DOI: 10.1590/1980-549720180015

ORIGINAL ARTICLE / *ARTIGO ORIGINAL*

Strategies used to link Health Information Systems for the follow-up of women with abnormal mammograms in the Brazilian Public Health System

Estratégias usadas no relacionamento entre Sistemas de Informações em Saúde para seguimento das mulheres com mamografias suspeitas no Sistema Único de Saúde

Jeane Glaucia Tomazelli[I], Vania Reis Girianelli[II], Gulnar Azevedo e Silva[III]

ABSTRACT: ***Introduction:*** Health Information Systems are potential instruments to analyze health situation; however, the non-compulsory filling of a single common field makes it difficult to link systems' data. This study aimed to describe and evaluate the adequacy of the strategies used to perform data linkage between databases from the Brazilian Public Health System (SUS) as to records for breast cancer control. ***Methods:*** The Breast Cancer Control Information Systems (SISMAMA), the Outpatient Information System (SIA, through Individualized Outpatient Service Production — BPA-I — and High-Complexity Outpatient Procedures Authorization Forms — APAC), the Hospital Information System (SIH), and the Mortality Information System (SIM) were linked probabilistically. The baseline was constructed by records with "suspected" and "highly suspected malignancy" from the second half of 2010. The linkage strategy included 15 steps. Registries with the national health service user card (CNS) or social security number (SSN) were used to estimate the sensitivity of the strategy, considering matches between records identified in the initial steps as gold standard, when these fields were used as key for blocking. ***Results:*** Using CNS and the SSN as a linkage strategy allowed to identify the high proportion of true matches across databases in which these variables were inputted: 47.3% in follow-up mammography records, 41.4% in SIH, and 45.5% in APAC. The sensitivity of the linkage strategy was 100%. ***Conclusion:*** The study showed that the strategies were satisfactory and the use of CNS and SSN allowed many matches, even without critical proceedings and with the possibility of linkage between databases based on information from only a few identification fields.

Keywords: Health Information Systems. System integration. Breast neoplasms.

[I]Instituto Nacional de Câncer José Alencar Gomes da Silva – Rio de Janeiro (RJ), Brazil.
[II]Escola Nacional de Saúde Pública Sérgio Arouca, Fundação Oswaldo Cruz – Rio de Janeiro (RJ). Brazil.
[III]Institute for Social Medicine, Universidade do Estado do Rio de Janeiro – Rio de Janeiro (RJ), Brazil.
Corresponding author: Jeane Glaucia Tomazelli. Divisão de Detecção Precoce e Apoio à Organização de Rede – Instituto Nacional de Câncer. Rua Marquês de Pombal, 125, Centro, CEP: 22230-240, Rio de Janeiro, RJ, Brasil. E-mail: jtomazelli@inca.gov.br

Conflict of interests: nothing to declare – **Financial support:** none.

RESUMO: ***Introdução:*** Sistemas de Informação em Saúde (SIS) são instrumentos potenciais para análise da situação de saúde, mas a não obrigatoriedade de preenchimento de um campo comum único dificulta sua integração. O objetivo deste estudo foi descrever as estratégias utilizadas para relacionar bases de dados do Sistema Único de Saúde (SUS) que contenham registros para o controle do câncer de mama e avaliar a adequação da estratégia empregada. ***Metodologia:*** Foram relacionados probabilisticamente o Sistema de Informação do Controle do Câncer de Mama (SISMAMA), o Sistema de Informação Ambulatorial (SIA, por meio do Boletim de Produção Ambulatorial Individualizado — BPA-I — e da Autorização de Procedimentos Ambulatoriais de Alta Complexidade — APAC), o Sistema de Informação Hospitalar (SIH) e o Sistema de Informação sobre Mortalidade (SIM). A base de referência foram registros de mamografia suspeita e altamente suspeita de malignidade do segundo semestre de 2010. A estratégia de relacionamento incluiu 15 passos. Os registros com Cartão Nacional de Saúde (CNS) ou Cadastro de Pessoa Física (CPF) foram utilizados para estimar a sensibilidade da estratégia, tendo como padrão-ouro os pares de registros identificados nos passos iniciais, que usaram esses campos como chave de blocagem. ***Resultados:*** A utilização do CNS e do CPF como estratégia de relacionamento permitiu identificar elevada proporção de pares verdadeiros nas bases nas quais existiam essas variáveis: 47,3% nas mamografias de seguimento, 41,4% no SIH e 45,5% na APAC. A sensibilidade da estratégia utilizada foi de 100%. ***Conclusão:*** O estudo mostrou que as estratégias utilizadas foram satisfatórias e que a utilização do CNS e do CPF permitiu a identificação de muitos pares, mesmo com a ausência de crítica destes e a possibilidade de realizar o relacionamento entre bancos com poucos campos de identificação.

Palavras-chave: Sistemas de Informação em Saúde. Integração de sistemas. Neoplasias de mama.

