RESEARCH ON BIOMEDICAL ENGINEERING
rbejournal.org
Volume 31, Number 3, p. 196-207, 2015

Original Article
DOI: http://dx.doi.org/10.1590/2446-4740.0666

Integrating Brazilian health information systems in order to support the building of data warehouses

Sergio Miranda Freire*, Rômulo Cristovão de Souza, Rosimary Terezinha de Almeida

Abstract **Introduction**: This paper's aim is to develop a data warehouse from the integration of the files of three Brazilian health information systems concerned with the production of ambulatory and hospital procedures for cancer care, and cancer mortality. These systems do not have a unique patient identification, which makes their integration difficult even within a single system. **Methods**: Data from the Brazilian Public Hospital Information System (SIH-SUS), the Oncology Module for the Outpatient Information System (APAC-ONCO) and the Mortality Information System (SIM) for the State of Rio de Janeiro, in the period from January 2000 to December 2004 were used. Each of the systems has the monthly data production compiled in dbase files (dbf). All the files pertaining to the same system were then read into a corresponding table in a MySQL Server 5.1. The SIH-SUS and APAC-ONCO tables were linked internally and with one another through record linkage methods. The APAC-ONCO table was linked to the SIM table. Afterwards a data warehouse was built using Pentaho and the MySQL database management system. **Results**: The sensitivities and specificities of the linkage processes were above 95% and close to 100% respectively. The data warehouse provided several analytical views that are accessed through the Pentaho Schema Workbench. **Conclusion**: This study presented a proposal for the integration of Brazilian Health Systems to support the building of data warehouses and provide information beyond those currently available with the individual systems.

Keywords: Record linkage, Health information systems, Database integration, Data warehouse, Neoplasms.

Introduction

In Brazil, the incorporation of new technologies requires the knowledge about their health benefits and economic and organizational impacts on the system (Brasil, 2011a). Cancer care is one of the areas with the highest volume and pressure for the inclusion of new technologies worldwide. In Brazil, this scenario is no different and can be observed in the list of items recommended by the Comitê Nacional de Incorporação de Tecnologias no Sistema Único de Saúde (SUS) (Brasil, 2011b). For the period from July 26th 2012 to November 8th 2013, 22% of the recommendations to incorporate technologies into SUS were aimed at cancer care (Comissão..., 2013).

In addition the price of new drugs and equipment is increasing and they do not necessarily bring proportional benefits and, in some cases, may also introduce new risks to health and have a relevant impact on health services. Thus, for the incorporation of new technologies, it is essential to know the benefits of the already incorporated technologies to the system in order to allow the comparison between the new and what is already in use.

In this sense, it is necessary to generate performance indicators (effectiveness, cost, adverse events among others) of the technologies in use and not only production indicators. This is possible due to the availability and access to SUS health information systems (mortality information system – SIM in Portuguese, hospital information system - SIH-SUS in Portuguese, outpatient information system - SIA-SUS in Portuguese, among others) (DATASUS, 2013) and the possibility of integrating the data records of these systems (Freire et al., 2012; Gomes and Almeida, 2004; Pires, 2011; Queiroz et al., 2009; Santos and Gutierrez, 2008). In cancer care, studies have demonstrated the possibility of linking records for generating indicators for the estimation of the required infrastructure, the follow up of clinical guidelines and the effectiveness of a screening program (Bastos, 2011; Costa, 2005; Gomes and Almeida, 2009).

Despite the efforts of the Department of Informatics of SUS in making data publicly available and some academic applications that make use of this data, the health system manager is still unable to use those indicators in daily practice. This is because those applications require technical knowledge and capabilities not found in the management environment. Therefore it is necessary to develop tools and environments to facilitate access to the data, and the construction of customized indicators, that will meet the daily needs

*e-mail: sergio@lampada.uerj.br

Received: 23 September 2014 / Accepted: 15 July 2015

of the different actors involved in the decision process in each technology stage in the health system.

Data warehouses (AD) are central data repositories that facilitate the production of queries, standard and/or ad hoc reports, easy browsing, data downloading and statistical analyses (Immon, 2005; Kimbal, 2008). In health, besides the management aspect, these tools are also used for clinical research (Chute et al., 2010; Lyman et al., 2008; Muranaga et al., 2007; Prevedello et al., 2010).

This study aimed to integrate the APAC-ONCO, SIH-SUS and SIM databases by means of probabilistic record linkage in order to build a data warehouse having the patient with cancer as the base for the integration and modelling of the data warehouse.

Methods

Databases integration overview

The study used data on cancer care of outpatients and inpatients for the years 2000-2004, provided by the Secretaria de Estado de Saúde do Estado do Rio de Janeiro, Brazil. Each of the systems has the monthly data production compiled in dbase files (dbf).

The ETL (Extract, Transform and Load) phase of the data warehouse implementation consisted of merging all the monthly files of the same system, record linking the systems and then loading the data warehouse. These steps will be explained below.

All the files pertaining to the same system were read into a table in a MySQL Server 5.1 (Oracle, 2013) scheme, called sisonco_etl, generating three tables: apac, aih and sim, corresponding respectively to the APAC-ONCO, SIH-SUS and SIM systems. The apac, aih and sim tables contained respectively 559,698, 4,360,917 and 712,172 records.

The tables of sisonco_etl were integrated by means of record linkage techniques explained in detail in Freire et al. (2012). A summary of the methodology and the differences regarding the process used in (Freire et al., 2012) are presented here.

The linkage process had the following steps: 1) cleaning and standardization of the tables, 2) deterministic linkage, 3) probabilistic linkage, and 4) linkage evaluation.

Data cleaning and standardization

All tables were subjected to a process of cleaning and standardization, generating the standardizedApac, standardizedAih and standardizedSim tables corresponding respectively to the apac, aih and sim tables. For the purpose of blocking only, the first letter of the second name, if not a vowel, was added to the end of the common first names “Maria”, “Ana”, “Adriana”, “Marcia”, “Joao”, and “Jose”, and to the beginning of the common surnames “Silva”, “Souza”, “Santos” and “Oliveira”, before being phonetically coded according to Freire et al. (2012).

In the cleaning of the address variable, values which indicated absence of address, accented characters, type of address and address complements were removed as well as words expressing titles.

