DOI: 10.1590/1980-5497201600040006

ARTIGO ORIGINAL / *ORIGINAL ARTICLE*

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*

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

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]

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.

[I]Faculdade de Saúde Pública, Universidade de São Paulo – São Paulo (SP), Brasil.
[II]Harvard University – Cambridge (MA), United States of America.
[III]Fundação Sistema Estadual de Análise de Dados – São Paulo (SP), Brasil.
[IV]Instituto de Estudos de Saúde Coletiva, Universidade Federal do Rio de Janeiro – Rio de Janeiro (RJ), Brasil.
Autor correspondente: Stela Verzinhasse Peres. Rua Benjamim de Laborde, 131, Jardim São Ricardo, CEP: 05143-140, São Paulo, SP, Brasil. E-mail: sverzinhasse@yahoo.com.br
Conflito de interesses: nada a declarar – **Fonte de financiamento:** o Registro de Câncer de Base Populacional do município de São Paulo disponibilizou um computador para realização dos linkages e contratou dois auxiliares de pesquisa.

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.

INTRODUÇÃO

Os Registros de Câncer de Base Populacional (RCBP) são instituições que têm por finalidade promover a vigilância epidemiológica do câncer e contribuir para o planejamento dos serviços de saúde[1,2]. O papel desses serviços no entendimento da magnitude do evento — devido ao seu fardo econômico, psicológico e social — se faz essencial para o direcionamento de políticas de saúde, tornando o controle da qualidade da informação na qual as ações são respaldadas de suma importância.

Cabe a essas instituições, além da incidência, produzir informações sobre mortalidade e sobrevida, considerando que essas são dificilmente observadas devido à ausência do *status* vital do paciente — vivo ou morto — e da data da última informação[1,3].

Completitude, acurácia e proporção de dados perdidos são indicativos da qualidade de um sistema de informação, com diversos fatores podendo contribuir para a existência de problemas, como inconsistência e baixa qualidade, múltiplas fontes de dados, recursos computacionais, financeiros e humanos limitados, representação de dados complexos, volume de dados armazenados, evolução das necessidades de dados, regras de entrada de dados muito restritas ou anuladas, entre outros[4].

Em específico, a avaliação da qualidade do RCBP do Município de São Paulo (RCBP-SP) segue os padrões da *International Agency for Research on Cancer* (IARC), a saber: diagnóstico histopatológico (> 70%), notificação somente por atestado de óbito (< 20%), idade ignorada (< 10%), localização inespecífica (< 10%) e razão mortalidade/incidência (entre 20 e 30%)[1,2].

O RCBP-SP foi fundado em 1969, sendo um dos mais antigos e importantes do país devido à sua longevidade e cobertura. Dentre suas práticas, está a colaboração com as políticas públicas do município, subsidiando o planejamento e estabelecendo prioridades no controle do câncer, por meio das análises de distribuição e de tendências no município. Ao longo de sua existência, esse registro apresentou períodos de descontinuidade na coleta e, somente em 1997, tal atividade deu-se de maneira ininterrupta[5]. Limitação de recursos humanos e financeiros dificultou a completitude das variáveis relativas ao *status* vital — data do óbito e da última informação —, bem como a agregação de novas, como nome da mãe e causa básica do óbito.

Atualmente, os casos novos e a atualização do *status* dos pacientes são coletados de 301 fontes de notificação, sendo que 245 delas são visitadas periodicamente. Essa coleta é realizada em hospitais, clínicas, serviços de verificação de óbito e laboratórios de anatomia patológica. Os dados dos Registros de Câncer Hospitalar (RHC) são fornecidos pela Fundação Oncocentro de São Paulo (FOSP)[1].

No tocante, por meio das técnicas de *linkages* probabilístico e determinístico amplamente utilizadas para diferentes estudos[6-12], foi proposta a reestruturação de uma base de dados mais consistente e completa, a partir de dados existentes em outras bases de dados, a fim de melhorar esse sistema de informação epidemiológico de câncer no município de São Paulo, propiciando informações mais fidedignas e, dessa maneira, contribuindo para ações públicas assertivas.