INTRODUCTION

Health Information Systems (SIS) are potential instruments to analyze the health situation, planning, programming, and evaluation, as they keep record of epidemiological, assistance and vital statistic information. Implemented at different moments in Brazil since the 1970s, the SIS were developed to serve different purposes. Their characteristics are diverse, from coverage, which can be universal, such as the Mortality Information System (SIM), to services belonging to the Brazilian Public Health System (SUS) such as: Outpatient Information System (SIA) and Hospital Information System (SIH). As to SIS, the filling of some fields is not uniform, just like the pointing of information that enables individuals' socioeconomic characterization[1-4].

By linking the SIS, one can obtain longitudinal follow-up of individuals in the SUS care network. This linkage can not be done directly because there is not a common field in SIS that identifies the individuals and whose filling is not compulsory. Although this sole identification in SUS has been gradually discussed and implemented through the national health service user card (CNS)[5-7], in practice, the non-adoption of this identifier across all SIS poses difficulties to health care evaluation. Among the nationally implemented SIS, CNS is mandatory only for High-Complexity Procedures Authorization (APAC). In 2011, duplicate records, that is, individuals with more than one CNS, were estimated to be higher than 30%[8], which added more obstacles to the evaluation process.

Other difficulties were: the process of decentralized CNS registration and problems of synchronization with the national card user database, making it not representative of all individuals registered in the system[4].

Methodologies that allow to identify a user at different moments of their care are a resource that has been used to overcome this difficulty. The development of computational routines for SIS data linkage aiming to keep track of SUS users' trajectory and to evaluate care is central to support planning[9-12]. Gomes Júnior et al.[13] proposed data linkage routines based on APAC oncology module information. Adaptations[14] and proposals for algorithm automation[15] are currently being developed.

In Brazil, several studies have applied the probabilistic linkage technique using the software RecLink®[16-23], and some report satisfactory sensitivity and specificity[17-19]. It is important, however, to consider that the accuracy of data linkage is related to the number of identifiers to be compared, as well as to the quality of their completion[17]. Overall, false matches (false positives) tend to occur when there are few fields for comparison, incompleteness of fields used for comparison, and homonyms. On the other hand, unidentified matches (false negatives) are more related to typing errors and information filled incorrectly[24].

There are two official information systems within SUS that keep record of early cancer screening tests: Cervical Cancer Control Information System (SISCOLO) and Breast Cancer Control Information System (SISMAMA)[25]. Most studies that applied data linkage technique used the "cytology" and "anatomopathological" modules of the SISCOLO database[20,23,26,27]. Some specific studies using both SISCOLO and SISMAMA data have been conducted using the RecLink®[20] software along with programming resources of other statistical software packages[26,27]. Only one study performed data linkage between SISCOLO and SIM[23]. There are no Brazilian studies available that link data from different SIS in order to keep track of women screened for cervical and breast cancers so one can to evaluate actions for the control of these diseases.

This study aimed to describe and evaluate the adequacy of the strategies used to perform probabilistic data linkage between databases from the SIS as to records for breast cancer control.

METHODS

This is a descriptive study on the methodology used to link data from SIS databases about breast cancer screening in the city of Rio de Janeiro. The period of study encompassed July 2010 through December 2012, starting one year after the implementation of SISMAMA[28] and ending before its replacement by the Cancer Information System (SISCAN)[29].

SIS data relating to the follow-up of women with suspected malignant mammography were used. SISMAMA (mammography and breast anatomopathological exam), SIA (diagnostic investigation and treatment procedures) and SIH (surgical treatments and hospitalizations for external radiotherapy) information was granted by the Municipal Department of Health. SISMAMA data were complemented with information from the national database provided by the National Cancer Institute José Alencar Gomes da Silva (INCA), after a preliminary analysis identify the absence of data regarding the initial months of the study. SIM data were provided by the State Bureau of Vital Records and Health Statistics (SES-RJ/SVS/CGVS/ADVITAIS) and refer to all deaths in the State along this period.

Monthly files were separately generated from municipal SISMAMA database for mammography exams and for anatomopathological examinations in the period of interest available (May 2011 to December 2012). From the national base of the SISMAMA, only the exams of residents of the city of Rio de Janeiro were used, with monthly mammography files from July 2010 to December 2012 and annual anatomicopathological exams. The anatomopathological examinations being available from January 2010 on allowed the exclusion of prevalent cases[30]. A correspondence between the fields of both SISMAMA sources (municipal and national) was then made. Afterward, the bases were linked by year and duplicates were removed.

The reference database was taken from SISMAMA's "mammography" module for the second semester of 2010. Records of women residing in the city of Rio de Janeiro with mammography suspicious (BI-RADS® 4) or highly suspicions (BI-RADS® 5) of malignancy[31] in one breast. Files lacking information on patient's mother's name and address were excluded. As the size of fields differs between the databases, the fields "patient's name", with 30 and 50 characters, and "mother's name", with 30 and 45 characters were created. After selection of records for reference, the others composed a file named "2010 follow-up mammograms", which was used as a comparison database in data linkage.