Deterministic linkage

In all tables, two variables were added, id and uid, for the purpose of identification of records. id is an integer type, auto-incremented, and the primary key of each table. The variable uid, which is also an integer type, works as a unique identifier of the patient, and receives values during the linkage process according to the algorithms explained in Freire et al. (2012).

After the standardization, the standardizedApac was deterministically linked with itself using a key composed of the name, birth date, gender and individual taxpayer number (CPF in Portuguese) fields, thus generating the deterministicAPAC table. A similar process was performed on the standardizedAIH table, generating the deterministicAIH table. These tables were then probabilistically linked, as explained in the next section.

Probabilistic linkage

A probabilistic linkage process was performed, according to the following steps: 1) probabilistic linkage of the apac table with itself (internal linkage); 2) internal probabilistic linkage of the aih table; 3) probabilistic linkage of the apac and aih tables; and 4) probabilistic linkage of the apac and sim tables.

The main obstacle to the integration of the databases is the absence of a key that allows the identification of records pertaining to a single individual in the different systems in a deterministic way, in which case probabilistic record linkage techniques may be employed (Fellegi and Sunter, 1969; Newcombe et al., 1969). In the Fellegi and Sunter approach (1969), scores are calculated for each pair of records based on the contribution of each variable such as name, birth date and gender to the probability (P) of making an accurate classification. For each variable i, let:

$$m_i = \text{P(variable values agree | true pair)} \qquad (1)$$

$$u_i = \text{P(variable values agree | false pair)} \qquad (2)$$

$$1 - m_i = \text{P(variable values disagree | true pair)} \qquad (3)$$

$$1 - u_i = P(\text{variable values disagree} \mid \text{false pair}). \tag{4}$$

If the variable values agree, the weight for agreement is given by:

$$w_{ai} = \log_2 (m_i / u_i) \tag{5}$$

If the variable values disagree, the weight for disagreement is given by:

$$w_{di} = \log_2 [(1 - m_i) / (1 – u_i)] \tag{6}$$

A score for each pair is the sum of the weights for each variable used in the matching process under the independence assumption. The greater the score, the greater the likelihood that the pair of records belongs to the same individual.

All variables that were compared in the probabilistic linkage process are of the string type. In order to estimate m_i for the internal linkage of the aih table, an internal probabilistic linkage of the deterministicAih table was performed, using values of m_i of a previous work (Freire et al., 2012) or "guessed" (in the case of gender, address, patient record number, and individual taxpayer number). The resulting pairs were sorted in descending order of score and all pairs whose scores were above a certain threshold were considered to be true links (set T). For each of the compared variables, the number of pairs in T where they were equal was counted and the value of m_i was established as this count divided by the number of pairs in T. To estimate u_i, 1,000 records were randomly selected from the deterministicAih table, which were paired to each other and, for each variable, the number of agreements was divided by the total number of pairs. For the internal linkage of the deterministicApac table, the m_i values in Freire et al. (2012) were used, and the u_i values were estimated in a similar manner as the u_i values for the deterministicAih table, except that a random sample of the deterministicApac table was used instead. The values of m_i and u_i for each variable are shown in Table 1.

The following algorithms were used for comparing the values of the variables: equity, flexible equity, Winkler's adjusted weight, as explained in Freire et al. (2012), and frequency-based string. In the latter algorithm, if one of the strings is null then the weight is zero. When the two strings are not null, they are compared using the Winkler string comparator. If the value is above a certain threshold (0.9 in this work), then the weight is given by the $\log_2$ (1/name frequency) (Gill, 2001). Otherwise, the weight is given by w_{di}.

The equity algorithm was used for the comparison of the variables year of birth, month of birth, day of birth, gender, address, individual taxpayer number, and patient's medical record number. The flexible equity was used for the comparison of the patient's middle initials and mother's middle initials variables. The Winkler's adjusted weight was used for the comparison of the following variables in the apac internal linkage process: patient's first name, patient's middle name, patient's last name, mother's first name, mother's middle name and mother's last name. Finally, the frequency-based string algorithm was used for the comparison of the patient's first name, patient's middle name, patient's last name, mother's first name, mother's middle name and mother's last name variables in all other linkage processes.

Internal probabilistic linkage of the apac table

This linkage was carried out without any blocking. The comparison variables were: patient's first name, patient's middle name, patient's last name, mother's

Table 1. Estimated values of m_i e u_i for the processes of probabilistic record linkage of the apac and aih tables.

Variable	apac		aih	
	m_i	u_i	m_i	u_i
Year of Birth	0.92	0.018	0.956	0.016
Month of Birth	0.94	0.084	0.981	0.084
Day of Birth	0.91	0.033	0.971	0.034
Patient's Middle Initials	0.83	0.042	0.998	0.038
Patient's first name	0.92	0.029	0.954	0.0081
Patient's last name	0.93	0.021	0.990	0.034
Patient's middle name	0.9	0.008	0.970	0.009
Mother's first name	0.89	0.049	-	-
Mother's second name	0.9	0.008	-	-
Mother's last name	0.91	0.021	-	-
Mother's middle Initials	0.83	0.083	-	-
Gender	-	-	0.992	0.564
Address	-	-	0.719	0.00054
Individual taxpayer number	-	-	0.26	0.00001
Medical Record Number	-	-	0.775	0.00010

first name, mother's middle name and mother's last name, patient's middle initials and mother's middle initials, year of birth, month of birth and day of birth.

The results of the linkage process are stored in a results table that contains for each pair of compared records, their primary keys and the comparison score. An inspection table is generated from the results and deterministicApac tables and arranged in the descending order of the scores. This table was inspected by one of the authors who established a cutoff point above which all pairs correspond to true matches. The grey zone was not manually inspected and the uid values were established according to the algorithm presented in (Freire et al., 2012). This same procedure was carried out in all steps of the probabilistic linkage presented later. All pairs below the cutoff point which had the same values for the individual taxpayer number were manually inspected.

Internal probabilistic linkage of the aih table

The internal probabilistic linkage of the aih table was performed in three steps with the following blocking variables: 1) soundex of patient's first name and soundex of patient's last name; 2) soundex of patient's first name and year of birth; and 3) soundex of patient's last name and year of birth. The comparison variables were: patient's first name, patient's middle name, patient's last name, patient's middle initials, year of birth, month of birth, day of birth, gender, address, patient's identifier and patient's medical record number.

