DOI: 10.1590/1980-5497201600040006

ORIGINAL ARTICLE / *ARTIGO ORIGINAL*

Quality and completeness improvement of the Population-based Cancer Registry of São Paulo: linkage technique use

Melhora na qualidade e completitude da base de dados do Registro de Câncer de Base Populacional do município de São Paulo: uso das técnicas de linkage

Stela Verzinhasse Peres[I], Maria do Rosário Dias de Oliveira Latorre[I], Luana Fiengo Tanaka[I], Fernanda Alessandra Silva Michels[II], Monica La Porte Teixeira[III], Claudia Medina Coeli[IV], Márcia Furquim de Almeida[I]

ABSTRACT: The availability of large computerized databases on health turned the record linkage technique into an alternative for different study designs. This technique provides the creation of more complete databases, at low operational costs. ***Objective:*** The aim of this study was to improve the quality of information and data completeness through probabilistic and deterministic record linkage between Population-based Cancer Registry of São Paulo (PBCR-SP) for incident cancer cases, death database and drugs/medical procedures database. ***Methods:*** We used the database of the PBCRP-SP composed of 343,306 incident cancer cases in the municipality of São Paulo in the period between 1997 and 2005 with ages raging from under 1 to 106 years old, from both sexes. Three databases were used for linkage, namely Improvement Program for Mortality Information in São Paulo city (PRO-AIM), Authorization of Procedures of High Complexity/Cost of Outpatient Information System from the Unified Health System (APAC-SIA/SUS), and Foundation State System of Data Analysis (Foundation SEADE). Crude incidence (CIR) and mortality rates (CMR) were calculated and overall survival analysis was performed using the Kaplan–Meier method. ***Results:*** After record linkage, it was possible to observe gain of 4.3% for the CIR and 25.8% for CMR. The overall survival analysis showed that before record linkage there was an underestimation of the probability of being alive for all variables ($p < 0.001$). ***Conclusion:*** The linkage techniques contributed with the improvement of the quality of RCBP-SP information both on completeness of data, as in defining the vital status of the patient. In addition, the results found in this study reflect the ability of databases when worked jointly, providing subsidies for various types of studies and information for planning policies and strategic actions.

Keywords: Health information systems. Medical record linkage. Neoplasms. Survival analysis. Incidence. Mortality.

[I]School of Public Health, *Universidade de São Paulo* – São Paulo (SP), Brazil.

[II]Harvard University – Cambridge (MA), United States of America.

[III]Fundação Sistema Estadual de Análise de Dados – São Paulo (SP), Brazil.

[IV]*Instituto de Estudos de Saúde Coletiva, Universidade Federal do Rio de Janeiro* – Rio de Janeiro (RJ), Brazil.

Corresponding author: Stela Verzinhasse Peres. Rua Benjamim de Laborde, 131, Jardim São Ricardo, CEP: 05143-140, São Paulo, SP, Brasil. E-mail: sverzinhasse@yahoo.com.br

Conflict of interests: nothing to declare – **Financial support:** the Population-based Cancer Registry in the city of São Paulo provided a computer for the linkage and hired two research assistants.

RESUMO: A disponibilidade de grandes bases de dados informatizadas em saúde tornou a técnica de *linkage* uma alternativa para diferentes tipos de estudos, proporcionando a geração de uma base de dados mais completa e de baixo custo operacional. ***Objetivo:*** Melhorar a qualidade e a completitude dos casos incidentes de câncer por meio dos *linkages* probabilístico e determinístico entre o Registro de Câncer de Base Populacional de São Paulo (RCBP-SP), o banco de dados de óbitos e de Autorização e Procedimentos de Alta Complexidade. ***Método:*** Foi utilizado o banco de dados do RCBP-SP, composto de 343.306 casos de câncer incidentes no município de São Paulo entre 1997 e 2005, com idades entre 1 e 106 anos, de ambos os sexos. Para o *linkage* foram utilizadas três bases de dados, a saber: do Programa de Aprimoramento de Mortalidade no Município de São Paulo (PRO-AIM), da Fundação SEADE e da Autorização e Procedimentos de Alta Complexidade/Custo do Sistema de Informação Ambulatorial do Sistema Único de Saúde (APAC-SIA/SUS). Foram analisadas os coeficientes brutos de incidência (CBI) e mortalidade (CBM) e a sobrevida global pela técnica de Kaplan-Meier. ***Resultados:*** Após o *linkage*, verificou-se um ganho de 4,3% para a CBI e 25,8% para a CBM. Na análise de sobrevida global antes do *linkage* havia uma subestimação da probabilidade de estar vivo para todas as variáveis analisadas ($p < 0,001$). ***Conclusão:*** As técnicas de *linkage* contribuíram para a melhora da qualidade da informação do RCBP-SP tanto na completitude das variáveis quanto na definição do *status* vital do paciente, refletindo a capacidade das bases de dados, quando trabalhadas de maneira conjunta, de fornecerem subsídios para diversos tipos de estudos e informações para o planejamento de ações políticas e estratégicas.