Assim, o objetivo deste estudo foi melhorar a qualidade da informação do RCBP-SP e avaliar a completitude, no período de 1997 a 2005, por meio dos coeficientes brutos de incidência (CBI) e de mortalidade (CBM) por câncer e pela sobrevida global acumulada por câncer, segundo sexo, faixa etária e topografia antes e após o processo de *linkage* de base de dados. As bases de dados utilizadas para o *linkage* foram a do Programa de Aprimoramento de Mortalidade do Município de São Paulo (PRO-AIM), a de óbito do Estado de São Paulo fornecida pela Fundação Sistema Estadual de Análise de Dados (Fundação SEADE) e a da Autorização e Procedimentos de Alta Complexidade do Sistema de Informação Ambulatorial do Sistema Único de Saúde (APAC-SIA/SUS), referente a todos os procedimentos de alta complexidade e custo.

MÉTODOS

Esta pesquisa avaliou a coorte de novos casos de câncer do RCBP-SP no período de 1997 a 2005, composta por 343.306 casos incidentes de acordo com a Classificação Internacional de Doenças para Oncologia (CID-O) — 3° Edição (C00.0 a C80.9) — no município de São Paulo, entre indivíduos de 1 a 106 anos de idade, de ambos os sexos.

A base de dados do PRO-AIM foi utilizada para o *linkage*, compreendendo 767.752 óbitos, exceto os fetais, ocorridos entre 1997 e 2007, no município de São Paulo. A escolha de dois anos além do seguimento do RCBP-SP foi para captar um número maior de óbitos,

além de este ser o período disponibilizado pela Coordenação de Epidemiologia e Informação (CEINFO). Os óbitos ocorridos no estado foram disponibilizados pela Fundação SEADE, totalizando 2.308.081, exceto os fetais, ocorridos entre os anos de 2000 e 2009. A escolha deste período também foi pela disponibilidade e o uso dessa base é justificado pela taxa de evasão de óbitos de 4,3 % do Município[13].

A terceira base de dados relacionada foi a APAC-SIA/SUS, que originalmente continha 31.743.533 registros. Porém, o processo de identificação de pacientes que apareceram mais de uma vez foi realizado, considerando que, para o *linkage* probabilístico, essa base de dados deveria apresentar um único registro para cada paciente, contendo a data da última vez em que um procedimento ou medicamento foi solicitado[14]. Dessa forma, o banco final foi composto por 863.735 pacientes, de todas as idades e de ambos os sexos, que realizaram qualquer tratamento/procedimento de alta complexidade e custo entre o período de agosto de 2003 e dezembro de 2007 no município de São Paulo. A escolha da APAC-SIA/SUS se deve à necessidade de identificar os pacientes vivos, pois aí encontram-se aqueles que receberam ou realizaram qualquer procedimento de alta complexidade e custo. Foram solicitados à Secretaria Estadual de Saúde de São Paulo (SES-SP) os dados referentes à APAC-SIA/SUS para o período de 1997 a 2007, porém só foi possível obter as informações do período de agosto de 2003 a dezembro de 2007, disponibilizadas pela Secretaria Municipal de Saúde de São Paulo (SMS-SP).

A técnica probabilística foi aplicada na melhoria da completitude das variáveis e na identificação dos óbitos entre a base de dados do PRO-AIM e o RCBP-SP e, para a avaliação, o *status* vivo entre a APAC-SIA/SUS e o RCBP-SP. O método determinístico foi aplicado na completitude das variáveis para a identificação dos óbitos pela Fundação SEADE e na identificação de casos novos.

