QUALITY OF DATA FROM CENTRAL AND DEPARTMENTAL INPATIENT DATABASES

A Comparative Study

Bernardo Marques[1,2], Eliana Sousa[1,2], Tiago Silva-Costa[1,2], Ricardo Correia[1,2] and Alberto Freitas[1,2]
[1]*Department of Health Information and Decision Sciences, Faculty of Medicine, University of Porto, Porto, Portugal*
[2]*CINTESIS - Center for Research in Health Technologies and Information Systems, University of Porto, Porto, Portugal*

Keywords: Data quality problems, Administrative data, Hospital information systems.

Abstract: This paper is a preliminary study over the problems resulting from the integration of a departmental information system database over a central database. This work will allow the comparison between the quality of the data collected for clinical purposes by a medical department, and the data collected for administrative and epidemiological purposes in a central hospital database. It is expected that the different purposes for these two data collections can have an impact on data consistency, namely on it completeness and detail of information, among other data quality problems. We expect to detect the type of variables that are better recorded in each information system, by calculating and comparing the quality of similar variables. We also expect to detect differences between both systems in the registries of the same patients. This paper can play an important role for better understanding the quality of the integration of departmental systems in the general hospital information system, pointing out some limitations about consistency and information flow. It is also our goal to suggest some recommendations and strategies to prevent data quality problems and to improve communication between central and departmental databases.

1 INTRODUCTION

Over the past years we have been witnessing an improvement of medical registries along with the development of even more capable digital systems and warehouse capacity. The exponential growth of information has led to an intensification of interest in exploring the information collected, not only for clinical decisions and research studies but also for hospital management. The information value is strongly dependent on the quality of the data contained in the registry (Arts et al., 2002). Therefore, studies regarding data quality are now even more relevant as the utilization of these databases increase in magnitude and importance (Freitas et al., 2010b). Particularly, in Portugal, many efforts have been done to study the scale of data quality issues over hospital databases and their implications to decision makers, administrators and researchers (Freitas et al., 2010a); (Silva-Costa et al., 2007); (Silva-Costa et al., 2010).

Regarding central databases in health care arena, the Portuguese National Health Service (NHS) has, since 1990, a system called SONHO for the management of hospital patients. This system allows the registry of patients and departments, as pharmacy, blood or surgery, and is used in all NHS public hospitals. The integration of this system had a positive impact both in productivity and improvement of diagnostic techniques (Dismuke and Sena, 1999). This system have been collecting data systematically as patients flow over the Portuguese public hospitals, gathering huge amounts of data ready to be explored.

Apart from the fact that SONHO is not accessible to every health professionals or researchers, one problem of this system is that the database model is so complex and the amount of data is so big that studies over this information are as yet quite limited. Thereby directors and staff from hospital departments have been working with developing teams implementing and integrating different information systems over SONHO (Cruz-Correia, 2010). There are multiple advantages for the integration of these systems, namely the easy access to the collected information, different database structure and more specific information for

Marques B., Sousa E., Silva-Costa T., Correia R. and Freitas A..
QUALITY OF DATA FROM CENTRAL AND DEPARTMENTAL INPATIENT DATABASES - A Comparative Study.
DOI: 10.5220/0003759901950200
In *Proceedings of the International Conference on Health Informatics* (HEALTHINF-2012), pages 195-200
ISBN: 978-989-8425-88-1

clinical purposes. The integration of these information systems has also some inherent risks or disadvantages in particular if the communication with the central system (SONHO) is not as effective as it should be. This fact can lead to several data quality problems and/or to two different sets of data instead of two sets with the same information but with different purposes (Cruz-Correia et al., 2006).

This paper is a preliminary study over the referred problems resulting from the integration of a departmental system database over a central administrative database. This work should allow the comparison between the quality of the data collected by medical departments for clinical purposes, and the data collected in central hospital databases for administrative and epidemiological purposes. It is expected that the different purposes for these two data collections have an impact on data consistency, namely on it completeness, detail of information and other quality problems.

We expected to detect the type of variables that are better recorded in each information system, by calculating and comparing the quality of similar variables. We also expected to detect differences between both systems in registries of the same patients.