The Individualized Outpatient Service Production (BPA-I) contains diagnostic procedures for breast cancer (biopsy or surgical specimen anatomopathological examination). Data were extracted from the fields "procedure performed", "anatomopathological examination of the breast" and "frozen/paraffin-embedded anatomopathological exam, except for cervix and breast", by both surgical specimen or biopsy. The frozen/paraffin-embedded procedure was included because the system does not have critical proceedings for International Disease Classification (ICD). APAC keeps record of authorized chemotherapy, hormone therapy and radiation therapy procedures. From the "main procedure" field, we extracted records that admit ICD for breast cancer and selected non-continuity records.

Surgical records were selected from SIH, based on the field "procedure performed": simple and radical mastectomy with axillary lymphadenectomy, segmentectomy/quadrantectomy/sectorectomy, sectorectomy/quadrantectomy with lymph node dissection, oncological or non-oncological; and hospitalization for external radiation therapy. The records dated from "August 2011" were not available and could not be retrieved.

For SIA and SIH, ICD-10 registries selected addressed breast cancer (C50), in-situ breast cancer (D05), benign breast disease (D24), disease of uncertain behavior (D48.6) and genitourinary tract disease (N60-N64), as well unfulfilled fields. From the annual files of SIM, female and non-fetal death reports were selected, except when lacking the woman's name and address.

A registration field was inserted in each database, with a sequence of letters followed by the year and the numerical order of the record on the base, identifying the examination. When searching all databases, the accent marks and cedillas were removed. Female records were selected and a manual review of male records was performed to identify possible coding errors. Changes were made to records in which the patient's name was not gender-dubious.

The annual databases of each SIS, except for the mammography module of SISMAMA 2010, were unified and residual duplicates were removed. Blank and improperly filled fields (i.e. repeated numbers, less/more characters than established) of variables social security number (SSN) and CNS were filled with a code created to reference the base, thus avoiding matches with blank fields upon linkage and optimizing its processing.

The analyses were performed on the program R version 3.1.1[31].

LINKAGE BETWEEN DATABASES

The program RecLink® version 3.1[32] was used for data linkage, following the steps of standardization, blocking and matching. For matching, the parameters proposed in the RecLink®[32] Manual were used, along with information from the fields "name", "mother's name" and "date of birth" (DB). "CNS" and "SSN" fields were used as blocking key for steps 1 and 2, respectively. Step 3 included the fields: DB, *soundex* codes for patient's (FN and LN) and mother's (FM and LM) first and last names. In the remaining 12 steps, the strategies became progressively less restricted. SIA and SIM databases do not have the field "SSN".

The reference baseline was initially linked to mammograms from 2010 to identify women who repeated the examination within the semester. The first exam remained in reference database and the others were included in a file called "2010 follow-up mammograms". Then, the reference base was linked to other bases, and matches with score above zero were assessed. At each step, only records not classified as true matches were maintained for comparison.

The classification of matches as true, by manual review, abided by the following criteria: name, DB and mother's name. The records in which two of these fields were unfulfilled were not considered for peer evaluation, except when the patient's name and address (street, number and neighborhood) were identical. Matches were: when the mother's name was totally different/absent, but the fields "patient's name", "DB" and "address" were the same/similar; when the mother's name and/or address were absent/different, but the name was considered rare (a foreign name, for example) and the DBs were the same. Rare first and last names or abbreviation/absence of middle or last name were defined as similar.

Since pathological records of breast in SISMAMA must correspond to those reported in the BPA-I[25], a linkage was made between both bases to identify exams that were not in SISMAMA. As BPA-I has few fields and does not have a field for mother's name, the strategy of using the national registration of health establishments (CNES), "appointment date/ outcome" and "ICD-10 in exam" in association with other fields available in this database (CNS, FN, LN and DB) was adopted. ICD-10 was used to match names and DB. After this linkage, the reference base was linked to BPA-I records that did not match SISMAMA's anatomopathological exams, using name rarity and the rule that the anatomopathological examinations' date should be after the mammogram as criteria.

The proportion of incompleteness of CNS and SSN fields in each SIS was calculated, along with respective percentage variation (PV) between 2010 and 2012. The number of matches between the anatomopathological exams in SISMAMA and BPA-I was then presented, and the percentage of types of procedures that did not match was calculated. The number of matches between the reference base and the other bases was presented, as well as the number and percentage of matches classified as true in each step. The sensitivity of probabilistic relationship strategy was calculated based only on CNS and/or SSN found in databases, having matches identified in steps 1 and 2 as the gold standard.