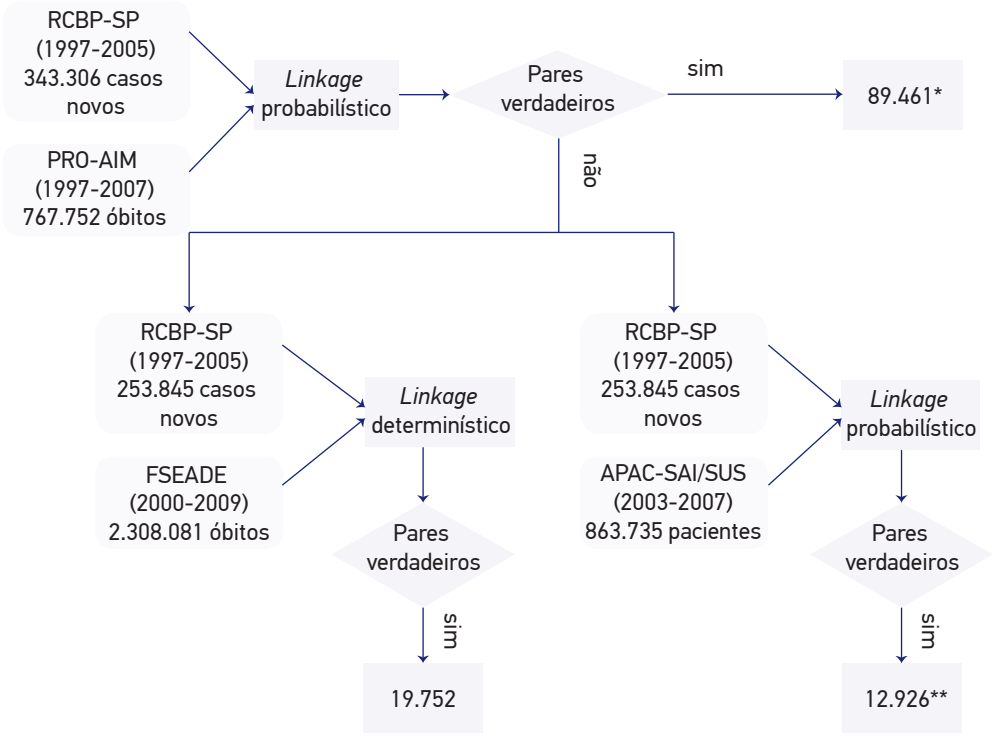
As etapas foram executadas no programa Reclink III, versão 3.1.6[15], para as bases de dados do RCBP-SP, do PRO-AIM e da APAC-SIA/SUS. As variáveis utilizadas no processo de pareamento foram o nome do paciente e a data ou ano de nascimento. Quanto à blocagem, as variáveis foram sexo, *soundex* do primeiro nome (PBLOCO), *soundex* do último nome (UBLOCO) e ano de nascimento. Como variáveis confirmatórias para o aceite de um par verdadeiro, quando disponível, foram usadas as variáveis: endereço, nome da mãe, causa básica de óbito e tipo de procedimento/tratamento. A combinação entre os campos de pareamento e de blocagem resultaram em 14 estratégias de *linkage*. Foram considerados pares verdadeiros, entre o *linkage* do RCBP-SP *versus* PRO-AIM, os escores de concordância entre 20,8 e 23,3, e para o *linkage* entre o RCBP-SP *versus* APAC-SIA/SUS, os escores entre 20,7 e 23,0. Os escores intermediários, entre a discordância total (-8,90 e -6,1, respectivamente, PRO-AIM e APAC-SAI/SUS) e os valores de concordância, foram revistos manualmente por uma equipe de três pesquisadores[15].

Para a base de dados da Fundação SEADE, todo o processo foi desenvolvido em *Visual Basic*, que conecta as bases hospedadas em *SQL Server*. Nesse processo, foram elaborados dois critérios para a seleção de pares verdadeiros: igualdade (preenchimento idêntico) e semelhança (concordância ≥ 80% — revisão visual)[16]. Na seleção automática os pares

que apresentaram igualdade total nas variáveis nome, data de nascimento, endereço e data do óbito, ou nas variáveis nome, data de nascimento e endereço, foram aceitos como pares verdadeiros.

A Figura 1 representa as etapas do *linkage* conforme a entrada das bases de dados. O primeiro *linkage* foi realizado entre o RCBP-SP e o PRO-AIM, após a retirada dos casos encontrados e classificados como óbito. Foram realizados concomitantemente o *linkage* entre o RCBP-SP e a Fundação SEADE e entre o RCBP-SP e a APAC-SIA/SUS.