Palavras-chave: Sistemas de informação em saúde. Registro médico coordenado. Neoplasias. Análise de sobrevida. Incidência. Mortalidade.

INTRODUCTION

Population-based Cancer Registry (RCBP) are institutions that aim at promoting epidemiological cancer survey and contributing with the planning of health services[1,2]. The role of these services to understand the magnitude of the event — due to its economic, psychological, and social burden — is essential to address health policies, making the control of the quality of information, on which the actions are based, very important.

Besides their incidence, these institutions should produce information about mortality and survival, considering that this type of data are hardly observed due to the lack of the vital status of the patient — dead or alive — and the date of the last record[1,3].

Completeness, accuracy, and proportion of lost data are indications of the quality of an information system. Several factors can contribute with the existence of problems, such as inconsistency and low quality, multiple sources of data, limited computer, financial and human resources, representation of complex data, volume of stored data, evolution of need for data, rules of data entry that are very restricted or annulled, among others[4].

Specifically regarding the assessment of quality of PBCR in the city of São Paulo (RCBP-SP), it follows the guidelines from the International Agency for Research on Cancer (IARC), as follow: histopathological diagnosis (> 70%), notification only through death certificate (< 20%), ignored age (< 10%), unspecific location (< 10%), and mortality/incidence ratio (between 20 and 30%)[1,2].

RCBP-SP was founded in 1969, being one of the oldest and more important in the country due to its longevity and coverage. Among its practices there is the collaboration with public policies of the city, subsidizing the planning and establishing priorities in cancer control, through analyses of distribution and tendencies in the city. Throughout its existence, this registry presented periods of interruption in the collection, and only in 1977 this activity was uninterrupted[5]. Limited human and financial resources made it difficult to have complete variables regarding the vital status — date of death and last data — as well as the aggregation of new data, such as mother's name and basic cause of death.

Currently, the new cases and the updated status of the patients are collected from 301 sources of notification; 245 of them are periodically visited. This collection is conducted in hospitals, clinics, services of death verification, and laboratories of pathological anatomy. Data from the Hospital Cancer Records (RHC) are provided by Fundação Oncocentro de São Paulo (FOSP)[1].

By using probabilistic and deterministic record linkage, widely used in different studies[6-12], the restructure of a more consistent and complete database was proposed, based on existing data form other databases, in order to improve this system of cancer epidemiological information in the city of São Paulo. Therefore, it would be possible to provide reliable information, thus contributing with assertive public actions.

So, the objective of this study was to improve the quality of information of RCBP-SP and assess its completeness from 1997 to 2005, through crude incidence rates (CIRs) and crude mortality rates (CMRs) by cancer, and global accumulated survival rates by cancer, according to sex, age group, and topography before and after the database linkage process.

The databases used for the linkage were that of the Software of Improvement of Mortality in the city of São Paulo (PRO-AIM), death in the State of São Paulo, provided by the foundation State System of Data Analysis (Foundation SEADE), and the Authorization and Procedures of High Complexity in the Outpatient Information System of the Unified Health System (APAC-SIA/SUS), referring to all high complexity/cost procedures.

METHODS

This study assessed the cohort of new cancer cases in RCBP-SP from 1997 to 2005, composed of 343,306 incident cases, according to the International Classification of Diseases for Oncology (ICD-O) — 3rd Edition (C00.0 to C80.9) — in the city of São Paulo, among individuals aged from 1 to 106 years old, both sexes.

The database of PRO-AIM was used for the linkage including 767,752 deaths, except fetal ones, occurred between 1997 and 2007, in the city of São Paulo. The choice of analyzing two years besides the follow-up of RCBP-SP was in order to capture a higher number of deaths. Also, this was the period made available by the Coordination of Epidemiology

and Information (CEINFO). Deaths in the state were provided by Foundation SEADE, accounting for 2,308,081, except for fetal ones, occurred between 2000 and 2009. This period was also chosen for the availability and use of this base, justified by the evasion of deaths of 4.3% in the city[13].

