DETECTION OF INCONSISTENCIES IN HOSPITAL DATA CODING

Juliano Gaspar, Fernando Lopes and Alberto Freitas
CIDES – Department of Health Information and Decision Sciences, Porto, Portugal
CINTESIS – Center for Research in Health Technologies and Information Systems, Porto, Portugal
Faculty of Medicine, University of Porto, Porto, Portugal

Keywords: Inconsistencies, Errors, Data quality problems, Hospital databases.

Abstract: Introduction: Health professionals need data, in sufficient quantity and quality, and tools that can manage the vast amount of available data. They need help for data management and appropriate support for decision making.
Aim: The focus of this study is to develop a prototype that can contribute to the identification of data quality problems in clinical and administrative data.
Methods: Methods involve the definition of requisites and business rules, the prototype development and testing, and the realization of two studies using the prototype.
Results: Studies performed using the prototype resulted in the detection of many data problems and inconsistencies. Amongst those we can point out, for instance, that 82,000 (15%) episodes had '*diagnostic code does not exist in ICD-9-CM table*' and that 783 (0,2%) episodes within '*female breast cancer*' had the variable *gender* equal to '*male*'.
Discussion: This prototype, besides contributing to the detection of data quality problems, is also expected to be an incentive to the improvement of information system architectures. It shows the importance of the development of mechanisms to detect and validate data in health environments.

1 INTRODUCTION

Medical databases generally involve all sorts of data and store a considerable size of information (Kumar et al., 2008). Data consists, generally, of records with different type of information and with diverse characteristics such as age, blood type, weight, clinical images, diagnosis, lab results, among many other patient details (Chandola et al., 2009). In the latest years, with an exponential growth in hospital information, the volume of electronic clinical records has also significantly increased. Impaired with this increase, new interests emerged in the analysis of such information, thus driving it not only to be used as a source to make decisions but also to support hospital management (Freitas et al., 2005).

Amongst the various health services, hospital focus gains a particular interest because it gathers a higher complexity and procedures cost. This makes hospital databases the main focus of various analysis (Freitas et al., 2005). The display of reliable and trustworthy information, obtained from solid data, is vital to help health professionals and technicians in the process of decision making and also help in higher levels of hospital management. In this context, Information Systems are responsible for making, analyzing and disseminating such data (Pinto, 2010).

1.1 Data Quality

The data value is directly proportional to its quality. So, the higher the quality is, the greaters its utility (Arts et al., 2002). It is important to consider that, to insure data quality, certain basic rules must be followed such as coherency, accuracy, integrity and consistency (Tayi and Ballou, 1998); (Wang, 1998); (Silva-Costa et al., 2010). Thus, the importance of rectify integrity problems, normalize values, fill or identify missing values, identify redundant information, among other processes in database management (Barateiro and Galhardas, 2005).

The quality of data, accordingly to some authors, is a relative concept and is directly correlated to the

Gaspar J., Lopes F. and Freitas A..
DETECTION OF INCONSISTENCIES IN HOSPITAL DATA CODING.
DOI: 10.5220/0003757301890194
In *Proceedings of the International Conference on Health Informatics* (HEALTHINF-2012), pages 189-194
ISBN: 978-989-8425-88-1

information purpose. The authors use the "fitness for use" concept that emphasizes the importance of users' point of view and their judgment in information assertiveness (Tayi and Ballou, 1998); (Wang, 1998); (Olson, 2003). Especially in medical databases, this concept reinforces itself, because data can have sufficient quality for economic analysis but insufficient to clinical or epidemiological studies (Silva-Costa, 2010); (Cruz-Correia et al., 2009).

In the scenario of hospital department, electronic data quality problems (DQP) gains a higher attention because the information has a vital importance for patient care (Daniel et al., 2008). However, due to a high level of complexity and integrity inherent to this activity, maintaining a high quality database is not an easy chore.