As frequências absolutas e relativas foram calculadas para análise estatística, assim como os CBI e CBM por câncer antes e após o *linkage*. O denominador foi a população total do município de São Paulo para o período estudado, segundo sexo e faixa etária. A sobrevida global acumulada foi realizada pelo estimador do produto limite de *Kaplan-Meier* e a comparação entre antes e após utilizou o teste de *log rank*. O tempo de sobrevida foi calculado como a diferença entre a data da última informação ou data do óbito e a data de diagnóstico. Nesta análise, o tempo foi truncado em sete anos, sendo considerados como censura os pacientes que vieram a óbito após esse período. Para que houvesse um tempo mínimo de cinco anos de seguimento, apenas pacientes com diagnóstico de câncer entre 1997 e 2002 foram incluídos. Na análise por topografia foram incluídos os cânceres de maior incidência,



*Excluídos para o próximo processo de *linkage*; **dos pares verdadeiros identificados neste *linkage*, 1.720 pacientes vieram a óbito e foram identificados no *linkage* entre o Registro de Câncer de Base Populacional do Município de São Paulo (RCBP-SP) e a Fundação SEADE.

Figura 1. Fluxograma do processo de linkage probabilístico e determinístico entre as bases de dados.

a saber: estômago, cólon-reto, pulmão, mama, colo do útero e próstata. Foi assumido um nível descritivo de 5% para significância estatística. Os dados foram analisados no programa *Statistical Package for Social Science* (SPSS), versão 17.0.

Esta pesquisa foi aprovada pelo Comitê de Ética em Pesquisa da Faculdade de Saúde Pública da Universidade de São Paulo (n° 0086.0.207.000-08) e pela SMS-SP (n°0064.0.162.000-09).

RESULTADOS

Ao final de cada *linkage* (Figura 1), observa-se o total de pares relacionados, totalizando 122.139 registros. A Tabela 1 apresenta a completitude da base de dados do RCBP-SP antes e após o *linkage*. Na coluna "antes", estão apresentados os valores iniciais do RCBP-SP, e

Tabela 1. Número de casos com informações não ignoradas no Registro de Câncer de Base Populacional do Município de São Paulo e seus respectivos ganhos em frequência absoluta e relativa. Registro de Câncer de Base Populacional do Município de São Paulo, 1997 a 2005.

Linkage	Antes	Após	≠
	n	n	percentual
PRO-AIM			
Data de nascimento	220.176	224.719	2,1
Endereço	81.057	137.118	69,2
Nome da mãe	1.801	44.990	2.398,1
Data do óbito	103.910	120.895	16,3
APAC-SIA/SUS			
Data de nascimento	220.176	221.968	0,8
Endereço	81.057	88.988	9,8
Nome da mãe	1.801	14.477	703,8
Data da última informação	11.462	20.628	80,0
Fundação SEADE			
Data de nascimento	220.176	223.288	1,4
Endereço	81.057	101.209	24,9
Nome da mãe	1.801	22.264	1.136,2
Data do óbito	103.910	119.893	15,4
GERAL			
Data de nascimento	220.176	229.470	4,2
Endereço	81.057	163.310	101,5
Nome da mãe	1.801	76.332	4.138,3
Data do óbito	103.910	136.626	31,5
Data da última informação	11.462	20.628	80,0

Linkage: valores observados entre o *linkage* com o RCBP-SP; PRO-AIM: Programa de Aprimoramento de Mortalidade no Município de São; APAC-SIA/SUS: Autorização e Procedimentos de Alta Complexidade do Sistema de Informação Ambulatorial do Sistema Único de Saúde; Fundação SEADE: Fundação Sistema Estadual de Análise de Dados; GERAL: total de valores observados entre os *linkages* com o RCBP-SP.

na coluna "após" é apresentada a somatória entre quantidade de informação inicial com as informações completadas. Ao comparar os momentos antes e após no *linkage* do RCBP-SP com a base de dados do PRO-AIM, foi verificado um ganho de 69,2% para a variável endereço e 16,3% para a data do óbito.

Quanto ao *linkage* com a base de dados da Fundação SEADE, verifica-se ganhos de 24,9 e 15,4%, respectivamente, para as variáveis endereço e data de óbito. Em relação ao *linkage* com a base de dados da APAC-SIA/SUS, a variável data da última informação teve um ganho na completitude dos dados de 80,0%. No geral, ressalta-se que todos os relacionamentos foram importantes na estratégia para a completitude das informações da base de dados do RCBP-SP. Quanto ao seguimento, a data do óbito apresentou um ganho de 31,5% (Tabela 1).