The third database was APAC-SIA/SUS, which originally contained 31,743,533 records. However, the process of identification of patients who showed up more than once was conducted considering that, for the probabilistic linkage, this database should present a single record per each patient, containing data about the last time a procedure or medication was requested[14]. Therefore, the final base comprised 863,735 patients, of all ages and both sexes, who underwent any high complexity/cost treatment/procedure between August 2003 and December 2007, in the city of São Paulo. The choice of APAC-SIA/SUS is owed to the need of identifying living patients, because these are the ones who underwent or performed ay high complexity/cost procedure. Data referring to APAC-SIA/SUS were requested to the State Secretariat of Health of São Paulo (SES-SP) for 1997–2007; however, it was only possible to obtain the information from August 2003 to December 2007, provided by the Municipal Secretariat of Health in São Paulo (SMS-SP).

The probability technique was applied to improve the completeness of variables and to identify the deaths between the databases of PRO-AIM and RCBP-SP, and, for the evaluation, the living status between APAC-SIA/SUS and RCBP-SP. The deterministic method was applied in the completeness of the variables to identify the deaths by Foundation SEADE and in the identification of new cases.

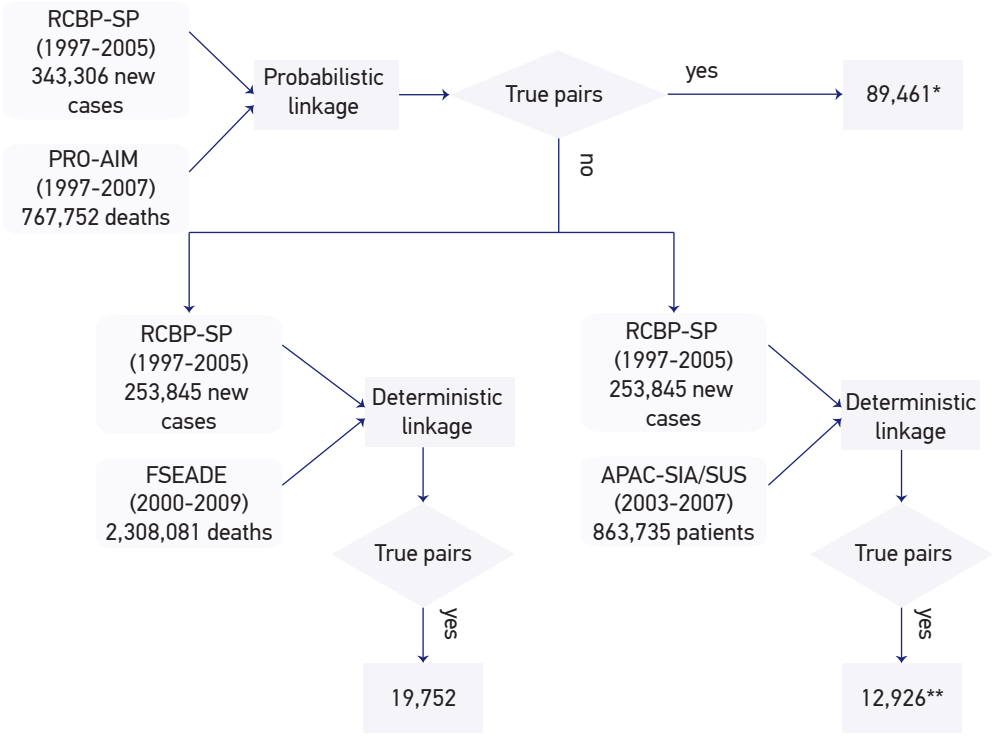
The stages were executed in the software Reclink III, version 3.1.6[15], for databases from RCBP-SP, PRO-AIM, and APAC-SIA/SUS. The variables used in the matching process were the name of the patient and date or year of birth. As to blocking, the variables were sex, soundex of the first name (PBLOCO), soundex of the last name (UBLOCO), and year of birth. As confirmatory variables for accepting a real match, when available, the following were used: address, mother's name, basic cause of death, and type of procedure/treatment. The combination between the matching and blocking fields resulted in 14 linkage strategies.

The following were considered as true matches between the linkage of RCBP-SP versus PRO-AIM: agreement scores between 20.8 and 23.3; and for linkage between RCBP-SP versus APAC-SIA/SUS, scores between 20.7 and 23.0. Intermediate scores between total disagreement (-8.90 and -6.1, respectively, PRO-AIM and APAC-SIA/SUS) and agreement values were manually revised by a team of three researchers[15].

For the database of the SEADE Foundation, the entire process was developed in Visual Basic, connecting bases hosted by a SQL server. In this process, two criteria were elaborated for the selection of real matches: equality (identical fulfilling) and similarity (agreement ≥ 80% — visual revision)[16]. In the automatic selection, pairs presenting total equality in the variables name, date of birth, address, and date of death, or in the variables name, date of birth and address, were accepted as real matches.