This paper can play an important role for better understanding the quality of the integration of departmental systems in the general hospital information system, pointing out some limitations about consistency and information flow. It is also our goal to suggest some recommendations and strategies to prevent data quality problems and to improve communication between central and department databases.

2 METHODS

This study has been developed at Hospital São João (HSJ), one of the biggest Central Hospitals of Portuguese NHS. It is also a teaching hospital where research teams develop and integrate numerous information systems at different hospital departments. The majority of information systems are integrated with the central system SONHO, therefore this study aims to evaluate, measure and compare the data quality between SONHO and the information systems available at HSJ. The study started by selecting the departmental information systems to be considered (e.g.: obstetrics, intensive care, pneumology and haematology). Then, common variables to the departmental and the central database will be studied to check consistency and other quality issues.

As referred before, this paper presents preliminary results of a comparative study; that is, the results presented focus over one of the available information systems.

The information system selected was ObsCare (VirtualCare), an application running on obstetric department to register and manage all obstetric episodes occurred. This system was integrated in HSJ obstetric department in 2004 and since then it is collecting daily data, namely from parturient and newborns.

In this study we used a simple method. We started with an individual analysis over each field at each table in each system. This first process aimed to evaluate the individual data quality in both systems. After this characterization and after understanding the individual problems, we have merged equal tables from both systems so the comparison field by field could start. The merging process was made based on patient's *sequential number*.

To the individual analysis we have selected all episodes registered in both systems. As SONHO has been collecting data since 1997 and ObsCare only started in 2004, the number of observations in each table will be greater in SONHO. Therefore, for the comparison, we will be analysing only the common registries in both systems, i.e., episodes from January 2004 until 20 of July 2011.

To run the study over these information systems, authorizations were granted by the obstetric department's director. The research team involved in this analysis consisted in informatics specialists, developers of the ObsCare system and statisticians.

The tools used in this study were Excel 2007 and SPSS 19.

3 RESULTS

In this section we will present the results of our preliminary study. We will start presenting the results of the individual analysis of data quality over each table.

Table 1 shows the results of the individual analysis over the episodes table of both systems.

Even with a larger set of data in the episodes table, in SONHO no data quality problems were detected. On the other hand the same table in ObsCare evidences some quality problems. The problem with more expression is the missing values in variable *administrative discharge date*. As showed in Table 1, all registries in this variable have missing value. After checking with the developers of ObsCare we understood that this variable is not filled in this system. The reason for this is simple, when a patient leaves the obstetric department the clinical discharge date is registered in ObsCare. The *administrative discharge date* is filled by the administration staff in SONHO when the patient leaves the hospital. This happens because this patient can be admitted in other departments after leaving the obstetric department and before leaving the hospital. The problem is that ObsCare does not receive any information from SONHO about the patient after he leaves the department. Thus the system does not know the patient's administrative discharge date. This is not a data quality problem but an integration problem.

Table 1: Data quality problems observed in episodes tables.

Data quality problems detected	ObsCare		SONHO	
	N	%	N	%
Episodes table				
Total registries examined	30,985	-	51,410	-
Missing admission date	65	0.21	-	-
Missing clinical discharge date	2,853	9.21	-	-
Missing administrative discharge date	30,985	100	-	-
Missing/Invalid admission responsible	2,006	6.47	-	-
Missing/Invalid discharge responsible	3,000	9.68	-	-

Nevertheless, some real quality problems were detected in other variables. 65 (0.21%) registries with no *admission date* were detected. It should not be possible to register any patient in the system without filling this variable. The same was observed with the *clinical discharge date*, 2 853 (9.21%) registries with missing values. Other detected problems were the missing or invalid values in variables related to the *admission responsible* and *discharge responsible*. In these variables a numeric code is registered identifying the doctor responsible for admission/discharge. Those cases, which are filled with 0, are considered invalid. Thus, we detected, in *admission* and *discharge responsible*, 2 006 (6.47%) and 3 000 (9.68%) missing/invalid values respectively. Once again it reveals that no mechanisms are used in ObsCare to validate or control this process. In these variables, zero or blank values should not be accepted.

In Table 2 we present the observed data quality problems in identification tables.

The most relevant result that we can extract form Table 2 is the missing values for the variable *patient number*. This is an important variable for the identification of the patient and both systems present a high percentage of missing values. Other variables like *contact* or *marital status* also present high number of missing values but in these cases they are not as important for the identification/notification of patients, nevertheless these are data quality problems that should not occur in these systems.