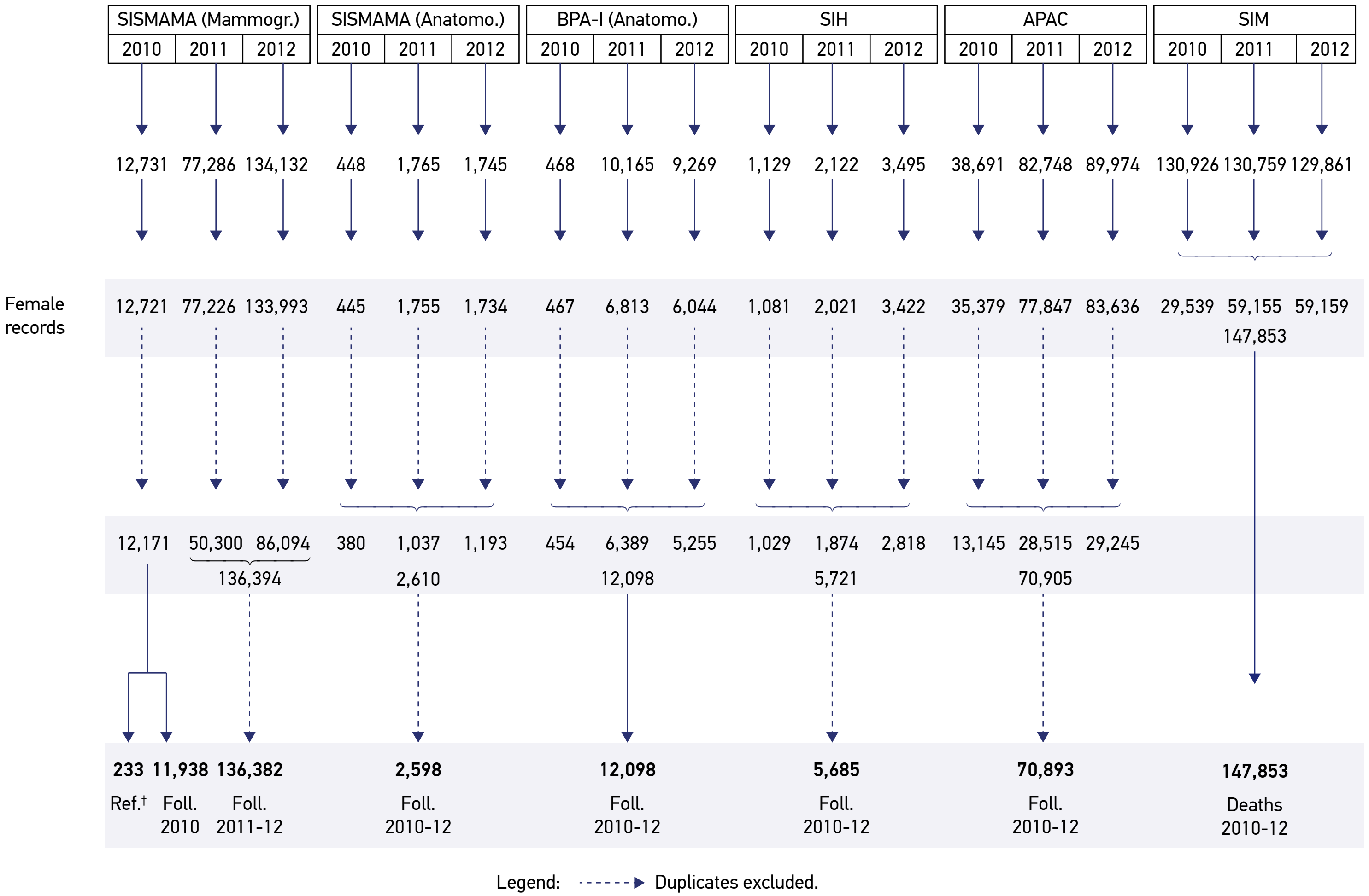
The study was approved by the Research Ethics Committees of Institute of Social Medicine *Universidade do Estado do Rio de Janeiro* (Opinion no. 1,105,945), the city of Rio de Janeiro (Opinion no. 1,162,544) and INCA (Opinion no. 1,139,738).

RESULTS

SISMAMA had 5,565 more mammograms and 146 more anatomopathological exams in the city of Rio de Janeiro than nationwide. Reviewing the "gender" field allowed the identification of six mammography exams incorrectly coded (SISMAMA), five anatomopathological (SISMAMA), 59 in BPA-I, 40 in SIH, 576 in APAC and 4 in SIM.

The reference was made up of 233 women (Figure 1). Follow-up mammograms, after the cleansing process, totaled 11,938 in 2010 and 136,382 from 2011 to 2012. The base with the smallest number of records was anatomopathological exams in SISMAMA (n = 2,598), followed by SIH (n = 5,685), anatomopathological findings in BPA-I (n = 12,098), APAC (n = 70,893) and SIM (n = 147,853).

The CNS field had a reduction in the proportion of incompleteness across all systems, being more significant in SIH: from 30.6% in the second half of 2010 to 3.6% in 2012 (percentage variation, PV = 88.2%). However, in 2012, this percentage was still high in SISMAMA (mammography: 70.6%, anatomopathological exam: 88.5%), SIM (98.5%) and BPA-I (69.7%). In APAC, the filling of this field is mandatory. The SSN field had a drop of proportion of incompleteness in SIH and among anatomopathological records of SISMAMA (PV = 34.6 and 6.6%, respectively), but it was still elevated in the last year



*Anatomopathological exams registered in SISMAMA from January to December 2010.

†Reference: BI-RADS® 4 or 5 screening mammography records.