Figure 1 represents the linkage stages according to the database entry. The first linkage was conducted between RCBP-SP and PRO-AIM, after the removal of cases found and classified as death. The linkage between RCBP-SP and Foundation SEADE and between RCBP-SP and APAC-SIA/SUS was simultaneous.

Absolute and relative frequencies were calculated for statistical analysis, as well as CIR and CMR by cancer before and after the linkage. The denominator was the total population of the city of São Paulo for the analyzed period, according to sex and age group. The accumulated global survival rate was performed by the Kaplan–Meier product-limit estimator, and the comparison between before and after used the log rank test. Time of survival was calculated as the difference between date of last data or date of death and date of diagnosis. In this analysis, time was blocked in seven years, considering as censorship patients who died after this period. To accomplish minimum time of five years of follow-up, only patients diagnosed with cancer between 1997 and 2002 were included. In the analysis by topography, cancers of highest incidence were included, as follows: stomach, colon-rectum, lung, breast, uterus, and prostate. A 5% descriptive level was established for statistical significance. Data were analyzed in the software Statistical Package for Social Sciences (SPSS), version 17.0.



*Excluded for the next linkage process; **of the true pairs identified in this linkage, 1,720 patients died and were identified in the linkage between the Population-Based Cancer Registry in the city of São Paulo (RCBP-SP) and Foundation SEADE.

Figure 1. Flowchart of the probabilistic and deterministic linkage process between databases.

This study was approved by the Research Ethics Committee in the Public Health School at Universidade de São Paulo (n. 0086.0.207.000-08) and SMS-SP (n°0064.0.162.000-09).

RESULTS

At the end of each linkage (Figure 1), it is possible to observe the total number of related pairs, accounting for 122,139 records. Table 1 presents the completeness of the RCBP-SP database before and after the linkage. The column "before" presents the initial values of RCBP-SP, and the column "after" shows the sum between the amount of initial information

Table 1. Number of cases with non-ignored information in the Population-Based Cancer Registry in the city of São Paulo and their respective gains in absolute and relative frequency. Population-Based Cancer Registry in the city of São Paulo, 1997 to 2005.

Linkage	Before	After	≠
	n	n	percentage
PRO-AIM			
Date of birth	220,176	224,719	2.1
Address	81,057	137,118	69.2
Mother's name	1,801	44,990	2.398.1
Date of death	103,910	120,895	16.3
APAC-SIA/SUS			
Date of birth	220,176	221,968	0.8
Address	81,057	88,988	9.8
Mother's name	1,801	14,477	703.8
Date of last data	11,462	20,628	80.0
Foundation SEADE			
Date of birth	220,176	223,288	1.4
Address	81,057	101,209	24.9
Mother's name	1,801	22,264	1.136.2
Date of death	103,910	119,893	15.4
GENERAL			
Date of birth	220,176	229,470	4.2
Address	81,057	163,310	101.5
Mother's name	1,801	76,332	4.138.3
Date of death	103,910	136,626	31.5
Date of last data	11,462	20,628	80.0

Linkage: values observed between the linkage with RCBP-SP; PRO-AIM: Improvement Program for Mortality Information in São Paulo; APAC-SIA/SUS: Authorization of Procedures of High Complexity/COST of Outpatient Information System from the Unified Health System; Foundation SEADE: Foundation State System of Data Analysis; GENERAL: total of values observed between linkages with RCBP-SP.

with the completed data. By comparing the before and after moments in the RCBP-SP linkage with the database of PRO-AIM, a 69.2% gain was observed for the variable address, and 16.3% for date of death.

As to the linkage with the database from Foundation SEADE, 24.9 and 15.4% gains were observed, respectively, for the variables address and date of death. Regarding linkage with the database from APAC-SIA/SUS, the variable date of last data gained 80.0% in completeness of data. In general, it is important to mention that all relations were essential for the strategy of completeness of information in the RCBP-SP database. As to follow-up, date of death presented a 31.5% gain (Table 1).

From the linkage between the databases of RCBP-SP versus PRO-AIM versus Foundation SEADE, the variable cause of death was aggregated. It is important to mention that, of the related cases, the basic cause of death was mostly cancer (86.4%). In the final status of patients registered in the RCBP-SP database, there was a 3.4% gain in new cases, going from 343,306 to 354,957 incident cases. Considering the three sources of information at the end of the process, there was a reduction in the loss of follow-up both for the living status (before = 3.3%, and after = 5.8%) and the death status (before = 30.3%, and after = 38.5%). However, 55.7% remained without information as to current status.