The missing values detected in the *post code* can be easily explained. In some cases administrative staff filled wrongly the post code as part of the *address* variable. However this problem should be avoided for a better quality of data for future analysis or usage.

In ObsCare we also detected some cases of missing values in *birth date*. There are few cases but the validation mechanisms should not let this happens.

In ObsCare, as we can see in Table 2, there are 2 cases of missing *gender*, 1 of missing *patient name*, 28 cases of missing *process number* and 4 of missing *address*. SONHO also has 41 missing values in the *address* variable. Their occurrence is marginal but can work as alerts for problems with the system for future versions.

Table 2: Data quality problems observed in identification tables.

Data quality problems detected	ObsCare		SONHO	
	N	%	N	%
Identification table				
Total registries examined	23,994	-	35,966	-
Missing patient number	3,390	14.1	5,111	14.2
Missing birth date	111	0.46	-	-
Missing post code	89	0.37	113	0.31
Missing contact (tel.)	10,919	45.5	5,772	16.1
Missing marital status	790	3.29	1,099	3.06
Missing gender	2	0.01	-	-
Missing address	4	0.02	41	0.11
Missing name	1	0.00	-	-
Missing process number	28	0.12	-	-

Other problems, not presented in Table 2 but that focused our attention, were some inconsistencies in several values. For example, if we are analysing only obstetric episodes and the identification table just register the identification of the parturient (female) all registries should have female as *gender*.

However, we have detected 15 registries of males in ObsCare and 14 in SONHO. Another inconsistence detected with the patient's *gender* is that this is a numerical variable registered in the database with 1 (Male) and 2 (Female), but we have detected in ObsCare 275 (1.15%) representations with 'F'. This is truly an inconsistence but in this case it is a database problem.

In *marital status* the possible values are: single, married, divorced, widow and cohabiting couples. We have detected 871 (3.6%) cases registered with 'Unknown', 115 (0.5%) with 'Other' and these are not possible values in the form field for this variable. In addition, 14 cases completely out of standard were detected.

Table 3 summarizes the data quality results for the newborn tables. Analysing the detected problems for apgar variables, it is evident the lack of registries for these variables in SONHO, as the missing values are 9 228 (25.2%) for *apgar1*, and 8 887 (24.27%) for *apagar5*. In ObsCare, only *apgar10* score has a high percentage of missing values revealing that the tenth minute measure is not as important as the other two measures.

Table 3: Data quality problems observed in newborn tables.

Data quality problems detected	ObsCare		SONHO	
	N	%	N	%
Newborn table				
Total registries examined	21,225	-	36,611	-
Missing delivery type description	305	1.44	-	-
Missing son inpatient number	-	-	88	0.24
Missing son sequential number	-	-	81	0.22
Missing Apgar1	58	0.27	9,228	25.2
Missing Apgar5	65	0.31	8,887	24.3
Missing Apgar10	11,885	56.0	-	-
Invalid delivery type	305	1.44	-	-
Invalid fetal presentation	305	1.44	-	-

Again in newborn tables, as we verified in identification tables, there are variables with invalid values. The *delivery type*, for instance, is a string variable with possible values: 'Eutocic', 'Forceps', 'Vacuum', 'Cesarean', 'At home', 'In Pré-hospital transportation' and 'Unknown'. We detected, in this variable, 305 (1.44%) registries with different representations than those listed. As a result those 305 registries have missing values in the *delivery type description* variable because the database does not have correspondence for these delivery types. The same happens with the variable *fetal presentation* where 305 registries with invalid values were detected. In addition, we have detected other problems such as 1 registry with 0 *weight* in ObsCare and 2 registries in SONHO.

As the individual analysis of tables in both systems is complete the next phase is to compare registries between both systems. For this comparison missing values will be excluded. Before presenting the results of the comparison it is important to refer some differences between variable representations in both systems. For example the variable *fetal presentation* has in SONHO the possible values: 'T' (Transverse), 'C' (Cephalic) and 'P' (Pelvic) while in ObsCare instead of 'T' there's an 'E' for '*Espádua'*. The two terms have the same meaning but in a database architecture point of view the same values should be used in both systems.