Figure 1. Databases organized by source according to year and filters used, city of Rio de Janeiro, July 2010 * to December 2012.

of study (34.6 and 93.4%, respectively). There was, however, an increase in unfulfilling of this field in mammograms (PV = 17.3%). The other systems do not encompass this information (Table 1).

Upon data linkage between BPA-I and SISMAMA, 99.8% (2,276/2,280) of matches were classified as true. Among unidentified records in BPA-I, 546 were breast-specific procedures (biopsy and surgical specimen), corresponding to a loss of 19.3% in SISMAMA, as the procedure for "surgical specimen or biopsy except for cervix and breast" is not expected to be recorded in the system (Table 2).

Table 3 shows matches classified as true in linkage process. Among mammograms from the second half of 2010, 9 follow-up and mammograms were identified, as well as 220 screening mammograms for subsequent years. As to anatomopathological exams, 30 were found in SISMAMA and 7 in BPA-I. As for treatment, 70 surgeries were recorded in SIH and 455 in APAC. In SIM, 20 records were found.

In reference base, 33.0% of registries had the CNS field filled in, while 44.6% had SSN information. Using these fields as blocking key allowed the classification of matches as true in: 47.3% follow-up mammogram matches (2011-2012), 41.4% in SIH and 45.5% in APAC. True matches were not found using these blocking keys for follow-up mammograms in the second half of 2010 and for anatomopathological exams in SISMAMA. Steps 3 and 4 also presented a high proportion of true matches: data linkage with follow-up mammograms and pathological exams in SISMAMA dated from the second half of 2010, in which no matches were formed in previous steps, identified

Table 1. Percentage of unfulfilled fields for National Health Card and social security number by reference[a] and year, and percentage variation between 2010 and 2012, city of Rio de Janeiro.

Reference	2010			2011			2012			PV 2010–2012	
	n	% not informed		n	% not informed		n	% not informed			
		CNS	SSN		CNS	SSN		CNS	SSN	CNS	SSN
Mammography[b]	12,171	82.6	80.7	50,300	80.6	88.7	86,094	70.6	94.7	14.5	-17.3
Anatomopathological exam[b]	380	99.7	100.0	1,037	100.0	92.4	1,193	88.5	93.4	11.2	6.6
BPA-I (SIA)	454	100.0	-	6,389	85.5	-	5,255	69.7	-	30.3	-
SIH/SUS	1,029	30.6	52.9	1,874	34.9	49.8	2,818	3.6	34.6	88.2	34.6
APAC (SIA)	13,145	0.0	-	28,515	0.0	-	29,245	0.0	-	0.0	-
SIM	29,539	99.5	-	59,155	99.3	-	59,159	98.5	-	1.0	-

[a]After exclusion of duplicates of the year; [b]exams from the Breast Cancer Control Information System (SISMAMA); BPA-I: Individualized Outpatient Service Production Bulletin; SIA: Outpatient Information System; SIH: Hospital Information System; SUS: Brazilian Public Health System; APAC: High-Complexity Procedures Authorization; SIM: Mortality Information System; PV: percentage variation.

88.9 and 86.7% of true matches, respectively. The proportion of matches classified as true in steps 1, 2 and 3 was 91.4% for follow-up mammograms (2011-2012) and 92.3% in APAC records.

All matches with CNS and SSN information (steps 1 and 2, respectively) were detected in the complementary analysis, when probabilistic linkage strategies were used (sensitivity = 100%). Table 4 shows the number and percentage of matches in each step of this analysis. Step 3, whose blocking strategy was more restrictive, brought up more than 90% of true matches across all databases.

DISCUSSION

The proportion of mammograms with alterations (1.9%, 233/12,721) found in the reference was considered high, because, the percentage across Brazil, in the same year, was 1.4%, which is close to the results of the evaluation on SISMAMA implementation by Passman et al.[33].

The difference in SISMAMA for the city of Rio de Janeiro, despite the fact that data were available only as of May 2011, results from all examinations being performed in the city, while the national base has the records of patients who are residents in the municipality only. In addition, one must also consider the possibility of loss in routing flow between the municipal, state and national levels of databases[25].

The absence of records from August 2011 in SIH and the lack of records in SISMAMA were a limitation of this study, since some histopathological investigations may not have been included.

The identification of records of female patients coded as males indicates that restricting the field "gender" may hinder true matches in linkage process. An alternative to manual

Table 2. Anatomopathological records from the Individualized Outpatient Service Production Bulletin before and after probabilistic linkage, from July 2010 to December 2012, in the city of Rio de Janeiro.

Anatomopathological exam	Before linkage		After linkage			
			Identified		Not identified	
	n	%	n	%	n	%
Surgical specimen (breast)	873	7.2	723	82.8	150	17.2
Biopsy (breast)	1,949	16.1	1,553	79.7	396	20.3
Surgical specimen or biopsy (except breast or cervix)	9,276	76.7	0	0	9,276	100.0
Total	12,098	100.0	2,276	18.8	9,822	81.2

Table 3. Matches (n1), matches classified as true (n2) and percentage by data linkage step and database.