It is also important to mention that Table 2 presents CIR and CMR before and after linkage. To calculate these coefficients, non-melanoma skin cases were excluded (C44.0 to C44.9). In this analysis, numerators were the total of new cases (before = 272,644 and after = 284,280) and the total of deaths (before = 95,899 and after = 120,652), occurred between 1997 and 2005. A single record of each patient was considered to calculate CMR. CRI, which was 288.2 new cases per 100,000 inhabitants, turned to 300.5 after linkage between databases, so there was a 4.3% gain in the coefficient. By analyzing the CMR, a 25.8% difference is observed in general (Table 2).

As to accumulated survival rates, the odds were underestimated. Before linkage, the probability of being alive in seven years was 7.8%, increasing to 13.0% (Table 3). Likewise, underestimation was observed for both genders. Among men, after two years of follow-up, there was a percentage difference of 59.3%, in which the probability of being alive before the linkage was 26.3%, increasing to 41.9% after the relation ($p < 0.001$). The same was observed for all age groups ($p < 0.001$) before and after linkage. The probability of being alive after

Table 2. Crude incidence and crude mortality rates by cancer (per 100,000 inhabitants) before and after linkage, except for non-melanoma skin cancer. Population-Based Cancer Registry in the city of São Paulo, 1997 to 2005.

Coefficients	Before		After		≠
	n	coefficient	n	coefficient	percentage
CIR	272,644	288.2	284,280	300.5	4.3
CMR*	95,899	101.4	120,652	127.5	25.8

CIR: crude incidence rate; CMR: crude mortality rate; *deaths found in the databases of the Improvement Program for Mortality Information in São Paulo (PRO-AIM), in the Foundation State System of Data Analysis (Foundation SEADE) and the 56 cases of the Authorization of Procedures of High Complexity/COST of Outpatient Information System from the Unified Health System (APAC-SIA/SUS).

seven years went from 2.3%, before linkage in the age group of 0–14 years, to 17.6% after linkage (Table 3). For the age group of 60–84 years, the probability of survival before the linkage was 29.9%, whereas after the relation this probability was 45.8%.

Regarding topographies (Table 4), there is a statistically significant underestimation in the probabilities of survival for all analyzed periods ($p < 0.001$). The survival probability for prostate cancer was 5.2% before the linkage, and after it increased to 17.3%.

DISCUSSION

Our results indicated that the application of both techniques contributed with the improvement in the quality of information in RCBP-SP, both regarding completeness of variables and in the definition of the vital status of patients. CIR and CMR also improved, as well as the survival analyses that brought another profile of the prognosis.

Table 3. Accumulated global survival rate, according to sex and age group, before and after linkage, blocked in seven years. Population-Based Cancer Registry in the city of São Paulo, 1997 to 2005.

Variables	Linkage	No. of cases	No. of deaths	Probability of accumulated global survival (% in years)					p-value (K-M)
				1st	2nd	3rd	5th	7th	
Generaç	Before	56,237	48,613	48.8	31.9	22.7	12.3	7.8	< 0.001
	After	78,831	64,120	61.1	46.6	37.0	23.2	13.0	
Sex									
Female	Before	27,865	22,900	54.7	37.6	27.6	16.4	11.8	< 0.001
	After	39,415	30,379	65.7	51.2	41.2	27.0	16.4	
Male	Before	28,372	25,713	43.0	26.3	18.0	8.3	3.5	< 0.001
	After	39,416	33,741	56.5	41.9	32.8	19.5	9.7	
Age Group (years)									
0 – 14	Before	711	648	45.6	26.3	16.2	6.5	2.3	< 0.001
	After	1,035	756	59.7	44.3	34.6	24.5	17.6	
15 – 29	Before	1,408	1,147	52.3	34.8	24.6	14.7	9.9	< 0.001
	After	1,940	1,403	63.1	48.7	39.2	27.5	18.8	
30 – 44	Before	5,532	4,300	58.0	40.6	30.8	19.8	15.0	< 0.001
	After	7,675	5,637	67.2	52.6	42.9	28.9	18.7	
45 – 59	Before	14,503	12,374	50.9	33.1	24.1	13.9	9.5	< 0.001
	After	19,548	15,591	61.9	46.5	37.0	24.0	14.2	
60 – 84	Before	31,157	27,716	46.4	29.9	20.7	10.3	5.7	< 0.001
	After	44,239	37,007	59.9	45.8	36.2	22.2	11.7	
≥ 85	Before	2,526	2,386	38.5	22.1	14.7	5.2	1.9	< 0.001
	After	4,021	3,699	54.9	39.7	30.7	15.0	6.1	

K-M: Kaplan–Meier test.