The variable *gender* in the ObsCare newborn table is of string type with values 'F' and 'M' while in SONHO the same variable is numeric with values 2 and 1 respectively. A similar problem was detected in the variable *delivery type* and respective description. In ObsCare the *delivery type* is a string variable while in SONHO it is numeric and the possible values are different.

Even inside the same system there are different representations for the same variables. In ObsCare, the variable *gender* in the identification table is numeric while in newborn's table, as already referred, is a string variable.

For the comparison between identification tables, cases were merged based on their *sequential number*. During this process we have detected several registries in ObsCare with no correspondence in SONHO and vice-versa. At total 2 101 of these registries were detected in ObsCare and 142 in SONHO. These cases were also excluded from the comparison results presented in Table 4. As we can observe in Table 4 only 21 893 of the 23 994 registries from identification table in ObsCare were considered common in both tables.

In the common registries we detected 3 575 (16.33%) differences in the *contact* number registered in both systems. Differences were detected also in *patient number* and *process numbers* with 2 913 (13.31%) and 2 503 (11.43%) cases respectively.

The highest differences were detected in *address* and *marital status*. The differences between *marital status* can be partially explained due to the different possible values for this variable in both systems.

With the *address* it is not so simple to explain the detected differences without a specific tool to measure character string differences.

Other detected differences were observed in the variable *names*. With a lookup process to measure the differences it was possible to check that most cases differ because of a single surname. In SONHO the majority of the *names* in these cases appear with one more surname than in ObsCare's registries. We also detected some misspelling errors or differences in some letters of the *names*.

Table 4: Identification's tables comparison.

	N	%
Total common registries	21,893	-
Different contact (tel.)	3,575	16.3
Different patient number	2,913	13.3
Different process number	2,503	11.4
Different names	1,088	4.97
Different address	8,689	39.7
Different gender	1	0.00
Different birth date	189	0.86
Different post code	2,405	11.0
Different marital status	6,450	29.5

In Table 5 we can find a summary of the results for the comparison between newborns tables in both systems. As in the comparison of identification tables, we detected some cases where the merging process could not join both tables. In total, 850 ObsCare registries have no correspondence in SONHO and 210 newborn registries from SONHO have no correspondence in ObsCare.

Table 5: Newborn's tables comparison.

	N	%
Total common registries	20,375	-
Different delivery type	1,161	5.7
Different delivery type description	1,162	5.7
Different fetal presentation	563	2.8
Different birth date	1,594	7.8
Different weight	364	1.8
Different Apgar1	178	0.9
Different Apgar5	189	0.9
Different gender	152	0.7
Different live born (Y/N)	8	0.0

In Table 5 it is possible to observe that 1 594 (7.8%) cases have different *birth date* registered in both systems. In *delivery type* and respective descriptions, although the differences in possible values referred before, during this comparison process we have forced that similar delivery type values where considered the same. For example, we forced the correspondence between 'Eutocic' in ObsCare and SONHOS's values 'Eutocic – Twins', 'Eutocic – Pelvic' and 'Eutocic'. All other values were forced likewise when possible. So, the detected differences for these two variables are effective differences, more precisely 1 161 (5.7%) for *delivery type* and 1 162 (5.7%) for the respective description.

The same technique of forcing equalities was used for the variable *fetal presentation*, but in this case it was only necessary to force the 'E' in ObsCare to be the same as 'T' in SONHO, for the reasons already explained before. Even though we have detected 563 (2.8%) differences in *fetal presentation*. With not as much significance as the already mentioned differences, but with no less importance, there is the difference between *weights* with 364 (1.8%) cases, *apgar1* with 178 (0.9%) and *apgar5* with 189 (0.9%) cases. These are numeric values measured only once, so it is hard to understand the reasons why these values have differences in both systems. We also detected differences in the registries of the newborn's *gender* and in 8 cases the registries do not match in the variable *live birth*.

Next, we present the last comparison table with the results of the comparison between the episode tables from both systems.

Table 6: Episode's tables comparison.