1.2 Medical Data Coding Guidelines

There's many disparity in health policies and hospital funding between different countries. Despite those differences, there are common methods, characteristics and needs in all hospital environments which make it possible to define a set of international rules to analyze and compare different hospitals in different countries, having in consideration its' health records (Pinto, 2010).

After being collected and coded, medical data concerning patients, usually has billing information frequently used only for administrative and financial purposes. This does not prevent them from being usefull to other areas such as health quality and health care (Price et al., 2003).

These databases have a huge statistical potential, due to their large volume and similarity between hospitals, intitutions or even countries. This comparation between hospitals is due, mainly, to the use of a core group of variables named internationaly as "Minimum Basic Data Set" (MBDS) (Aylin et al., 2007); (Romano et al., 1995).

A part of this coded data is generated in the hospital discharge codification process. In this process, the information collected from clinical records, mainly diagnosis and procedures, is generally coded according to the "International Classification of Diseases" (ICD). It is also recorded the "Diagnostic Related Group" (DRG) in which episode is grouped. The use of ICD-9-CM[1], along these past years, has had many purposes such as supporting the health services payment, studding their costs, evaluate quality, planning future needs and helpping clinical research (Ginde et al., 2008), among others.

In Portugal, systematic codification of hospitalizations in 1989 (Lopes, 2010) and it was addopted the use of DRG as a hospital "case mix" measurement. Nowadays, the main hospital information systems have databases made of clinical and administrative data, which were compiled accordingly to the European established MBDS (Oliván, 1997).

In Portugal, the Health Ministery defines and publishes regulations that, not only, determine the prices to be charged by the health delivered in the National Health Service (NHS), but also clearly distinguishes between what should be an inpatient or an ambulatory episodes, among other rules for the codification in DRG.

The evaluation of clinical coding quality is essential for the correct attribution of hospital funding, as hospital evaluation and funding theirselves depend in DRG grouping (Pinto, 2010). So, all MBDS potential for hospital management and research can be compromised if clinical codification does not have certain levels of quality.

2 AIM

The aim of this study is to develop a prototype for the detection of data quality problems and inconsistencies in inpatient and ambulatory episodes in hospital databases. The prototype is intended to detect common errors in health related databases and also contextualize errors in specific medical specialties.

3 METHODS

Prototype and the studies regarding its use the development process follow the following predefined steps:

- Analyze the central database structure and the coding guidelines;
- Identify the business rules, define prototype requisites, implement and realize the tests;
- Perform a two studies including data from the last 10 years, the first from a hospital and second including data concerning a pre-determined disease diagnosed in all NHS hospitals.

[1]ICD-9-CM: International Classification of Diseases, Ninth revision, Clinical Modification.

The database used is composed of hospital episodes, data ambulatory and inpatient data, from medical or surgical specialties, concerning 96 Portuguese hospitals with discharges between the years 2000 and 2009, making a total of approximately 12 million records. Access to this data was possible due to the ACSS[2] collaboration in the HR-QoD[3] project developed by the Department of Health Information and Decision Sciences of Faculty of Medicine of University of Porto. All data used was anonymous and clinical data was coded in ICD-9-CM.

Based on the defined requisites and business rules, auxiliary tables were created to store information from the DQP detected such as error codes and the errors quantity detected in each hospital and year. Business rules contain, rules that validate whether the information contained in a record is consistent, meaning that it does not present incongruences between the values of a variable in relation to values of other variables in the same episode. If the values are accordingly to regulations and guidelines of that codifications used in Portugal, such as ICD-9-CM. These were created and schematized as showed in figure 1.

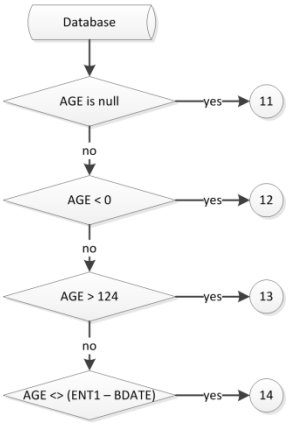

Figure 1: Business rule to variable "AGE".