After the steps above, all pairs of the results table below the cutoff point and with identical values in certain variables were manually inspected and classified as true matches or not. The sets of pairs to be examined were obtained from queries, each with a different set of fields, namely:

- street, street number, address complement, zip code, city, hospital's identifier;
- street name, number, address complement, zip code, city and AIH number;
- street, street number, address complement, zip code, city and medical record number;
- street, street number, address complement, zip code, city and individual taxpayer number;
- street, street number, address complement, zip code and city;
- street, street number, address complement and zip code;
- street, street number and address complement;
- medical record number and AIH number.

The three steps of the probabilistic linkage process above were then repeated, now excluding the address, individual taxpayer number and patient's medical record number variables in the comparisons.

External probabilistic linkage of the apac and aih tables

The linkage of the deterministicApac and deterministicAih tables was performed in three steps. The blocking variables in each step were: 1) soundex of patient's first name and soundex of patient's last name; 2) soundex of patient's first name and year of birth; and 3) soundex of patient's last name and year of birth. The comparison variables were: patient's first name, patient's middle name, patient's last name, patient's middle initials, year of birth, month of birth, day of birth and gender.

External probabilistic linkage of the apac and sim tables

The deterministicApac and standardizedSim tables were linked in five steps, with the following blocking variables: 1) soundex of patient's first name and soundex of mother's first name; 2) soundex of patient's last name and soundex of mother's last name; 3) soundex of patient's first name and soundex of mother's last name; 4) soundex of patient's last name and soundex of mother's first name; and 5) month and day of birth. The comparison variables were: patient's first name, patient's middle name, patient's last name, patient's middle initials, mother's first name, mother's middle name, mother's last name, mother's middle initials, year of birth, month of birth, day of birth, and gender.

All pairs of records with the apac date after the death date were removed from inspection and records in the sim table associated to more than one patient in the deterministicApac table were manually inspected, as well as the associations between a patient and more than one record in the sim table.

The linkage process was performed on a computer with an Intel Quad Core (2.40 GHz, 2 x 4MB L2 Cache, 1066 MHZ) processor running Ubuntu Linux and accessing the databases in an Intel Xeon 3.2 Ghz, 3.5 GB RAM HP ML server running Linux Red Hat operational system.

Linkage evaluation

The quality of the linkage processes was evaluated by generating samples composed of all pairs of records which had the same values for the variables: street name, street number, zip code and state. Each sample was manually inspected to obtain the proportion of true pairs (sensitivity) and of true non-pairs (specificity) detected by the linkage process. The 95% confidence intervals for all measures of sensitivity and specificity were obtained with OpenEpi version 2.3.1 (Dean et al., 2007).

Data warehouse modeling and implementation

An intermediate database, sisonco_staging_area, was created during the ETL process. At the end of the linkage process the following tables were generated in the sisonco_staging_area: apac_ni and aih_ni, corresponding respectively to the tables apac and aih, excluding all identifying information (names, address, individual taxpayer number and medical record number); apac_aih_links containing all uid pairs associated through the linkage of the apac and aih tables; and apac_sim_links containing all uid pairs associated through the linkage of apac and sim tables. The sisonco_staging_area database was completed with other tables, such as time (presentation), gender, International Classification Diseases codes, and healthcare providers. Excluding the time table, the others were created based on data obtained from DATASUS.

The data warehouse was implemented following the dimensional model, and Kimball's star schema approach was adopted for modeling purposes. In the dimensional model, there are one or more central tables, called fact tables, that contain the processes metrics. A fact table is linked to others, called dimension tables (presentation, gender, healthcare providers, for instance), which give a context to the analyzed metrics.

When the ETL process was finished, data from the sisonco_staging_area was used as input for SISONCO_DW database.

The sisonco_staging_area and the SISONCO_DW databases were implemented with the MySQL database management system and Pentaho (Pentaho..., 2014), which is an open-source software that consists of the following tools:

- Pentaho Data Integration (PDI) – used to extract, transform and load the database (ETL). In this work, this tool was not used due to the fact that the linkage process was performed through another software.
- Pentaho Schema Workbench (PSW) – used to create XML schemas that are used by the Pentaho OLAP engine in order to process the searches based on the Multidimensional Expressions (MDX) language through a relational database and, thus, execute the analytical views (cubes). The analytical views in this work were based on the data available in the three information systems and in previous works (Costa, 2005; Gomes and Almeida, 2009);
- Pentaho Report Designer (PRD) –used for the creation of predefined reports;
- Pentaho Metadata Editor (PMEs) – in this work, the technical documentation corresponding to the back room metadata (Kimball and Caserta, 2004) was not developed. On the other hand, the PME was used to create an intuitive model which is presented to the end user at the moment of the creation of ad hoc reports. This way the end user does not need to know the table structures that compose the data warehouse.

An important question in the design of a data warehouse is the granularity definition, which defines the level of detail in which the information could be extracted from the data warehouse. Here the granularity that was available in the original dbf files was adopted. For instance, in relation to the temporal aspect of the procedures performed on the patients, the lowest level possible is the monthly level, because the SIH-SUS and APAC-ONCO systems record only the month and year when the procedure is performed.

This work has been approved by the research ethics committee of the Hospital Universitário Pedro Ernesto (CEP/HUPE – CAAe: 0153.0.228.000-07).

Results

Internal linkage of the apac table

The deterministic linkage resulted in a table with 78,717 records as compared to the original 559,698 records. The probabilistic linkage produced the results in Table 2.

Table 3 shows the quality of the linkage process. A total of 3,396 true links was found out of 3,398 identified by the linkage process, resulting in a sensitivity of 99.94 (99.79, 99.98). Since there were no false links, the specificity was 100% (99.9, 100) (Table 3 – apac).

Internal linkage of the aih table

The aih table had 70,963 records removed because the patients' names were not present or could not be identified, resulting in 4,289,954 records that were subjected to the linkage process. The deterministic linkage resulted in a table with 3,762,768 records. The probabilistic linkage produced the results in Table 2. In Table 3, a total of 2,906 true links was found out of 2,911 identified by the linkage process, resulting in a sensitivity of 99.83 (99.6, 99.93). Since there were no false links, the specificity was 100% (99.92, 100) (Table 3 – aih).