Steps and blocking keys	Mammography 2010			Mammography 2011-12			Anatomop. 2010-12			BPA-I 2010-12			SIH 2010-12			APAC 2010-12			SIM 2010-12		
	n1	n2	%	n1	n2	%		n2	%	n1	n2	%	n1	n2	%	n1	n2	%	n1	n2	%
CNS				83	83	37.7							24	24	34.3	208	208	45.7			
SSN				21	21	9.5							5	5	7.1						
FN + LN + DB + FM + LM	8	8	88.9	97	97	44.1	22	22	73.3							215	215	47.3			
FN + LN + DB				9	8	3.6	4	4	13.3	5	5	71.4	36	36	51.4	20	20	4.4	19	19	95.0
FN + DB + FM + LM	1	1	11.1	2	2	0.9	1	1	3.3							8	8	1.8			
FN + DB + LM																					
FN + DB + FM																					
FN + DB				1						11	0					1	1	0.2	1	1	5
FN + LN + PM + UM				7	6	2.7	2	2	6.7				2	2	2.9	2	2	0.4			
FN + LN + PM				5									1	0	0	3	1	0.2	1		
FN + LN				11	2	0.9				37	1	14.3	3	1	1.4	2			15		
FN + PM + LM				5															2		
FN + LM				6															10		
LN + DB				2	1	0.5	1	1	3.3	2			2	2	2.9				1		
DB				5						101	1	14.3							9		
Total	9	9	100.0	254	220	100.0	30	30	100.0	156	7	100.0	73	70	100.0	459	455	100.0	58	20	100.0

BPA-I: Individualized Outpatient Service Production Bulletin; SIH: Hospital Information System; APAC: High-Complexity Procedures Authorization; SIM: Mortality Information System; CNS: national health service user card; SSN: Social Security Number; FN: *soundex* code of female patient's first name; LN: *soundex* code of female patient's last name; DB: date of birth; FM: *soundex* code of mother's first name; LM: *soundex* code of mother's last name.

review would be to increase the number of steps in linkage process, including this field in some steps along with others in the blocking.

Incorrect fulfilling of SSN (more or less than 11 digits) and digit repetitions were identified, given the absence of a critical parameter in the systems. Problems with SSN were identified in another data linkage study that used national registries in APAC and SIH[12].

The four initial steps allowed us to identify most of the true matches. No anatomopathological exam, however, was identified in the two initial steps, due to the incompleteness of these fields in the histopathological database of SISMAMA and BPA-I.

The loss of anatomopathological exams recorded in SISMAMA was considered low (19.3%) when compared to BPA-I; and only matches addressing non-breast-specific surgical procedures or biopsies were found, signaling adequacy of the linkage strategy.

Table 4. Number and percentage of matches identified in probabilistic linkage from databases with registration of national health user card and/or social security number, according to step and database.

Step	Blocking keys	Database					
		SISMAMA 2010-2012		SIH 2010-2012		APAC 2010-2012	
		n	%	n	%	n	%
3	FN + LN + DB + FM + LM	98	94.2	28	96.6	198	95.2
4	FN + LN + DB	1	1.0	0	0.0	0	0.0
5	FN + DB + FM + LM	1	1.0	0	0.0	5	2.4
6	FN + DB + LM	0	0.0	0	0.0	0	0.0
7	FN + DB + PM	0	0.0	0	0.0	0	0.0
8	FN + DB	0	0.0	0	0.0	0	0.0
9	FN + LN + FM + LM	4	3.8	1	3.4	5	2.4
10	FN + LN + FM	0	0.0	0	0.0	0	0.0
11	FN + LN	0	0.0	0	0.0	0	0.0
12	FN + FM + LM	0	0.0	0	0.0	0	0.0
13	FN + LM	0	0.0	0	0.0	0	0.0
14	LN + DB	0	0.0	0	0.0	0	0.0
15	DB	0	0.0	0	0.0	0	0.0
Total		104	100.0	29	100.0	208	100.0

SISMAMA: Breast Cancer Control Information System; SIH: Hospital Information System; APAC: High-Complexity Procedures Authorization; FN: *soundex* code of female patient's first name; LN: *soundex* code of female patient's last name; DB: date of birth; FM: *soundex* code of mother's first name; LM: *soundex* code of mother's last name.

The linkage strategy was effective, as sensitivity was 100% in the analysis restricted to records that had CNS and social security information. Additionally, this analysis can contribute to the selection and/or prioritization of blocking strategies to be further used. The blocking key with FN, LM, FM and LM (step 9), for example, resulted in matches in most databases despite previous steps (steps 6 through 8) being unable to identify any.

This study was able to show the importance of SIS to evaluate breast cancer control actions. Each information system from SUS registers a stage of health care, and linking data from all of them allows to keep track of the follow-up of women with altered mammography while a single identifier is not made available for all SIS across the country.

CONCLUSION

The study showed that CNS and SSN allowed many matches, even without critical proceedings and being used for data linkage between databases with few identification fields and several blocking keys.