In the implementation of the prototype the following programming languages were used: PHP, HTML and JavaScript the Eclipse developing tools for PHP and Toad for Oracle were also used.

[2]ACSS: Administração Central do Sistema de Saúde (in portuguese).

[3]HR-QoD: Quality of data (outliers, inconsistencies and errors) in hospital inpatient databases: methods and implications for data modeling, cleansing and analysis.

4 RESULTS

With its' use it was possible to create reports with the detection of DQP and respective incongruences. These reports included the analysis of over 50 different types DQP in the studied data.

As result of a first study, data from a hospital concerning discharges between the years 2000 and 2009 (inclusive) was selected, totalizing 543,133 hospital episodes. Some of the problems detected are shown in table 1.

Table 1: DQP Detected.

Problem Description	N	%
Incorrect length of stay (LOS)	191,282	35.22
Diagnostic code not in ICD-9-CM table	82,422	15.18
Procedure code not in ICD-9-CM table	9,023	1.66
Principal diagnostic is null	85	0.02

The evolution, in the 10 years of determined detected DQP can be seen in graphics presented in figure 2 and 3.

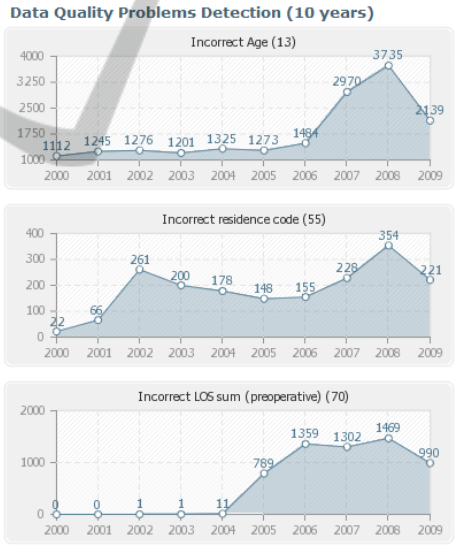

Figure 2: Data Quality Problems within 10 years.

In the figure 2 we can see that, for example, the growth (after 2007) errors in the variable 'AGE' (calculated incorrectly) and also that the total days in preoperative are incorrectly calculated after 2004.

The figure 3 show a gradual decline of errors related to 'weight at birth', the accentuated increase of 'repeated diagnosis' in the same episodes between 2007 and 2008, the gradual increase related to 'inter-hospital transfers' and identical volumes of errors between 2000 and 2004 related with 'LOS sum (preoperative)' and 'surgical date'.

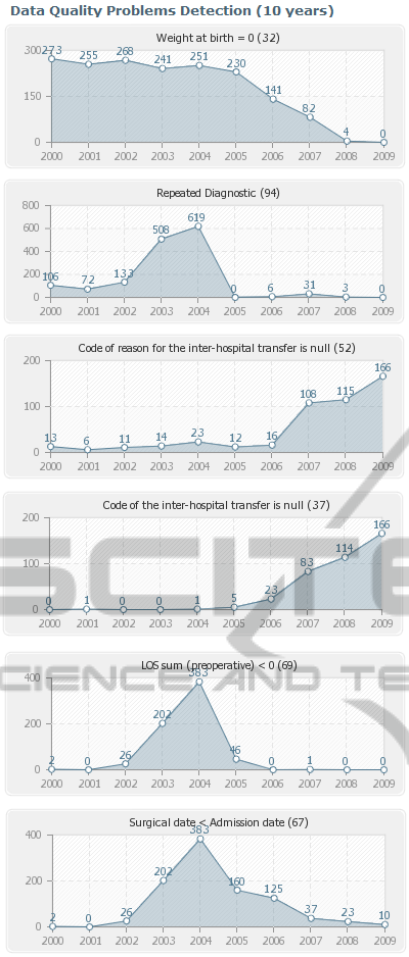

Figure 3: Data Quality Problems within 10 years.