Linkage of the apac and aih tables

The probabilistic linkage of the apac and aih tables generated a total of 72,541 pairs of records where the first element belongs to the apac table and the second element belongs to the aih table. A total of 33,636 patients in the apac table had records in the aih table. The quality evaluation resulted in a sensitivity of 99.39% (99.2, 99.53) and a specificity of 100% (99.98, 100) (Table 3 – apac-aih).

Linkage of the apac and sim tables

The probabilistic linkage of the apac and sim tables generated a total of 22,024 patients in the apac table who had records in the sim table. The quality evaluation resulted in a sensitivity of 96.50% (95.31, 97.39) and a specificity of 100% (99.61, 100) (Table 3 – apac-sim).

Linkage processing time

The probabilistic linkage processing time varies, according to the size of the linked tables and the blocking strategy. The internal linkage of the apac table took 4d 23h 42 min 23s. For the internal linkage of the aih table, the processing time varied from 5h 56min 47s to 22h 57min 38s. The linkage time of the apac and aih tables varied from 9 min 12s to 37min 4s. The linkage time of the apac and sim tables varied from 9 min 31s to 2h 56min 47s.

Table 2. Relationship between the number of records and the number of patients in the apac and aih tables.

Number of records	Number of Patients	
	apac	aih
1	6,931	1,793,054
2	9,266	401,112
3	6,846	127,636
4	3,088	56,629
5	2,918	27,591
6	3,100	16,568
7	1,765	9,892
8	2,024	6,604
9	1,393	4,680
10	1,160	3,398
>10	14,944	23,836
Total	53,435	2,471,000

SISONCO_DW

Figure 1 presents a reduced model of the sisonco_staging_area with the main entities that make up the database. A patient (patient table) is associated to zero or more hospitalizations (hospitalization table) and could have one or more cancers (case_cancer). A case of cancer is defined as all registers that have the same values for the diagnosis, sex, birthdate and CPF fields (Gomes and Almeida, 2004). Therefore the same patient may be associated to several cases of cancer. A case of cancer is associated with a type of cancer (icd), with a set of outpatient encounters (ambulatorial_treatment) each in a specific healthcare unit (healthcare_unit) in a certain month and year (presentation), and submitted to one or more procedures (ambulatorial_procedure). Each hospitalization is associated with one or more procedures. A view provide all encounters of a single patient allowing the user to get a history of the patient.

From the SISONCO_DW, several analytical views were generated through the Pentaho Schema Workbench. The metrics and dimensions of each analytic view are presented in a simplified version in Table 4. For the patient analytic view, for example, the following metrics are available: number of patients and number of cases of cancer, the average age, the number and the percentage of deaths. Each metric can be analyzed in relation to gender, age group and place of residence. Similar interpretations can be made for the other facts tables.

The dimensional model for cases of cancer is presented in Figure 2. The facts table (f_case_cancer) contains the following metrics: the number of cases, the number of new cases, mean age, diagnosis-treatment

Table 3. Sensitivity and specificity of the internal linkage processes of the apac and aih tables, the linkage of the apac and aih tables, and of the apac and sim tables. T^+ and T^- represent linked pairs and not linked pairs as established by the linkage process.

	apac	aih	apac-aih			apac-sim		
Link	Yes	No	Yes	No	Yes	No	Yes	No
T^+	3396	0	2906	0	8908	0	1184	0
T^-	2	3905	5	4778	55	17515	43	987
Total	3398	3905	2911	4778	8963	17515	1227	987
Sensitivity (95% CI)	99.94 (99.79; 99.98)		99.83 (99.6; 99.93)		99.39 (99.2; 99.53)		96.50 (95.31; 97.39)	
Specificity (95% CI)	100 (99.9; 100)		100 (99.92; 100)		100 (99.98; 100)		100 (99.61; 100)	

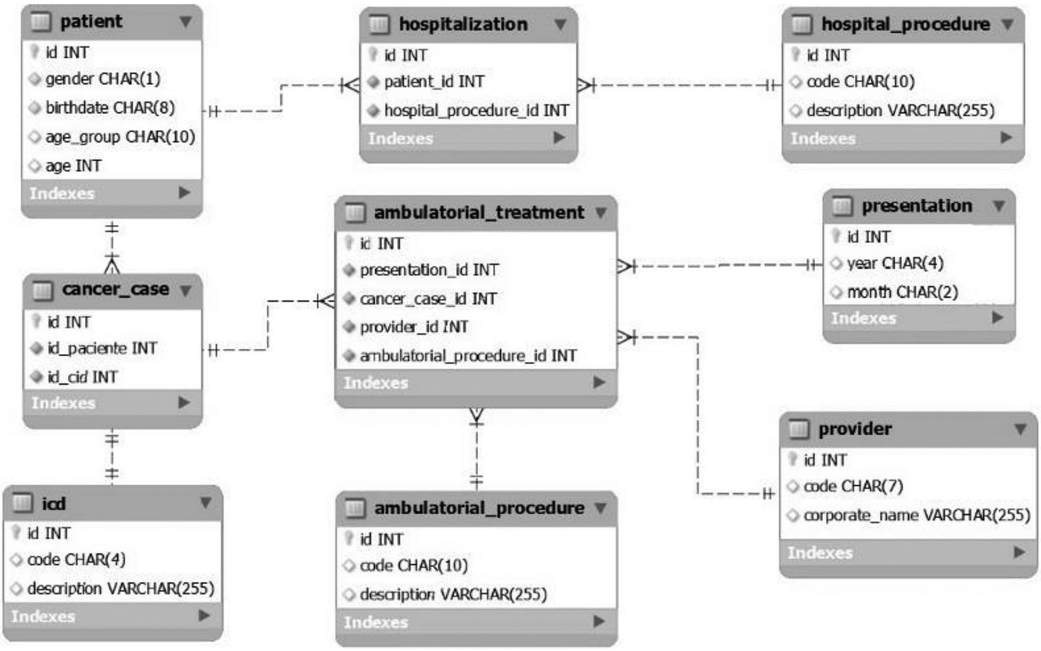


Figure 1. Partial entity-relationship model of the sisonco_staging_area.