A partir do *linkage* entre a base de dados do RCBP-SP *versus* PRO-AIM e *versus* Fundação SEADE, a variável causa do óbito foi agregada. Destaca-se que, dos casos relacionados, a causa básica de óbito, em sua maioria, foi decorrente ao câncer (86,4%). No *status* final dos pacientes registrados na base de dados do RCBP-SP, houve um ganho de 3,4% em casos novos, passando de 343.306 para 354.957 casos incidentes. Considerando as três fontes de informação ao final do processo, houve uma diminuição na perda de seguimento tanto para o *status* vivo (antes = 3,3% e após = 5,8%) quanto para o *status* óbito (antes = 30,3% e após = 38,5%). Entretanto, 55,7% permaneceram sem informação quanto ao *status* atual.

Destaca-se também que na Tabela 2 são apresentados os CBI e CBM, antes e após o *linkage*. Para o cálculo desses coeficientes, foram excluídos os casos de pele não melanoma (C44.0 a C44.9). Nesta análise, os numeradores foram o total de casos novos (antes = 272.644 e após = 284.280) e o total de óbitos (antes = 95.899 e após = 120.652), ocorridos entre os anos de 1997 e 2005. Ressalta-se que um único registro de cada paciente foi considerado no cálculo do CBM. O CBI, que era de 288,2 casos novos por 100.000 habitantes passou para 300,5 após o *linkage* entre as bases de dados, havendo um ganho de 4,3% no coeficiente. Analisando os CBM, verifica-se que, no geral, há uma diferença percentual de 25,8 % (Tabela 2).

Ao observar a sobrevida acumulada, havia uma subestimação nas probabilidades. Antes do *linkage* a probabilidade de estar vivo em sete anos era de 7,8%, passando para 13,0% (Tabela 3).

Tabela 2. Coeficientes brutos de incidência e mortalidade por câncer (por 100.000 habitantes) antes e após o *linkage*, excluídos os casos de câncer de pele não melanoma. Regitro de Câncer de Base Populacional do Município de São Paulo, 1997 a 2005.

Coeficientes	Antes		Após		≠
	n	coeficiente	n	coeficiente	percentual
CBI	272.644	288,2	284.280	300,5	4,3
CBM*	95.899	101,4	120.652	127,5	25,8

CBI: coeficiente bruto de incidência; CBM: coeficiente bruto de mortalidade; *óbitos encontrados nas bases de dados do Programa de Aprimoramento de Mortalidade no Município de São Paulo (PRO-AIM), na Fundação Sistema Estadual de Análise de Dados (Fundação SEADE) e mais os 56 casos da Autorização e Procedimentos de Alta Complexidade do Sistema de Informação Ambulatorial do Sistema Único de Saúde (APAC-SIA/SUS).

Da mesma forma, a subestimação foi observada para ambos os sexos. Entre os homens, após dois anos de seguimento, houve uma diferença percentual de 59,3%, na qual a probabilidade de estar vivo antes do *linkage* apresentava-se em 26,3% passando para 41,9% após o relacionamento ($p < 0,001$). O mesmo foi observado para todas as faixas etárias ($p < 0,001$) antes e após o *linkage*. Evidencia-se que a probabilidade de estar vivo ao sétimo ano passou de 2,3%, antes do *linkage* na faixa etária de 0 a 14 anos, para 17,6% após o *linkage* (Tabela 3). Para a faixa etária de 60 a 84 anos, a probabilidade de sobrevida antes do *linkage* era de 29,9%, enquanto que após o relacionamento esta probabilidade ficou em 45,8%.

Em relação às topografias (Tabela 4), verifica-se uma subestimação estatisticamente significativa nas probabilidades de sobrevida para todos os períodos analisados ($p < 0,001$). Destaca-se a probabilidade de sobrevida para o câncer de próstata que antes do *linkage* era de 5,2% e após o câncer passou para 17,3%.

Tabela 3. Sobrevida global acumulada, segundo sexo e faixa etária, antes e após o *linkage*, truncada em sete anos. Registro de Câncer de Base Populacional do Município de São Paulo, 1997 a 2005.