For the second study, the pathology Female Breast Cancer (FBC) was chosen due to a publication released in January 2011 by INE[4], stating that in Portugal 1,635 people died in the year 2009 (INE, 2011) due to FBC and according to WHO[5], FMC is the most frequent tumor worldwide. It's the main cause of female death internationally (Bastos et al., 2007); (WHO, 2011).

To this purpose, all episodes with diagnosis of FBC coded in ICD-9-CM with '174.x' were selected in all 96 Portuguese hospitals in the last 10 years, resulting 346,654 records.

Specific queries were applied in order to detect inconsistencies in the context of codification in this disease. Table 2 shows the most relevant inconsistencies detected.

[4]INE: Instituto Nacional de Estatística (in Portuguese)
[5]WHO: World Health Organization.

Table 2: Contextual DQP Detected.

Contextual Problem Description	N	%
'V58.0' in secundary diagnostics	66,875	19.34
'V58.1' in secundary diagnostics	72,281	20.90
'V10.x' in principal diagnostic	830	0.24
Episodes with 'gender' variable = 'male'	783	0.23

From this analysis we can point out 783 records with value "male" in the variable 'gender'. This fact represents a serious DQP because this study is specifically about FBC. Male Breast Cancer is coded in ICD-9-CM with '175.x', and so code '174.x' is restricted to female patients.

5 DISCUSSION

The algorithm was capable of identifying the most significant DQP existing in the database.

The "Incorrect length of stay (LOS)", one of the DQP detected in 35% of total episodes, observes some particular considerations: (a) the regulation number 839-A/2009 from the health ministry informs that "Global Length of Stay" (G-LOS) concerns the total number of days consumed by each patient within the hospital, considering the admission day and ignoring the discharge day (SNS, 2009); (b) a most detailed analysis of this inconsistency value showed that, despite the business rule is correct with the in vigor regulation, this value can be caused by inconsistencies in informatics applications that calculate the variable G-LOS; (c) some specialists make their calculations only with the value of the day in question; meanwhile other consider also the hours and minutes, which can cause one day difference in G-LOS.

DQP occurrences DQP "Diagnostic code not in ICD-9-CM table" and "Procedure code not in ICD-9-CM table", showed in table 1, are due essentially with the passing years, to the reviewing of some codes and the modification or others. However, the process associated to the management of these modifications is not performed efficiently. This process would have to store the different versions of changes made to the ICD-9-CM table used in each hospital as well as to store the version used in the codification process instant (Lopes, 2010).

Codification errors on principal diagnosis, such as the 85 cases presented on table 1, imply the classification of episodes with DRG 470 ("Ungroupable") in these cases, the hospital is not financially compensated for the episode costs (Silva-Costa et al., 2010).

The gradual decrease of DQP related with 'weight at birth', presented on figure 1, could be associated with data validation rules, performed during data entry, by the different electronic applications used on NHS hospitals during the last years.

In figure 3 a gradual increment related to patient transfers between hospitals after the year 2006 can be observed. DQP "Code of the reason for the inter-hospital transfer is null" and "Code of the inter-hospital transfers is null", raise the issue that the lack of these codes could be related to regulations that the health ministry posted in 2003: (a) a hospital that transfers a patient to other, by "lack of resources", can only receive up to 50% of the correspondent DRG; (b) the hospital that receives a transferred patient is limited, in its' classification, to a restrict set of DRG.

Despite that these regulations already existed in 2003, only after 2006 all NHS hospitals had these regulations. Regardless of the cause, it is a fact that this problem has obvious implications in hospital reimbursement level.