Table 4. Analytic views of the SISONCO_DW.

Group	Metrics	Dimensions
Patients	Number of patients Number of cases of cancer Average age Number of deaths Percentage of deaths	Gender Age group Place of residence
Cases of Cancer	Number of cases Number of new cases Diagnosis-treatment interval Number of deaths Average age Diagnosis-death interval Percentage of deaths	Year Type of cancer Age group Gender Staging Metastasis Place of residence Provider
Outpatient services	Number requested procedures Number approved procedures Number irradiated fields Charged value Reimbursed value Average charged value	Year Treatment modality Provider Reason for discharge
Hospital admissions	Number of patients	Year Hospital Diagnosis on admission Reason for discharge Link to APAC

interval, diagnosis-death interval, the number of deaths and percentage of deaths. This table is linked to the dimensional tables year (dim_presentation), type of cancer (dim_cid), gender (dim_gender), age group (dim_age_group), staging of disease (dim_staging), metastasis (dim_metastasis), place of residence (dim_place) and healthcare provider of the first treatment (dim_healthcare provider).

Figure 3 shows a partial analytic view of the cases of cancer with the number of cases, the number of new cases, and average age metrics.

Figure 4 shows the history of cases of cancer, in which the user selects the year of first APAC, type of cancer, gender and age-group and gets a list of cases of cancer that meet the chosen criteria. For each case of cancer, the following data is available: age,

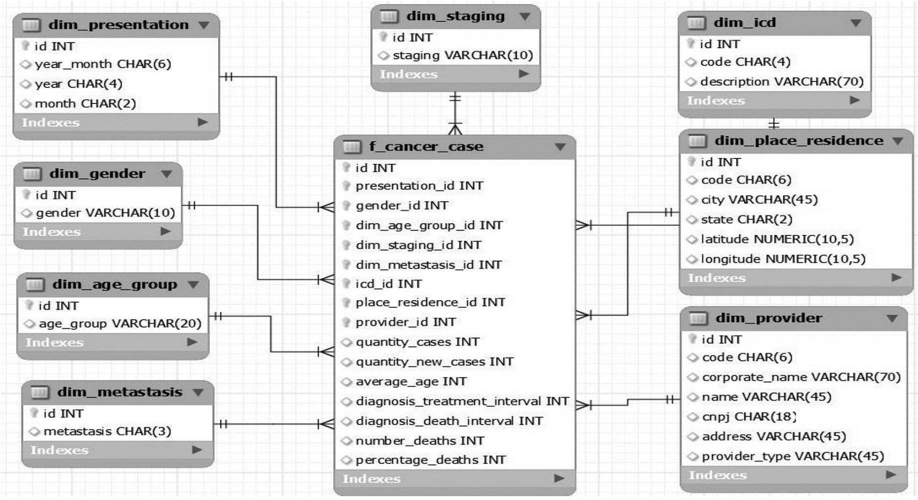
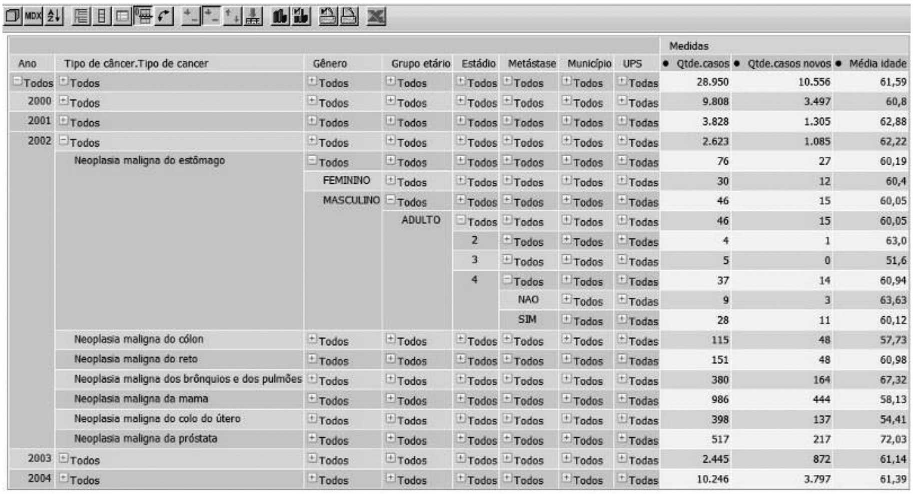
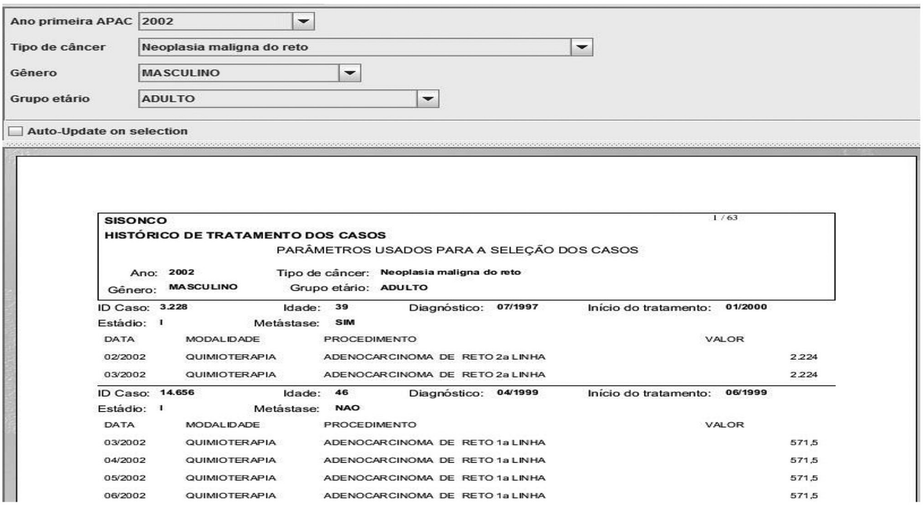


Figure 2. Dimensional model for the cases of cancer analytic view.