It is worth to remember that the choice for the type of linkage used considered the objectives of the relation and its epidemiological implications. The deterministic linkage was used to identify new cases. This technique is based on the use of rules to classify the links formed in true pairs. The development of rules by experts on the subject and on the basis may lead to accurate results[17]. The probabilistic linkage is known for being less transparent, given the fact that the power of discrimination of the pairs is based on scores built according to the discriminatory capacity of each field, being vulnerable to false positive and negative errors.

However, this technique becomes a viable solution when the records to be related require quality. The variables may present errors and flaws in information, lack of a univocal indicator or few variables to relate the records in databases with many cases[17-19].

In this study, the results were analyzed by the identification of true pairs, according to completeness and quality of the databases. For deaths, the linkage process between the databases of RCBP-SP versus PRO-AIM detected that 81% already had the date of death. RCBP-SO annually conducted the process of probabilistic linkage with PRO-AIM. However, this relation was only carried out with cases presenting death by cancer, in order to conclude the follow-up and check for possible incident cases that were not detected by RCBP-SP. It is worth to mention that this process was conducted in accordance with the recommendations of IARC, which consider that patients who were not identified in the death database are alive[2]. Besides the complemented information, other data were added, such as mother's name, address, and basic cause of death. The substantial gain for the

Table 4. Accumulated Global Survival rate, according to topography, before and after linkage, blocked in seven years. Population-Based Cancer Registry in the city of São Paulo, 1997 to 2005.

Topography	Linkage	No. of cases	No. of deaths	Probability of accumulated global survival (% in years)					p-value (K-M)
				1st	2nd	3rd	5th	7th	
Stomach	Before	4,990	4,877	25.7	11.2	6.0	1.9	0.5	< 0.001
	After	6,029	5,693	36.8	23.0	16.6	9.0	3.9	
Colon and Rectum	Before	5,643	5,373	46.9	26.9	14.6	4.1	0.9	< 0.001
	After	7,318	6,630	56.7	39.3	26.6	13.3	6.2	
Lung	Before	5,622	5,554	29.0	13.1	7.0	2.3	0.5	< 0.001
	After	6,543	6,317	37.2	21.6	14.3	7.0	2.6	
Breast	Before	7,414	5,188	74.8	59.0	46.5	30.4	23.5	< 0.001
	After	10,963	7,106	82.0	69.6	58.6	41.2	27.8	
Uterus	Before	4,477	3,758	55.8	34.4	24.3	14.5	12.1	< 0.001
	After	6,242	4,858	66.7	48.6	37.8	24.8	16.3	
Prostate	Before	3,485	3,005	63.1	45.1	32.8	14.7	5.2	< 0.001
	After	6,326	4,850	77.5	64.1	52.7	32.8	17.3	

K-M: Kaplan-Meier test.

variable mother's name is essential to improve the quality of RCBP-SP database, since it is the main way to distinguish homonyms.

As to the completeness of the variable address, conducted via death records and data from APAC-SIA/SUS, some considerations are important because the address of death or current address may not be the same as that of the time of diagnosis, which implies on the mistaken spatial distribution of cancer cases. In this context, this information is not indicated for georeferencing analyses. On the other hand, this piece of information can indicate the internal migration for cases occurred in the proximities of treatment units (deaths per occurrence), and for living patients it is relevant in case there is the need for an active search.

The basic cause of death, be it cancer or not, appeared as a new variable. Of the deaths, 87.6% had cancer as a basic cause. However, this proportion of causes of death by cancer should be analyzed carefully, for two reasons. First, there may be information bias connected to the filling out of the Death Certificate (DC). A tendency is believed to exist, among deaths of patients with cancer, in which the doctor fills out the certificate and associates the cause of death to existing causes that were not responsible for the outcome[20]. Second, in the linkage process, the basic cause of death was one of the variables used as a parameter to confirm the identification of a real pair. However, it was not used in isolation to identify a related record. To be a real pair, at least two other conditions of the variables of relationship and confirmation were required[14].

One of the main study limitations were the filling out errors in the database of APAC-SIA/SUS. It is believed that the type of record used for charging may have influenced the quality, because, after the first record, the others — for the same patient — presented flawed information, such as short names, incomplete address, CPF, and mother's name missing, or even the use of the words like "THE SAME" as a reference to a previous record.

To assess the impact caused by the linkage process on the statistics of RCBP-SP, the numbers were analyzed before and after the process. CIR was underestimated in 4.3%, so, other studies that used the linkage technique to complete the information showed the differences in the number of registered events[21,22]. Regarding CMR, there was a 25.8% gain after the linkage process. In the study by Pereira et al.[23], conducted in Rio de Janeiro between 1999 and 2001, it was observed that there was a 20.0% underestimation in the coefficient of neonatal mortality. Likewise, the study by Rafael et al.[24], based on the database of the System of Mortality Information (SIM) and the System of Hospital Information (SIH) in the state of Maranhão, showed the underestimation in neonatal and children's mortality coefficient of 24.9 and 19.8%, respectively.

