instruções de preenchimento [Internet]. 2ª ed. São Paulo: Fundação Oncocentro de São Paulo; 2013 [accessed on Oct. 4, 2020]. Available at: http://www.fosp.saude.sp.gov.br:443/epidemiologia/docs/ManualRHC_2013.pdf
10. Takala J. Eliminating occupational cancer in Europe and globally [Internet]. [accessed on Dec, 4, 2020]. Available at: https://oshwiki.eu/wiki/Eliminating_occupational_cancer_in_Europe_and_globally
11. 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. http://dx.doi.org/10.1016/s1470-2045(09)70134-2
12. Kim SJ, Williams D, Cheresh P, Kamp DW. Asbestos-Induced Gastrointestinal Cancer: An Update. J Gastrointest Dig Syst 2013; 3: 135. http://dx.doi.org/10.4172/2161-069X.1000135
13. Camidge DR, Stockton DL, Bain M. Factors affecting the mesothelioma detection rate within national and international epidemiological studies: insights from Scottish linked cancer registry-mortality data. Br J Cancer 2006; 95(5): 649-52. http://dx.doi.org/10.1038/sj.bjc.6603293
14. Labréche F, Case BW, Ostiguy G, Chalaoui J, Camus M, Siemiatyki J. Pleural meothelioma surveillance: validity of cases from a tumour registry. Can Respir J 2012; 19(2): 103-7. https://doi.org/10.1155/2012/650935
15. Bittencourt SA, Camacho LAB, Leal MC. O Sistema de Informação Hospitalar e sua aplicação na saúde coletiva. Cad Saúde Pública 2006; 22(1): 19-30. https://doi.org/10.1590/S0102-311X2006000100003
16. 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
17. Mathias TAF, Soboll MLMS. Confiabilidade de diagnósticos nos formulários de autorização de internação hospitalar. Rev Saúde Pública 1998; 32(6): 26-32. https://doi.org/10.1590/S0034-89101998000600005
18. Machado JP, Martins M, Leite IC. Quality of hospital databases in Brazil: some elements. Rev Bras Epidemiol 2016; 19(3): 567-81. https://doi.org/10.1590/1980-5497201600030008
19. World Health Organization. ICD-10: International Statistical Classification of Diseases and Related Health Problems: tenth revision. 2ª ed. Genebra: WHO; 2004.
20. Percy C, van Holten V, Munir C, editores. CID-O: Classificação Internacional de Doenças para Oncologia. 3ª ed. São Paulo: EDUSP; 2005.
21. National Program of Cancer Registries. Registry Plus™ Link Plus Features and Future Plans [Internet]. Atlanta: Centers for Disease Control and Prevention; 2018 [accessed on Jan. 11, 2019]. Available at: https://www.cdc.gov/cancer/npcr/tools/registryplus/lp_features.htm
22. Campbell KM, Deck D, Krupski A. Record linkage software in the public domain: A comparison of Link Plus, The Link King, and a 'basic' deterministic algorithm. Health Informat J 2008; 14(1): 5-15. https://doi.org/10.1177/1460458208088855
23. Garvin JH, Herget KA, Hashibe M, Kirchhoff AC, Hawley CW, Bolton D, et al. Linkage between Utah All Payers Claims Database and Central Cancer Registry. Health Serv Res 2019; 54(3): 707-13. https://doi.org/10.1111/1475-6773.13114
24. Cohen JA. Coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20(1): 37-46. https://doi.org/10.1177%2F001316446002000104
25. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993; 46(5): 423-9. https://doi.org/10.1016/0895-4356(93)90018-v
26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33(1): 159-74. https://doi.org/10.2307/2529310
27. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971; 76(5): 378-82.
28. Altman DG. Practical statistics for medical research. Londres/Nova York: Chapman and Hall; 1991.
29. Feinstein AR, Cicchetti DV. High agreement butlow kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990; 43(6): 543-9. https://doi.org/10.1016/0895-4356(90)90158-l
30. 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
31. Melo ECP, Travassos C, Carvalho MS. Qualidade dos dados sobre óbitos por infarto agudo do miocárdio, Rio de Janeiro. Rev Saúde Publica 2004; 38(3): 385-91. https://doi.org/10.1590/S0034-89102004000300008

32. Nikiema J, Jouhet V, Mougin F. Integrating cancer diagnosis terminology based on logical definitions of SNOMED CT concepts. J Biomed Inform 2017; 74: 46-58. https://doi.org/10.1016/j.jbi.2017.08.013
33. Oliveira PPV, Silva GA, Curado MP, Malta DC, Moura L. Confiabilidade da causa basica de obito por cancer entre Sistema de Informacoes sobre Mortalidade do Brasil e Registro de Cancer de Base Populacional de Goiania, Goias, Brasil. Cad Saúde Pública 2014; 30(2): 296-304. https://doi.org/10.1590/0102-311X00024813
34. Peres SV, Latorre MRDO, Tanaka LF, Michels FAS, Teixeira MP, Coeli CM, et al. Melhora na qualidade e completitude da base de dados do Registro de Câncer de Base Populacional do município de São Paulo: uso de técnicas de linkage. Rev Bras Epidemiol 2016; 19(4): 753-65. https://doi.org/10.1590/1980-5497201600040006
35. Flight L, Julious SA. The disagreeable behaviour of the kappa statistic. Pharm Stat 2015; 14(1): 74-8. https://doi.org/10.1002/pst.1659
36. Lima CRA, Schramm JMA, Coeli CM, Silva MEM. Revisão das dimensões de qualidade dos dados e métodos aplicados na avaliação dos sistemas de informação em saúde. Cad Saúde Pública 2009; 25(10): 2095-109. https://doi.org/10.1590/S0102-311X2009001000002
37. 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
38. Queiroz OV, Guerra Júnior AA, Machado CJ, Andrade EIG, Meira Junior W, Acurcio FA, et al. Relacionamento de registros de grandes bases de dados: estimativa de parâmetros e validação dos resultados, aplicados ao relacionamento dos registros das autorizações de procedimentos ambulatoriais de alta complexidade com os registros de sistema de informações hospitalares. Cad Saúde Coletiva 2010; 18(2): 298-308.

Received on: 01/13/2021
Revised on: 04/07/2021
Approved on: 04/23/2021

Authors' contributions: Silva DRM participated in the planning, analysis, interpretation and discussion of results, and writing the article. Luizaga CTM participated in the analysis and interpretation of results and the critical review of the content. Toporcov TN participated in the discussion of results, writing and critical review of the content. Algranti E participated in the planning, interpretation and discussion of the results, and writing the article. All authors approved the final version to be published.