Variáveis	*Linkage*	n° de casos	n° de óbitos	Probabilidade de sobrevida global acumulada (% em anos)					Valor p (K-M)
				1°	2°	3°	5°	7°	
Geral	Antes	56.237	48.613	48,8	31,9	22,7	12,3	7,8	< 0,001
	Após	78.831	64.120	61,1	46,6	37,0	23,2	13,0	
Sexo									
Feminino	Antes	27.865	22.900	54,7	37,6	27,6	16,4	11,8	< 0,001
	Após	39.415	30.379	65,7	51,2	41,2	27,0	16,4	
Masculino	Antes	28.372	25.713	43,0	26,3	18,0	8,3	3,5	< 0,001
	Após	39.416	33.741	56,5	41,9	32,8	19,5	9,7	
Faixa Etária (anos)									
0 – 14	Antes	711	648	45,6	26,3	16,2	6,5	2,3	< 0,001
	Após	1.035	756	59,7	44,3	34,6	24,5	17,6	
15 – 29	Antes	1.408	1.147	52,3	34,8	24,6	14,7	9,9	< 0,001
	Após	1.940	1.403	63,1	48,7	39,2	27,5	18,8	
30 – 44	Antes	5.532	4.300	58,0	40,6	30,8	19,8	15,0	< 0,001
	Após	7.675	5.637	67,2	52,6	42,9	28,9	18,7	
45 – 59	Antes	14.503	12.374	50,9	33,1	24,1	13,9	9,5	< 0,001
	Após	19.548	15.591	61,9	46,5	37,0	24,0	14,2	
60 – 84	Antes	31.157	27.716	46,4	29,9	20,7	10,3	5,7	< 0,001
	Após	44.239	37.007	59,9	45,8	36,2	22,2	11,7	
≥ 85	Antes	2.526	2.386	38,5	22,1	14,7	5,2	1,9	< 0,001
	Após	4.021	3.699	54,9	39,7	30,7	15,0	6,1	

K-M: teste de Kaplan-Meier.

DISCUSSÃO

Nossos resultados indicaram que a aplicação de ambas as técnicas contribuíram para a melhora da qualidade da informação do RCBP-SP, tanto em termos da completitude de variáveis como na definição do *status* vital dos pacientes. O CBI e o CBM apresentaram um incremento, bem como as análises de sobrevida que trouxeram outro perfil do prognóstico.

Vale lembrar que a opção do tipo de *linkage* empregado considerou os objetivos do relacionamento e suas implicações epidemiológicas. Para a identificação dos casos novos, utilizou-se o *linkage* determinístico, uma técnica baseada na utilização de regras para a classificação dos *links* formados em pares verdadeiros. O desenvolvimento de regras por *experts* no tema e nas bases pode levar a resultados bastante acurados[17]. Quanto ao *linkage* probabilístico, pesa a crítica deste ser menos transparente, dado que o poder de discriminação dos pares é baseado em escores construídos segundo a capacidade discriminatória de cada campo, sendo vulnerável a erros falsos-positivos e negativos. No entanto, essa técnica torna-se uma solução viável quando existe comprometimento da qualidade dos registros a serem vinculados, podendo as variáveis apresentar erros e falhas na informação, ausência de um indicador unívoco ou poucas variáveis para a vinculação dos registros em bases de dados com grande número de casos[17-19].

Nesta pesquisa, os resultados foram analisados a partir da identificação dos pares verdadeiros, segundo a completitude e qualidade das bases de dados. Para os óbitos, o processo de *linkage* entre a base de dados do RCBP-SP *versus* PRO-AIM detectou que 81% já constavam

Tabela 4. Sobrevida global acumulada, segundo topografia, antes e após o *linkage,* truncada em sete anos. Registro de Câncer de Base Populacional do Município de São Paulo, 1997 a 2005.

Topografia	*Linkage*	n° de casos	n° de óbitos	Probabilidade de sobrevida global acumulada (% em anos)					Valor p (K-M)
				1°	2°	3°	5°	7°	
Estômago	Antes	4.990	4.877	25,7	11,2	6,0	1,9	0,5	< 0,001
	Após	6.029	5.693	36,8	23,0	16,6	9,0	3,9	
Cólon e Reto	Antes	5.643	5.373	46,9	26,9	14,6	4,1	0,9	< 0,001
	Após	7.318	6.630	56,7	39,3	26,6	13,3	6,2	
Pulmão	Antes	5.622	5.554	29,0	13,1	7,0	2,3	0,5	< 0,001
	Após	6.543	6.317	37,2	21,6	14,3	7,0	2,6	
Mama	Antes	7.414	5.188	74,8	59,0	46,5	30,4	23,5	< 0,001
	Após	10.963	7.106	82,0	69,6	58,6	41,2	27,8	
Colo do útero	Antes	4.477	3.758	55,8	34,4	24,3	14,5	12,1	< 0,001
	Após	6.242	4.858	66,7	48,6	37,8	24,8	16,3	
Próstata	Antes	3.485	3.005	63,1	45,1	32,8	14,7	5,2	< 0,001
	Após	6.326	4.850	77,5	64,1	52,7	32,8	17,3	