As to the analysis of global survival rates of patients in RCBP-SP, after the linkage the possibilities of survival increased throughout the years analyzed for both sexes, age groups, and topographies analyzed. In this case, it is possible to mention the importance of relating not only the death databases, but also the one of living patients. From the

living patients identified via APAC-SIA/SUS, information became more qualitied, which reflected on the survival probability throughout the years. However, it is necessary to mention that even after the inclusion of information regarding the patient's status, coming from the bases of PRO-AIM, Foundation SEADE, and APAC-SIA/SUS, 55.7% of the cases remained without follow-up data — date of the last information or date of death. Once the records of death are mandatory, taking part in the corresponding base, it is possible that the linkage process conducted by this study identified a significant number of deaths. On the other hand, data contained in the base of APAC-SIA/SUS are restricted to high cost and complex procedures conducted in patients followed-up by the Unified Health System (SUS); therefore, living patients who are not being treated are not part of the base and do not have the chance of being captured.

However, SMS-SP reports that the high complexity procedures recorded in this system, due to its high cost and the limited coverage by private insurance health plans, represent 90% of the procedures conducted in the southeast region. For the city of São Paulo, it is possible to observe that the coverage rate of health insurance plans is 59.8%; however, the study conducted by the Institute of Supplementary Health Studies (IESS) identified that 26% of the people who pay for private medical care also use SUS[25,26]. So, the periodical relation between RCBP-SP with other databases that can contribute with follow-up information, especially that of living patients, should be part of the strategy for its improvement.

By observing the proportion of loss in the follow-up, another important matter is that it can change according to the topography analyzed. In this study, topographies of higher incidence and mortality were chosen. However, it is important to mention that the more aggressive the cancer, the higher the chances of observing the patient between treatment and death, given the shorter follow-up time. But patients with thyroid and breast cancer, for instance, present higher survival rates, so they can interrupt the treatment or migrate to other regions.

CONCLUSION

Given the exposed, the conclusion is that all variables in RCBP-SP were more complete, as well as the improved quality of information using CIR, CMR, and survival rates, showing the effectiveness of probabilistic and deterministic linkages. It also showed that the results found in this study reflect the capacity of databases, when worked together, to provide subsidies for several types of studies and information to plan for political actions and strategies.

It is also important to mention that unlike other countries, little is done with survival analyses, using population-based data[12,27], especially in RCBP due to the lack of follow-up. In this context, the connection between databases of administrative profile is important, such as APAC-SIA/SUS, the System of Health Information for Basic Care (SISAB), the

Journal of Individualized Outpatient Production (BPA-I), the System of Cancer Information (SISCAN), the System of Decentralized Hospital Information (SIHD), and the Regional Electoral Court (TRE), with data regarding epidemiological information such as deaths, births, and notification of diseases. This connection improves and increases the possibilities of analysis, besides creating more qualified and less expensive information, in order to improve health indicators.