The continuous and progressive qualification of SIS is fundamental for the evaluation of health actions and programs. In addition to a mandatory single identifier, the standardization of minimum fields for qualification of an individual's records, such as "mother's name" in BPA-I, would increase the reliability of results. Using BPA-I as a source of information allowed us to find some records, despite the small number of identification fields. The inclusion of a critical system in fields not encompassed by rules, such as SSN and CNS, not allowing the registration of any more digits than established, would optimize the use of linkage tools. This is a very important step for epidemiological surveillance, taken by information centers of the Municipal and State Departments of Health responsible for the databases. Implementing this strategy could contribute to the continuous improvement in systems' quality and in the evaluation of health care programs.

REFERENCES

1. Brasil. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Regulação, Avaliação e Controle. Coordenação Geral de Sistemas de Informação. Manual técnico operacional SIA/SUS. Sistema de Informações Ambulatoriais. Brasília: Ministério da Saúde; 2009. 69 p.
2. Brasil. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Regulação, Avaliação e Controle. Coordenação Geral de Sistemas de Informação. Manual técnico operacional do Sistema de Informações Hospitalares. Brasília: Ministério da Saúde; 2010. 119 p.
3. Brasil. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Regulação, Avaliação e Controle. Coordenação Geral de Sistemas de Informação. APAC. Autorização de Procedimento Ambulatorial. Manual de operação do sistema. Versão 1.1. Brasília: Ministério da Saúde; 2013.
4. Brasil. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Regulação, Avaliação e Controle. Sistemas de Informação da Atenção à Saúde. Brasília: Ministério da Saúde; 2015. 166 p.
5. Brasil. Ministério da Saúde. Portaria GM nº 940, de 28 de abril de 2011. Regulamenta o Sistema Cartão Nacional de Saúde (Sistema Cartão). Brasília: Ministério da Saúde; 2011.

6. Brasil. Ministério da Saúde. Secretaria de Atenção à Saúde. Portaria Conjunta nº 2, de 15 de março de 2012. Dispõe acerca do preenchimento do número do Cartão Nacional de Saúde do usuário no registro dos procedimentos ambulatoriais e hospitalares. Brasília: Ministério da Saúde; 2012.
7. Brasil. Ministério da Saúde. Portaria nº 763, de 20 de julho de 2011. Dispõe acerca do preenchimento do número do Cartão Nacional de Saúde do usuário no registro dos procedimentos ambulatoriais e hospitalares. Brasília: Ministério da Saúde; 2011.
8. Brasil. Ministério da Saúde. Conselho Nacional de Secretários da Saúde. Nota técnica 22/2011. Proposta de consolidação do Cartão Nacional de Saúde – Cartão SUS. Atualização das notas técnicas 29/2010 e 32/2010 de 06/08 e 13/09/2010. Brasília: CONASS/progestores; 2011.
9. Magalhães VCL, Costa MCE, Pinheiro RS. Perfil do atendimento no SUS às mulheres com câncer de mama atendidas na cidade do Rio de Janeiro: relacionando os sistemas de informações SIH e APAC-SIA. Cad Saúde Coletiva. 2006; 14(2): 375-98.
10. Machado JP, Silveira DP, Santos IS, Piovesan MF, Albuquerque C. Aplicação da metodologia de relacionamento probabilístico de base de dados para a identificação de óbitos em estudos epidemiológicos. Rev Bras Epidemiol. 2008; 11: 43-54. http://dx.doi.org/10.1590/ S1415-790X2008000100004
11. Silveira DP, Artmann E. Acurácia em métodos de relacionamento probabilístico de bases de dados em saúde: revisão sistemática. Rev Saude Pública. 2009; 43(5): 875-82. http://dx.doi.org/10.1590/S0034-89102009005000060
12. Queiroz OV, Guerra Júnior AA, Machado CJ, Andrade EIG, Meira Júnior 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 de autorizações de procedimentos ambulatoriais de alta complexidade com os registros de sistemas de informações hospitalares. Cad Saúde Coletiva. 2010; 18(2): 298-308.
13. Gomes Jr SC dos S, Martino R, Almeida RT. Rotinas de integração das tabelas do sistema de autorização de procedimentos de alta complexidade em oncologia do Sistema Único de Saúde. Cad Saúde Coletiva. 2003; 11(2): 231-54.
14. Souza RC, Freire SM. Integração de dados ambulatoriais de quimioterapia e radioterapia registrados nas bases de dados do SUS. In: Congresso Brasileiro de Engenharia Biomédica, 24., 2014. Anais; 2014.
15. Souza RC, Freire SM, Almeida RT. Sistema de informação para integrar os dados da assistência oncológica ambulatorial do Sistema Único de Saúde. Cad Saúde Pública. 2010 Jun; 26(6): 1131-40. http://dx.doi.org/10.1590/S0102-311X2010000600007
16. Teixeira CLS, Bloch KV, Klein CH, Coeli CM. Método de relacionamento de bancos de dados do Sistema de Informações sobre Mortalidade (SIM) e das autorizações de internação hospitalar (BDAIH) no Sistema Único de Saúde (SUS), na investigação de óbitos de causa mal-definida no Estado do Rio de Janeiro, Brasil, 1998. Epidemiol Serviços Saúde. 2006; 15(1): 47-57. http://dx.doi.org/10.5123/S1679-49742006000100004
17. Fonseca MGP, Coeli CM, Lucena FFA, Veloso VG, Carvalho MS. Accuracy of a probabilistic record linkage strategy applied to identify deaths among cases reported to the Brazilian AIDS surveillance database. Cad Saúde Pública. 2010 Jul; 26(7): 1431-8. http://dx.doi.org/10.1590/S0102-311X2010000700022
18. Migowski A, Chaves RBM, Coeli CM, Ribeiro ALP, Tura BR, Kuschnir MCC, et al. Acurácia do relacionamento probabilístico na avaliação da alta complexidade em cardiologia. Rev Saúde Pública. 2011; 45(2): 269-75. http://dx.doi.org/10.1590/S0034-89102011005000012
19. Suzuki KMF. O uso de método de relacionamento de dados (record linkage) para integração de informação em sistemas heterogêneos de saúde: estudo de aplicabilidade entre níveis primário e terciário [tese]. Ribeirão Preto: Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo; 2012.
20. Girianelli VR, Thuler LCS, Silva GA. Qualidade do Sistema de Informação do Câncer do Colo do Útero no estado do Rio de Janeiro. Rev Saúde Pública. 2009; 43(4): 580-8. http://dx.doi.org/10.1590/S0034-89102009005000043
21. Mendes ACG, Lima MM, Sá DA, Oliveira LCS, Maia LTS. Uso da metodologia de relacionamento de bases de dados para qualificação da informação sobre mortalidade infantil nos municípios de Pernambuco. Rev Bras Saúde Matern Infant. 2012; 12(3): 243-9. http://dx.doi.org/10.1590/S1519-38292012000300004
22. Fávero JL, Cerqueira ACB, Fregona G, Prado TN, Werner RCD, Maciel ELN. Prevalência de tuberculose em profissionais da área de enfermagem obtida por método de relacionamento de banco de dados, 2000 a 2008, Espírito santo, Brasil. Rev Bras Pesq Saúde. 2012; 14(2): 31-7.
23. Girianelli VR. Fatores Associados ao Risco de Progressão para Câncer do Colo do Útero ou suas Lesões Precursoras em Mulheres com Exame de Papanicolaou Negativo: Um Estudo de Três Anos de Seguimento. Rio de Janeiro: Instituto Nacional de Câncer José Alencar Gomes da Silva; 2008.