	N	%
Total common registries	24,971	-
Different discharge date	1,509	6.0
Different admission date	1,182	4.7
Different admission responsible	4,198	16.8
Different discharge responsible	2,688	10.8

As we can observe in Table 6 there are many differences between date variables in both systems. The *discharge date* variable presents 1 509 (6.0%) cases of difference in registries and in *admission date* we have detected 1 182 (4.7%) differences. Also in the *admission* and *discharge responsible* we verified a high percentage of differences. These results show clearly some issues in the communication between the involved systems.

4 CONCLUSIONS

With the results presented in this paper it is clear that there are some issues needing improvement so the integration process can be as reliable and consistent as possible. At the end we think that these two systems work in an individual way and in fact there is no real integration between them. All registries are duplicated, i.e., each registry is introduced manually in both applications by different health professionals. That is a big concern in terms of data quality as this process can lead to different registries and even duplication of errors. This would be avoided if the communication between the systems was more effective reducing the source of errors.

By analysing the results of the individual quality of data produced by both systems, it is possible to understand that ObsCare need additional validation tools. In fact, there are tools implemented in this system but, as we observed in the presented results, they are not being as effective as desired. However it is patent that ObsCare, because of his purpose, has more detailed data, but not in a consistent and complete way. There is considerable amount of missing data, some variables have invalid values registered and, as we verified, there are different representations for the same variables.

The central system SONHO evidences less interest in collecting some specific variables as they are not as important for the system purpose. Nevertheless some detected data problems can be very useful to call the attention of the NHS so they can change the way data are collected, improving his completeness, consistence and detail.

Through the comparison, differences are clear between both systems. The differences were detected in every variable and table analysed. This proves that the integration failed as there is no really interaction. A better communication between both systems could conduct to more reliable information and save time in the introduction of data so that health professionals can have more time to be focused on patients and on research.

This is a preliminary study, and so all results collected and presented will be further explored during our future work. In the next steps of our research we will be working with developers to test and improve their validation tools and to implement an application to scan all data and check for these and other data quality problems. We would also like to extend this study to other departmental systems working at the HSJ central hospital.

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 modelling, cleansing and analysis (project PTDC/SAU – ESA /75660/ 2006).

REFERENCES

Arts, D. G., De Keizer, N. F. and Scheffer, G. J., 2002. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc*, 9, 600-11.

Cruz-Correia, R., Vieira-Marques, P., Ferreira, A., Oliveira-Palhares, E., Costa, P. and Costa-Pereira, A., 2006. Monitoring the integration of hospital information systems: How it may ensure and improve the quality of data. *Stud Health Technol Inform*, 121, 176-82.

Cruz-Correia, R. J., 2010. Implementation, monitoring and utilization of an integrated Hospital Information System--lessons from a case study. *Stud Health Technol Inform*, 160, 238-41.

Dismuke, C. E. and Sena, V., 1999. Has DRG payment influenced the technical efficiency and productivity of diagnostic technologies in Portuguese public hospitals? An empirical analysis using parametric and non-parametric methods. *Health Care Manag Sci*, 2, 107-16.

Freitas, A., Marques, B., Silva-Costa, T., Lopes, F., Garcia-Lema, I. and Costa-Pereira, A. Year. Data Quality issues in DRG databases. In: *26th PCS International Conference*, 2010a Munich.

Freitas, A., Silva-Costa, T., Marques, B. and Costa-Pereira, A. Year. Implications of data quality problems within hospital administrative databases. In: *12th mediterranean conference on medical and biological engineering and computing – medicon 2010*, 27-30 May 2010b Porto Carras, Chalkidiki, Greece.

Silva-Costa, T., Freitas, A., Jácome, J., Lopes, F. and Costa-Pereira, A. Year. A eficácia de uma ferramenta de validação na melhoria da qualidade de dados hospitalares. In: *CISTI - 2ª Conferência Ibérica de Sistemas e Tecnologias de Informação*, 21 a 23 de Junho 2007 Porto.

Silva-Costa, T., Marques, B. and Freitas, A. Year. Problemas de Qualidade de Dados em Bases de Dados de Internamentos Hospitalares. In: *5ª Conferência Ibérica de Sistemas e Tecnologias de Informação*, 16 a 19 de Junho 2010 Santiago de Compostela.

VirtualCare. VCOBS.GYN - ObsCare [Online]. Available: http://virtualcare.med.up.pt/index.php/en/Produtos/vcobsgyn-eng.html [Accessed].