								Medidas		
Ano	Tipo de câncer.Tipo de cancer	Gênero	Grupo etário	Estádio	Metástase	Município	UPS	● Qtde.casos	● Qtde.casos novos	● Média idade
⊟ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	28.950	10.556	61,59
2000	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	9.808	3.497	60,8
2001	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	3.828	1.305	62,88
2002	⊟ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	2.623	1.085	62,22
	Neoplasia maligna do estômago	⊟ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	76	27	60,19
		FEMININO	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	30	12	60,4
		MASCULINO	⊟ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	46	15	60,05
			ADULTO	⊟ Todos	⊞ Todos	⊞ Todos	⊞ Todas	46	15	60,05
				2	⊞ Todos	⊞ Todos	⊞ Todas	4	1	63,0
				3	⊞ Todos	⊞ Todos	⊞ Todas	5	0	51,6
				4	⊟ Todos	⊞ Todos	⊞ Todas	37	14	60,94
					NAO	⊞ Todos	⊞ Todas	9	3	63,63
					SIM	⊞ Todos	⊞ Todas	28	11	60,12
	Neoplasia maligna do cólon	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	115	48	57,73
	Neoplasia maligna do reto	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	151	48	60,98
	Neoplasia maligna dos brônquios e dos pulmões	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	380	164	67,32
	Neoplasia maligna da mama	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	986	444	58,13
	Neoplasia maligna do colo do útero	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	398	137	54,41
	Neoplasia maligna da próstata	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	517	217	72,03
2003	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	2.445	872	61,14
2004	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todos	⊞ Todas	10.246	3.797	61,39

Figure 3. Partial analytic visualization of the cases of cancer with the metrics: the number of cases, the number of new cases and average age.



Ano primeira APAC: 2002

Tipo de câncer: Neoplasia maligna do reto

Gênero: MASCULINO

Grupo etário: ADULTO

□ Auto-Update on selection

SISONCO 1 / 63

HISTÓRICO DE TRATAMENTO DOS CASOS

PARÂMETROS USADOS PARA A SELEÇÃO DOS CASOS

Ano: 2002 Tipo de câncer: Neoplasia maligna do reto

Gênero: MASCULINO Grupo etário: ADULTO

ID Caso: 3.228 Idade: 39 Diagnóstico: 07/1997 Início do tratamento: 01/2000

Estádio: I Metástase: SIM

DATA	MODALIDADE	PROCEDIMENTO	VALOR
02/2002	QUIMIOTERAPIA	ADENOCARCINOMA DE RETO 2a LINHA	2.224
03/2002	QUIMIOTERAPIA	ADENOCARCINOMA DE RETO 2a LINHA	2.224

ID Caso: 14.656 Idade: 46 Diagnóstico: 04/1999 Início do tratamento: 06/1999

Estádio: I Metástase: NAO

DATA	MODALIDADE	PROCEDIMENTO	VALOR
03/2002	QUIMIOTERAPIA	ADENOCARCINOMA DE RETO 1a LINHA	571,5
04/2002	QUIMIOTERAPIA	ADENOCARCINOMA DE RETO 1a LINHA	571,5
05/2002	QUIMIOTERAPIA	ADENOCARCINOMA DE RETO 1a LINHA	571,5
06/2002	QUIMIOTERAPIA	ADENOCARCINOMA DE RETO 1a LINHA	571,5

Figure 4. History of treatment of a case of cancer.

month and year at diagnosis, month and year of first treatment, stage, presence of metastasis, and a list of procedures with the treatment modality, year and month of application, description and value.

In addition to the analytic views shown in Table 4, it is possible to build ad hoc reports whereby the user, from a set of variables described as metadata, can select the variables used for grouping (e.g. type of cancer), give details (which will appear in the body of the report) and filter the outputs.

Discussion

This study showed a proposal for integrating the Brazilian health information systems and for organizing data in a data warehouse that is a powerful tool to support healthcare planning and management. The integrated systems extend the scope of analyses far beyond those available with the individual and fragmented systems. The unification of the chemotherapy and radiotherapy monthly tables allows the use of a new unit of analysis: the case of cancer. Thus, it is possible to obtain information such as the number of cases by type of cancer, gender, age, city of residence, staging and indication of metastasis, as well as the time between diagnosis and treatment, sequence of types of procedures and duration of treatment. In addition, the linkage of the APAC, AIH and SIM databases was useful for: the recovery of the history of the patient in the health care network, since it is possible to obtain all outpatient procedures and admissions to hospitals related to cancer cases; and the reduction in the effect of underreporting of deaths in the APAC databases.

The usefulness of the SISONCO-DW must be evaluated by healthcare managers. We have not carried out this evaluation in this study due to resource and time constraints to identify public health managers with the ability to make use of these resources, since it is recognized that the majority of health managers in Brazil do not make use of the public health information systems as a management tool. In this respect, it would also be desirable to incorporate more recent data in SISONCO-DW. However, this would require a new request to research ethics committees and to central authorities in order to get access to new datasets, which would depart from the main objective of the study.

In a typical DW, the ETL step is said to take up to seventy percent of the time devoted to its development (Herzog et al., 2007). This study is no exception, especially considering the linkage process that was the most time-consuming task, as illustrated by the linkage process times, which did not include those tasks that required human intervention and the further setting of uid values.

The SISONCO_DW load was performed only once. A problem that must be dealt with in future loads is the definition of the update type for the slowly change dimensions (SCD). One of the dimensional tables which requires this kind of definition is the dim_provider table (Figure 2), whose type attribute can change with time. Therefore, the DW model could be kept unaltered by simply overwriting the attribute value with new ones (SCD 1); or by including a new row in the table with the new data (SCD 2); alternatively, the model could be altered through the creation of a new attribute to keep the old value of the current attribute (SCD 3). This decision will impact on the management of the identification keys of the healthcare providers and may lead to a loss of historical information.

In Brazil, other data warehouses have been developed based on the SUS data files (Pires, 2011; Santos and Gutierrez, 2008). Santos and Gutierrez (2008) developed a data warehouse from SIA, SIH, SIM, and other systems for the State of São Paulo without identifying the patients and developed a bespoken tool, called MINERSUS, to support data mining as well. Pires (2011) advanced one step further and integrated the systems in the previous work. The SISONCO_DW differs from the work of Pires (2011) in that: 1) The SISONCO_DW record linkage process is based on a solid theoretical foundation, extensively used in the literature. The strategy adopted for record linkage in Pires (2011) was *ad hoc*, although with similar values for sensitivity and a slightly lower value for the specificity than the ones obtained in this study; 2) SISONCO_DW was implemented with Pentaho which has a strong open source community behind it, and can therefore benefit from the data mining tools available with Pentaho; 3) SISONCO_DW is not a general purpose tool, but focuses on the cancer problem.