K-M: teste de Kaplan-Meier.

com a data do óbito. Destaca-se que o RCBP-SP realizava anualmente o processo de *linkage* probabilístico com o PRO-AIM. Contudo, esse relacionamento era feito somente com os casos que apresentavam óbitos por câncer com a finalidade de fechar o seguimento e resgatar possíveis casos incidentes não detectados pelo RCBP-SP. Vale ressaltar que esse processo era realizado de tal forma em respeito às recomendações da IARC, as quais consideram que pacientes não identificados no banco de dados de óbitos estão vivos[2]. Além das informações complementadas, foram agregados os dados nome da mãe, endereço e causa básica do óbito. O ganho substancial para a variável nome da mãe é fundamental para a melhora na qualidade da base do RCBP-SP, pois é a principal forma para a diferenciação de homônimos.

Quanto à completitude da variável endereço, feita via registros de óbito e registros da APAC-SIA/SUS, algumas considerações fazem-se importantes, pois o endereço do óbito ou o endereço atual podem não ser o mesmo do momento do diagnóstico, implicando na distribuição espacial errônea dos casos de câncer. Nesse contexto, essa informação não é indicada para realizar análises de georreferenciamento. Por outro lado, essa informação pode ser um indicador de migração interna para os casos ocorridos nas áreas próximas de unidades de tratamento (óbitos por ocorrência) e, para os pacientes vivos, mostra-se relevante caso haja a necessidade de busca ativa.

A causa básica do óbito, sendo ela câncer ou não câncer, entrou como uma nova variável e, dos óbitos encontrados, 87,6% tiveram como causa básica o câncer. Todavia, essa proporção de causas de óbito por câncer deve ser analisada com cautela por dois motivos. Primeiramente, pode haver viés de informação atrelado ao preenchimento da Declaração de Óbito (DO). Acredita-se que há uma tendência, em óbitos de pacientes com câncer, do médico preencher o atestado, associando a causa da morte a causas existentes que não foram as responsáveis pelo desfecho[20]. Em segundo lugar, no processo de *linkage*, a causa básica do óbito foi uma das variáveis utilizadas como parâmetro de confirmação na identificação de um par verdadeiro. Porém, não isoladamente para identificar um registro vinculado. Para ser um par verdadeiro, eram necessárias ao menos mais duas condições das variáveis de relacionamento e de confirmação[14].

Uma das principais limitações deste estudo foram os erros de preenchimento da base de dados da APAC-SIA/SUS. Acredita-se que o tipo de registro utilizado para faturamento pode ter influenciado na qualidade, pois a partir do primeiro registro os demais, para aquele mesmo paciente, apresentavam falhas na informação como nomes abreviados, endereço incompleto, CPF zerado ou em branco, nome da mãe em branco ou com as palavras "O MESMO", fazendo referência há um registro prévio.

Para avaliar o impacto que o processo de *linkage* ocasionou nas estatísticas do RCBP-SP, estas foram analisadas antes e após o processo. O CBI estava subestimado em 4,3%, assim, outros estudos que fizeram uso da técnica de *linkage* para a completitude da informação mostraram as diferenças no número de eventos registrados[21,22]. Em relação ao CBM, houve um ganho de 25,8% após o processo de *linkage*. No estudo de Pereira et al.[23] realizado no Rio de Janeiro entre 1999 e 2001, foi verificado que havia uma subestimação de 20,0% no coeficiente de mortalidade neonatal. Da mesma forma, a pesquisa de Rafael et al.[24], que fez

uso das bases de dados do Sistema de Informação sobre Mortalidade (SIM) e do Sistema de Informações Hospitalares (SIH) no Estado do Maranhão, apresentou uma subestimação nos coeficientes de mortalidade neonatal e infantil de 24,9 e 19,8%, respectivamente.