24. Coeli CM. A qualidade do linkage de dados precisa de mais atenção. Cad Saúde Pública. 2015 Jul; 31(7): 1349-50. http://dx.doi.org/10.1590/0102-311XED010715
25. Brasil. Instituto Nacional de Câncer José Alencar Gomes da Silva. Coordenação Geral de Ações Estratégicas. Sistemas de Informação do Controle do Câncer de Mama e do Colo do Útero – Manual Gerencial. Rio de Janeiro: CEDC; 2011. 116 p.
26. Freire SM, Almeida RT, Cabral MDB, Bastos EA, Souza RC, Silva MGP. A record linkage process of a cervical cancer screening database. Comput Methods Programs Biomed. 2012; 108: 90-101. https://doi.org/10.1016/j.cmpb.2012.01.007
27. Cabral MDB. Proposta de relacionamento probabilístico dos registros da base de dados do programa de rastreamento do câncer do colo do útero [tese]. Rio de Janeiro: Universidade Federal do Rio de Janeiro; 2010.
28. Brasil. Ministério da Saúde. Definir como sistema de informação oficial do Ministério da Saúde, a ser utilizado para o fornecimento dos dados informatizados dos procedimentos relacionados ao rastreamento e a confirmação diagnóstica do câncer de mama, o Sistema de Informação do Controle do Câncer de Mama (SISMAMA). Portaria SAS n° 779 de 31 de dezembro de 2008. Brasília: Ministério da Saúde; 2008.
29. Brasil. Ministério da Saúde. Portaria GM nº 3.394 de 30 de dezembro de 2013. Institui o Sistema de Informação de Câncer (SISCAN) no âmbito do Sistema Único de Saúde (SUS). Brasília: Ministério da Saúde; 2013.
30. Tomazelli J. Avaliação das ações de detecção precoce do câncer de mama no Brasil: uma análise com base nos sistemas de informação em saúde [tese]. Rio de Janeiro: Instituto de Medicina Social da Universidade do Estado do Rio de Janeiro; 2016.
31. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013 [cited 2015 July 31]. Avaliable from: http://www.R-project.org/
32. Camargo Jr KR, Coeli CM. RecLink. Rio de Janeiro; 2007.
33. Passman LJ, Farias AM, Tomazelli JG, Abreu DM, Dias MB, Assis M, et al. SISMAMA: implementation of an information system for breast cancer early detection programs in Brazil. Breast. 2011; 20 (Suppl 2): S35-9. https://doi.org/10.1016/j.breast.2011.02.001

Received on: 09/14/2016
Final version presented on: 08/25/2017
Accepted on: 09/01/2017