The SISONCO_DW has the following advantages:

- integration of databases: the linkage of the SIA-SUS, SIH-SUS and SIM systems expanded the information available on patients treated for cancer in the National Health System;
- information generation: through analytical tools, the end user can dynamically execute queries, build reports, export data to other software such as spreadsheets, and visualize them in a web-based interface. The data repository can also be used for data mining;
- Documentation - the development of a data warehouse requires the definition of data models, which contributes to a documentary reference of the structures of tables that make

up the database. Besides a metadata layer allows an understanding of these structures without requiring knowledge of database modeling;

- Does not require either computer skills or SQL knowledge from the end user in order to access the information.

The linkage process used a similar linkage strategy to Freire et al. (2012), and produced similar sensitivities and specificities, although the sensitivity values are higher in this study. Instead of using a random sample of pairs to evaluate the quality of the linkage process, which is not viable in this context, the samples used for quality assessment were based on agreement in one variable, address, that was also used for calculating the scores and this may have inflated the quality measures. Considering, however, that the influence of this variable on the scores of true pairs is small, it is believed that the actual quality values are not very discrepant with respect to those obtained here.

The weighting by the frequency of names produced a much greater dispersion in the score values as well as resulted in lower scores for quite common names, preventing those pairs from being classified as pairs when there was no agreement on other variables. This may have led to a greater power for discriminating pairs from non-pairs, but it needs further investigation.

The use of other variables for calculating the scores, such as the patient medical record number, address, individual taxpayer number and gender, also contributed to a greater dispersion of values. In relation to the address in this study, only the core of the street name was used and the comparison was performed using a plain equality of the strings. Although it seems to be better than not using the address for the comparison, the linkage process may benefit from the segmentation of the address into its constituent parts, as suggested by Churches et al. (2002) and the use of more effective string similarity comparators.

The linkage process allowed the identification of the individuals and the follow-up of his/her treatment within the public healthcare system. Gomes et al. (2003) integrated the apac files using the case of cancer as the unit of analysis. In the present study, 53,434 patients were identified, corresponding to 56,102 cases of cancer (1.05% greater), which corroborates the hypothesis that, in the absence of a strong person's identifier, the case of cancer may be taken as a surrogate for the individuals in integrating the apac files.

The deterministic linkage substantially reduced the amount of records to be linked probabilistically for the apac table (85.94%). As for the aih table, the reduction was small (12,29%), similar to what happened in a previous work (Freire et al., 2012). This is due to the fact that, in the apac table, the individual taxpayer number is a strong identifier and a required variable in the system, a single patient have multiple entries in the system and a good quality in the entry of the variables individual taxpayer number, patient's name, patient's mother name, and date of birth. This is not true of the inpatient system, where the individual taxpayer number is not a required field and 41.8% of the patients have only one entry in the system. Even so, it was found in both systems different individuals with the same individual taxpayer number and with different values in the variables date of birth and gender over time. This is in agreement with findings in other studies (Martins and Travassos, 1998; Souza et al., 2008).

The authors envisage several points for further investigation:

- application of other record linkage strategies to improve the efficiency of the process, e.g. parallelism, segmentation of names and addresses;
- improvement of the source code and the graphical user interface of the software developed for the record linkage and study of its integration with Pentaho;
- incorporation of new ways of presenting the reports in the data warehouse such as graphics and maps;
- partial automation of the update process of SISONCO_DW, by downloading the newest files available through the DATASUS site and their integration into the current data warehouse by means of the record linkage process.

Acknowledgements

This work was supported by grant 402711/2007-6 from CNPq, call MCT- CNPq/ANS – No 25/2007. The authors are grateful to Nelson de Albuquerque Souza e Silva, Ilara Hämmerli Sozzi de Moraes and Manoel Roberto da Cruz Santos of the Secretaria de Estado de Saúde do Estado do Rio de Janeiro for the access to the databases. The authors are grateful to the anonymous reviewers' comments on the draft version of this manuscript.

References

Bastos EA. Estimativa da efetividade do programa de rastreamento do câncer do colo do útero no estado do Rio de Janeiro [dissertation] [internet]. Rio de Janeiro: Universidade Federal do Rio de Janeiro; 2011 [cited 2014

Aug 16]. Available from: http://objdig.ufrj.br/60/teses/coppe_m/EdianeDeAssisBastos.pdf.

Brasil. Lei 12.401, de 28 de abril de 2011. Altera a Lei 8.080, de 19 de setembro de 1990, para dispor sobre a assistência terapêutica e a incorporação de tecnologia em saúde no âmbito do Sistema Único de Saúde - SUS. Diário Oficial da República Federativa do Brasil [internet], Brasília, abr. 2011a [cited 2014 Aug 11]. Available from: http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12401.htm.

Brasil. Decreto 7.646, de 21 de dezembro de 2011. Dispõe sobre a Comissão Nacional de Incorporação de Tecnologias no Sistema Único de Saúde e sobre o processo administrativo para incorporação, exclusão e alteração de tecnologias em saúde pelo Sistema Único de Saúde - SUS, e dá outras providências. Diário Oficial da República Federativa do Brasil [internet], Brasília, dez. 2011b [cited 2014 Aug 11]. Available from: http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Decreto/D7646.htm.

Churches T, Christen KL, Zhu JX. Preparation of name and address data for record linkage using hidden markov models. Biomed Central Medical Informatics and Decision Making. 2002 [cited 2014 Aug 16]; 2(9). Available from: http://www.biomedcentral.com/content/pdf/1472-6947-2-9.pdf.

Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. Journal of the American Medical Informatics Association. 2010; 17(2):131-5. http://dx.doi.org/10.1136/jamia.2009.002691. PMid:20190054.