An analysis to identify the existing relation between errors found in years 2000 to 2004, in DQP sum (preoperative) < 0' and 'Surgical date < Admission date' (Figure 3), revealed that all related episodes had many patients that were operated in the emergency service and then admitted in the hospital. After 2004 the 'Admission date' to be considerate was the lesser date registered, being the emergency service entry date or the admittance date in the hospital services.

The identification of 783 cases coded with '174.x' and with the 'gender' variable equal to 'male', in the second study, is of extreme importance. Assuming that these records are actually male patients, this inconsistence represents a 0.23% impact in this study. However, if the focus of the studies was Male Breast Cancer, coded in ICD-9-CM as '175.x', these 783 cases wrongly coded would correspond to a total of 25% cases of Male Breast Cancer that may not be considered.

Official coding guidelines (ACSS, 2009) state that "in admittance of the patient exclusively to chemotherapy, imunohemoterapy or radiotherapy, the principal diagnosis code attributed should be 'V58.x', and secondary diagnosis should be the ones representing the cause for treatment".

After these directives, we can verify that in 139,156 records, 40.2% of total episodes of FBC (Table 2) present 'V58.0' or 'V58.1x' as secondary diagnosis codes. The average rate of this inconsistency is 64% per year. The directives also state that the code 'V10.x' ('personal history of malign cancer') should always be coded as a secondary diagnosis, however 830 records were found with 'V10.x' was the principal diagnosis.

6 CONCLUSIONS

This study helped the analysis and identification of existing DQP in a coded health database, and also helped in the determination of a profile for the coding of FBC episodes.

Acknowledge of anomalies in the data may allow an immediate improvement in their quality, when the correction is feasible. This acknowledgement is important for the management of health, both nationally and regionally, and also locally in each hospital or health facility.

The results are graphically simple and allow a direct reading of the most relevant values and are certainly an asset to help managing and controlling the efficiency of health facilities.

The visualization of DQP, in graphical format over 10 years, allowed users, only with a visual analysis of the evolution of the problem, to quickly identify possible relationships between the increase or decrease of a problem and the date of implementation of certain software in the hospital, or the adoption of a new standard for procedure..

Effective management of versions, also known as change management, applied both by managers and by the NHS hospital managers, in addition to companies that develop software for health, can help to prevent problems to happen again, as some of those presented in this study.

This study is a contribute to the detection of DQP, and also an incentive to the improvement of the architecture of existing information systems, emphasizing the importance of developing mechanisms and methods for detection and validation in health. This study also intends to alert for the impact of these errors in epidemiologic studies and, as well, in management evaluations and health related policies.

6.1 Future Works

In the next stages of this study, we intend to replicate the method used in this study to other pathologies and apply data mining algorithms to improve the detection of patterns, abnormal cases, possible errors or inconsistencies.

ACKNOWLEDGEMENTS

The authors would like to thank the support given by the research project HR-QoD - Quality of data (outliers, inconsistencies and errors) in hospital inpatient databases: methods and implications for data modeling, cleansing and analysis (project PTDC/SAU-ESA/75660/2006).

REFERENCES

ACSS, (2009). CID-9-MC Guidelines Oficiais para Codificação: Em vigor a partir de Outubro de 2009 (Lisboa, *Administração Central do Sistema de Saúde*, IP).

Arts, D., Keizer, N., and Scheffer, G.-J., (2002). Defining and Improving Data Quality in Medical Registries: A Literature Review Case Study, and Generic Framework. In: *J Am Med Inform Assoc* 9, 600-611.

Aylin, P., Bottle, A., and Majeed, A., (2007). Use of administrative data or clinical databases as predictors of risk of death in hospital: comparison of models. In: *BMJ* 334, 1044.

Barateiro, J., and Galhardas, H., (2005). A Survey of Data Quality Tools. In: *Datenbank-Spektrum* 5, 15-21.

Bastos, J., Barros, H., and Lunet, N., (2007). Evolução da Mortalidade por Cancro da Mama em Portugal (1955-2002). In: *Acta Med Port*, 139-144.