Quanto à análise de sobrevida global dos pacientes do RCBP-SP, após o *linkage* as probabilidades de sobrevida aumentaram ao longo dos anos analisados para ambos os sexos, faixas etárias e para as topografias analisadas. Pode-se destacar, neste caso, a importância de relacionar não só bases de dados de óbitos, e sim de pacientes vivos. Pois, a partir dos pacientes vivos identificados via APAC-SIA/SUS houve um ganho na qualidade da informação, o que refletiu nas probabilidades de sobrevida ao longo dos anos. No entanto, é necessário destacar que mesmo depois da inclusão das informações de *status* do paciente provenientes das bases PRO-AIM, Fundação SEADE e APAC-SIA/SUS, 55,7% dos casos permaneceram sem informação de seguimento — data da última informação ou data do óbito. Uma vez que os óbitos são registrados de forma compulsória e devem integrar as bases correspondentes, é possível que o processo de *linkage* realizado por este estudo tenha identificado parte importante dos óbitos. Por outro lado, os dados contidos na base da APAC-SIA/SUS restringem-se a procedimentos de alto custo e complexidade realizados nos pacientes em seguimento pelo Sistema Único de Saúde (SUS); portanto, pacientes vivos que não estão em tratamento não integram a base e não têm a oportunidade de serem capturados.

Todavia, a SMS-SP relata que os procedimentos de alta complexidade registrados nesse sistema, devido ao seu custo elevado e à limitada cobertura pelos planos privados de saúde, representam 90% do total de procedimentos realizados na região Sudeste. Para o município de São Paulo, verifica-se que a taxa de cobertura pelos planos de saúde é de 59,8%; no entanto a pesquisa realizada pelo Instituto de Estudos de Saúde Suplementar (IESS) identificou que 26% das pessoas que possuem assistência médica privada também utilizam o SUS[25,26]. Assim, o relacionamento periódico do RCBP-SP com outras bases de dados que possam contribuir com informações de seguimento, principalmente de pacientes vivos, deve ser parte da estratégia para o seu aprimoramento.

Ao observar a proporção da perda de seguimento, outro ponto importante é que esta pode variar conforme a topografia analisada. Nesta pesquisa, foram selecionadas as topografias de maior incidência e mortalidade, porém vale ressaltar que quanto maior a agressividade do câncer maior a probabilidade de observar o paciente entre o tratamento e o óbito, dado ao menor tempo de acompanhamento. Contudo, pacientes com câncer de tireóide e mama, por exemplo, apresentam sobrevidas elevadas, podendo interromper o acompanhamento ou migrar para outras regiões.

CONCLUSÃO

Pelo exposto, conclui-se que todas as variáveis do RCBP-SP apresentaram maior completitude do dado, assim como a melhora na qualidade da informação por meio dos CBI, CBM e sobrevida, mostrando a efetividade das técnicas de *linkages* probabilístico e determinístico.

Enfatiza-se também que os resultados encontrados nesta pesquisa refletem a capacidade das bases de dados, quando trabalhadas de maneira conjunta, de fornecer subsídios para diversos tipos de estudos e informações para o planejamento de ações políticas e estratégicas.

Destaca-se ainda que diferentemente de outros países, pouco se trabalha com análises de sobrevida utilizando bases de dados populacionais[12,27], principalmente em RCBP devido à ausência de seguimento. Nesse contexto, a vinculação entre bases de dados de perfil administrativo como APAC-SIA/SUS, Sistema de Informação em Saúde para a Atenção Básica (SISAB), Boletim de Produção Ambulatorial Individualizado (BPA-I), Sistema de Informação de Câncer (SISCAN), Sistema de Informações Hospitalares Descentralizados (SIHD), e Tribunal Regional Eleitoral (TRE) com bases que trazem informações epidemiológicas como óbitos, nascimentos e notificações de doenças, enriquecem e ampliam as possibilidades de análises, além de gerar informações de melhor qualidade e serem menos onerosas, buscando sempre o aprimoramento dos indicadores de saúde.

REFERÊNCIAS

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.

Recebido em: 02/09/2015
Versão final apresentada em: 11/02/2016
Aprovado em: 30/05/2016