Comissão Nacional de Incorporação de Tecnologias no Sistema Único de Saúde - Conitec. Decisões sobre a incorporação de tecnologias no Sistema Único de Saúde. Brasília; 2013 [cited 2014 Aug 16]. Available from: http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/leia-mais-o-ministerio/257-sctie-raiz/dgits-raiz/conitec/9027-relatorios-e-decisoes.

Costa MR. Comparação das condutas terapêuticas no tratamento ambulatorial das mulheres com câncer de mama [dissertation]. Rio de Janeiro: Universidade Federal do Rio de Janeiro; 2005.

DATASUS. Departamento de Informática do SUS, sistemas e aplicativos. Brasília; 2013 [cited 2013 Nov 16]. Available from: http://www2.datasus.gov.br/DATASUS/index.php?area=04.

Dean AG, Sullivan KM, Mir R. OpenEpi: Open Source Epidemiologic Statistics for Public Health. 2007 [cited 2014 Aug 16]. Available from: http://www.OpenEpi.com.

Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Association. 1969; 64(328):1183-210. http://dx.doi.org/10.1080/01621459.1969.10501049.

Freire SM, Almeida RT, Cabral MD, Bastos EA, Souza RC, da Silva MG. A record linkage process of a cervical cancer screening database. Computer Methods and Programs in Biomedicine. 2012; 108(1):90-101. http://dx.doi.org/10.1016/j.cmpb.2012.01.007. PMid:22341207.

Gill L. Methods for automatic record matching and linking and their use in national statistics. Norwich: Oxford University; 2001, [cited 2014 Aug 16]. (National Statistics Methodology Series, 25). Available from: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/gss-methodology-series--25--methods-for-automatic-record-matching-and-linkage-and-their-use-in-national-statistics.pdf.

Gomes SCS Jr, Almeida RT. Identificação de um caso novo de câncer no Sistema de Informação Ambulatorial do Sistema Único de Saúde. Cadernos Saúde Coletiva. 2004; 12:57-68.

Gomes SCS Jr, Almeida RT. Simulation model for estimating the cancer care infrastructure required by the public health system. Revista Panamericana de Salud Publica. 2009; 25(2):113-119. http://dx.doi.org/10.1590/S1020-49892009000200003.

Gomes SCS Jr, 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. Cadernos Saúde Coletiva. 2003; 11:231-54.

Herzog TN, Scheuren FJ, Winkler WE. Data quality and record linkage techniques. New York: Springer Science + Business Media, LLC; 2007.

Immon WH. Building the data warehouse. 4th ed. Wiley Publishing; 2005.

Kimbal R. The data warehouse lifecycle toolkit. Indianapolis: Wiley Publishing; 2008.

Kimball R, Caserta J. The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data. Indianapolis: Wiley Publishing; 2004.

Lyman JA, Scully K, Harrison JH Jr. The development of health care data warehouses to support data mining. Clinics in Laboratory Medicine. 2008; 28(1):55-71, vi. http://dx.doi.org/10.1016/j.cll.2007.10.003. PMid:18194718.

Martins M, Travassos C. Assessing the availability of casemix information in hospital database systems in Rio de Janeiro, Brazil. International Journal for Quality in Health Care. 1998; 10(2):125-33. http://dx.doi.org/10.1093/intqhc/10.2.125. PMid:9690885.

Muranaga F, Kumamoto I, Uto Y. Development of hospital data warehouse for cost analysis of DPC based on medical costs. Methods of Information in Medicine. 2007; 46(6):679-85. PMid:18066419.

Newcombe HB, Kennedy JM, Axford SJ, James AP. The use of medical record linkage for population and genetic studies. Methods of Information in Medicine. 1969; 8(1):7-11. PMid:5780257.

Oracle. MySQL database. 2013 [cited 2014 Aug 16]. Available from: http://www.oracle.com/br/products/mysql/index.html

Pentaho Corporation. Pentaho. 2014 [cited 2014 Aug 16]. Available from: http://www.pentaho.com.

Pires FA. Ambiente para extração de informação epidemiológica a partir da mineração de dez anos de dados do Sistema Público de Saúde [thesis]. São Paulo: Universidade de São Paulo; 2011.

Prevedello LM, Andriole KP, Hanson R, Kelly P, Khorasani R. Business intelligence tools for radiology: creating a prototype model using open-source tools. Journal of Digital Imaging. 2010; 23(2):133-41. http://dx.doi.org/10.1007/s10278-008-9167-3. PMid:19011943.

Queiroz OV, Guerra AA Jr, Machado CJ, Andrade ELG, Meira W Jr, Acúrcio FA, Santos W Fo, Cherchiglia ML. Building the national database on renal replacement therapy focused on the individual: probabilistic record linkage of death registries at the high complexity procedures authorization subsystem (Apac/SIA/SUS) and at the mortality information system (SIM) – Brazil, 2000-2004. Epidemiologia e Serviços de Saúde. 2009; 18(2):107-20.

Santos RS, Gutierrez MA. Ambiente computacional para extração de informações para a gestão da saúde pública por meio da mineração dos dados do SUS. Revista Brasileira de Engenharia Biomédica. 2008; 24(2):77-90. http://dx.doi.org/10.4322/rbeb.2012.050.

Souza RC, Pinheiro RS, Coeli CM, Camargo KR Jr. The Charlson comorbidity index (CCI) for adjustment of hip fracture mortality in the elderly: analysis of the importance of recording secondary diagnoses. Cadernos de Saude Publica. 2008; 24(2):315-22. http://dx.doi.org/10.1590/S0102-311X2008000200010. PMid:18278278.

Authors

Sergio Miranda Freire[1,2]*, Rômulo Cristovão de Souza[1], Rosimary Terezinha de Almeida[3]

[1] Programa de Pós-Graduação em Ciências Médicas, Universidade do Estado do Rio de Janeiro – UERJ, Rio de Janeiro, RJ, Brazil.

[2] Departamento de Tecnologias da Informação e Educação em Saúde, Universidade do Estado do Rio de Janeiro – UERJ, Av. Prof Manuel de Abreu, 2º andar, CEP 20550-170, Rio de Janeiro, RJ, Brazil.

[3] Programa de Engenharia Biomédica, Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa em Engenharia, Universidade Federal do Rio de Janeiro – UFRJ, Rio de Janeiro, RJ, Brazil.