REFERENCES

1. Brasil. Ministério da Saúde. Secretaria de Estado da Saúde de São Paulo. Registro de Câncer de Base Populacional de São Paulo. Câncer em São Paulo 1997-2008: incidência, mortalidade e tendência de câncer no Município de São Paulo. RCBP-SP 2011.
2. International Agency for Research on Cancer (IARC). Cancer incidence in five continents Volume-IX. IARC Scientific Publications No. 160. Lyon: IARC; 2007.
3. Roder D, Creighton N, Baker D, Walton R, Aranda S, Currow D. Changing roles of population-based cancer registries in Australia. Aust Health Rev 2015; 39(4): 425-8.
4. Data matching: concepts and techniques for record linkage, entity resolution and duplicate detection. Springer: Berlin; 2012.
5. Brasil. Ministério da Saúde. Instituto Nacional do Câncer. CONPREV. Secretaria de Estado da Saúde. Fundação Oncocentro de São Paulo. Secretaria Municipal da Saúde. PRO-AIM. Faculdade de Saúde Pública da Universidade de São Paulo (Departamento de Epidemiologia). Registro de Câncer no Brasil e sua história. São Paulo, Brasil; 2005.
6. The West of Scotland Coronary Prevention Study Group (WOSCOPS). Computerised record linkage: compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. J Clin Epidemiol 1995; 48(12): 1441-52.
7. Coeli CM, Blais R, Costa MCE, Almeida LM. Probabilistic linkage in household survey on hospital care usage. Rev Saúde Publica 2003; 37(1): 91-9.
8. Oberaigner W, Stühlinger W. Record linkage in the Cancer Registry of Tyrol, Austria. Methods Inf Med 2005; 44(5): 626-30.
9. Li B, Quan H, Fong A, Lu M. Assessing record linkage between health care and Vital Statistics databases using deterministic methods. BMC Health Serv Res 2006; 6: 48.
10. Machado JP, Silveira DP, Santos IS, Piovesan MF, Albuquerque C. Aplicação da metodologia de relacionamento probabilístico de base dados para a identificação de óbitos em estudos epidemiológicos. Rev Bras Epidemiol 2008; 11(1): 43-54.
11. Santos SLD, Silva ARV, Campelo V, Rodrigues FT, Ribeiro JF. Utilização do método *linkage* na identificação dos fatores de risco associados à mortalidade infantil: revisão integrativa da literatura. Cien Saude Colet 2014; 19(7): 2095-104.
12. Mitra D, Shaw A, Tjepkema M, Peters P. Social determinants of lung cancer incidence in Canada: a 13-year prospective study. Health Rep 2015; 26(6): 12-20.
13. Taniguchi MT, Pelaquin MHH, Latorre MRDO. Relacionamento probabilístico entre as bases de dados do registro de câncer de São Paulo e do sistema de informações de mortalidade municipal [trabalho de conclusão]. São Paulo: Faculdade de Saúde Publica da USP; 2006.
14. Peres SV, Latorre MRDO, Michels FAS, Tanaka LF, Coeli CM, Almeida MF. Determinação de um ponto de corte para a identificação de pares verdadeiros pelo método probabilístico de *linkage* de base de dados. Cad Saúde Colet 2014; 22(4): 428-36.
15. Coeli CM, Camargo Jr KR. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. Rev Bras Epidemiol 2002; 5(2): 185-96.
16. Moraes GH, Duarte EC. Análise da concordância dos dados de mortalidade por dengue em dois sistemas nacionais de informação em saúde, Brasil, 2000-2005. Cad Saúde Pública 2009; 25(11): 2354-64.
17. Waldvogel BC. Acidentes do trabalho: os casos fatais: a questão da identificação e da mensuração. Belo Horizonte: Segrac; 2002.
18. Machado CJ. A literature review of record linkage procedures focusing on infant health outcomes. Cad Saúde Pública 2004; 20(2): 362-71.

19. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. J Clin Epidemiol 2011; 64(5): 565-72.
20. Laurenti R, Mello Jorge MHP, Gotlieb SLD. Mortalidade segundo causas: considerações sobre a fidedignidade dos dados. Rev Panam Salud Publica/Pan Am J Public Health 2008; 23(5): 349-56.
21. Cavalcante MS, Ramos Jr AN, Pontes LRSK. Relacionamento de sistemas de informação em saúde: uma estratégia para otimizar a vigilância das gestantes infectadas pelo HIV. Epidemiol Serv Saúde 2005; 14(2): 127-33.
22. Drumond EF, Machado CJ, França E. Underreporting of live births: measurement procedures using the Hospital Information System. Rev Saúde Pública 2008; 42(1): 55-63.
23. Pereira APE, Gama SGN, Leal MC. Mortalidade infantil em uma amostra de nascimentos do município do Rio de Janeiro, 1999-2001; *"linkage"* com o Sistema de Informação de Mortalidade. Rev Bras Saúde Matern Infant 2007; 7(1): 83-8.
24. Rafael RAA, Ribeiro VS, Cavalcante MCV, Santos AM, Simões VMF. Relacionamento probabilístico: recuperação de informações de óbitos infantis e natimortos em localidades no Maranhão, Brasil. Cad Saúde Pública 2011: 27(7): 1371-9.
25. Brasil. Ministério da Saúde. Instituto de Estudos de Saúde Suplementar (IESS). Os custos do ressarcimento ao SUS. Saúde Suplementar em Foco. Informativo Eletrônico IESS; 2010.
26. Agência Nacional de Saúde Suplementar (ANS). Cadernos de Informação de Saúde Suplementar: Beneficiários, Operadoras e Planos; 2008. Disponível em: http://www.ans.gov.br (Acessado em 20 de outubro de 2014).
27. Tancredi MV, Holcman MM, Teixeira Jr AE, Farias NSO. Análise da sobrevida de pacientes com Aids no Estado de São Paulo. In: Dados para repensar a Aids no Estado de São Paulo: resultados da parceria entre o Programa Estadual DST/Aids e Fundação SEADE. São Paulo: DST/Aids, Fundação SEADE; 2010: 83-110.

Received on: 09/02/2015
Final version presented on: 02/11/2016
Accepted on: 05/30/2016