Chandola, V., Banerjee, A., and Kumar, V., (2009). Anomaly Detection: A Survey. In: *ACM Computing Surveys* 41.

Cruz-Correia, R., Rodrigues, P., Freitas, A., Almeida, F., Chen, R., and Costa-Pereira, A., (2009). Data Quality and Integration Issues in Electronic Health Records. In Information Discovery on Electronic Health Records. *In: CRC Data Mining and Knowledge Discovery Series*, H. V. C.a. Hall, ed., pp. 55-95.

Daniel, F., Casati, F., Palpanas, T., Chayka, O., and Cappiello, C., (2008). Enabling Better Decisions through Quality-aware Reports. In: *International Conference on Information Quality (ICIQ)* (USA).

Freitas, A., Brazdil, P., and Costa-Pereira, A., (2005). Mining Hospital Databases for Management Support. Paper presented at: *IADIS Virtual Multi Conference on Computer Science and Information Systems*, pp. 207-212.

Ginde, A. A., Tsai, C. L., Blanc, P. G., and Camargo, C. A., Jr., (2008). Positive predictive value of ICD-9-CM codes to detect acute exacerbation of COPD in the emergency department. In: *Jt Comm J Qual Patient Saf* 34, 678-680.

INE, (2011). Boletim mensal de estatística: Janeiro de 2011. Available in: *http://www.ine.pt. Access In:* 02/03/2011.

Kumar, V., Kumar, D., and Singh, R. K., (2008). Outlier Mining in Medical Databases: An Application of Data Mining in Health Care Management to Detect Abnormal Values Presented In Medical Databases. In: *IJCSNS International Journal of Computer Science and Network Security*, 272-277.

Lopes, F., (2010). Portal da Codificação e dos GDH. *Available in: http://portalcodgdh.min-saude.pt.* Access In*:* 31/10/2010.

Oliván, J. A. S., (1997). Sistemas de información hospitalarios: el CMBD. In: *Scire: representación y organización del conocimiento* 3, 115-130.

Olson, J. E., (2003). *Data Quality* - The Accuracy Dimension, *Morgan Kaufmann Publishers edn*.

Pinto, R., (2010). Sistemas de informações hospitalares de Brasil, Espanha e Portugal - Semelhanças e diferenças. In: *FIOCRUZ* (Rio de Janeiro, Escola Nacional de Saúde Pública Sergio Arouca), pp. 162.

Price, J., Estrada, C. A., and Thompson, D., (2003). Administrative Data Versus Corrected Administrative Data. In: *Am J Med Qual* 19, 38-44.

Romano, P. S., Zach, A., Luft, H. S., Rainwater, J., Remy, L. L., and Campa, D., (1995). The California Hospital Outcomes Project: using administrative data to compare hospital performance. In: *Jt Comm J Qual Improv* 21, 668-682.

Silva-Costa, T., (2010). Indicadores de Produção Hospitalar - Uma forma de medir a produção dos hospitais Portugueses (Porto, *Faculdade de Medicina da Universidade do Porto*), pp. 172.

Silva-Costa, T., Marques, B., and Freitas, A., (2010). Problemas de Qualidade de Dados em Bases de Dados de Internamentos Hospitalares. Paper presented at: *5ª Conferência Ibérica de Sistemas e Tecnologias de Informação* (Santiago de Compostela).

SNS, (2009). Portaria n.º 839-A/2009 - Ministério da Saúde de Portugal, M.d. Saúde, ed. (Lisboa, Diário da República), pp. 4978-(4972) a 4978-(4124).

Tayi, G. K., and Ballou, D. P., (1998). Examining Data Quality. *In: CACM* 41, 54-57.

Wang, R. Y., (1998). A Product Perspective on Total Data Quality Management. In: *CACM* 41, 58-65.

WHO, (2011). World Health Organization. Available in: *http://www.who.int/en/. Access In:* 20/07/2011.