Exploring Complexity in Health: An Interdisciplinary Systems Approach
A. Hoerbst et al. (Eds.)

doi:10.3233/978-1-61499-678-1-332

Assessment of Automatically Exported Clinical Data from a Hospital Information System for Clinical Research in Multiple Myeloma

Viviana TORRES[a,1], Mauricio CERDA[b], Petra KNAUP[c] and Martin LÖPPRICH[c,d]
[a] *Program of Master in Medical Informatics, Faculty of Medicine, Universidad de Chile, Santiago*
[b] *Program of Anatomy and Developmental Biology, Biomedical Neuroscience Institute (BNI), Center of Medical Informatics and Telemedicine (CIMT), ICBM, Faculty of Medicine, Universidad de Chile, Santiago*
[c] *Institute of Medical Biometry and Informatics, Heidelberg University, Germany*
[d] *Department of Internal Medicine, Division of Hematology/Oncology/Rheumatology, Heidelberg University Hospital, Germany*

Abstract. Purpose: An important part of the electronic information available in Hospital Information System (HIS) has the potential to be automatically exported to Electronic Data Capture (EDC) platforms for improving clinical research. This automation has the advantage of reducing manual data transcription, a time consuming and prone to errors process. However, quantitative evaluations of the process of exporting data from a HIS to an EDC system have not been reported extensively, in particular comparing with manual transcription. In this work an assessment to study the quality of an automatic export process, focused in laboratory data from a HIS is presented. Methods: Quality of the laboratory data was assessed in two types of processes: (1) a manual process of data transcription, and (2) an automatic process of data transference. The automatic transference was implemented as an Extract, Transform and Load (ETL) process. Then, a comparison was carried out between manual and automatic data collection methods. The criteria to measure data quality were correctness and completeness. Results: The manual process had a general error rate of 2.6% to 7.1%, obtaining the lowest error rate if data fields with a not clear definition were removed from the analysis ($p<10E-3$). In the case of automatic process, the general error rate was 1.9% to 12.1%, where lowest error rate is obtained when excluding information missing in the HIS but transcribed to the EDC from other physical sources. Conclusion: The automatic ETL process can be used to collect laboratory data for clinical research if data in the HIS as well as physical documentation not included in HIS, are identified previously and follows a standardized data collection protocol.

Keywords. electronic data capturing, automatic process, data collection, hospital information system

[1] Corresponding Author: vatorresm@gmail.com

1. Introduction

High data quality is critical in clinical research and it is ensured by established standards compliance [1]. Good Clinical Practice (GCP) is the most recognized standard to ensure the scientific quality and international acceptance of a clinical study [2]. GCP defines a series of research protocols focusing on the complete cycle of a clinical study, starting from the design, data collection, data entry, data verification and analysis [3]. For instance, GCP indicates that in all these stages the information collected has to be well documented [2].

Traditionally specialists in the area of clinical research have used paper-base case report forms (CRFs) to collect relevant data to carry out scientific studies [4, 5]. Nowadays, with the wide spread of information technologies, the use of electronic data capture (EDC) software has optimized the data collection process [6]. On the other hand, the important increase of clinical information in electronic format in hospital information systems (HIS) provides the opportunity to reuse the data for scientific research. This could facilitate data entry, reduce duplicated efforts and cost and optimize information use for research [2]. However, the transference of data from HIS to EDC systems remains a challenge. Certain limitation exists at organizational level that restricts the reuse of the data. Some of these limitations are the heterogeneous data in syntax and semantics, limited data logistic (use of different formats) and legal privacy limitations [7]. In this context assuring a high data quality extracted from a HIS for scientific purpose is an important problem that needs to be studied. Therefore, the main objective of this study is to assess whether the quality of an automatic process, reusing exported laboratory data would be appropriated for scientific purposes.

2. Methods

This study was performed in the Department of Multiple Myeloma of the University Hospital Heidelberg. This department uses a scientific database to enter clinical data for research purposes. Entered data is about demographic, diagnostic, laboratory and treatment data of patients with multiple myeloma. Currently the transfer of data from the HIS i.s.h.med (Industry Solution for Healthcare, a product of SAP), to the scientific database must be done manually by trained specialists, which is time consuming and prone to transcription errors.

A retrospective study was carried out to assess the quality of laboratory data. A subset of 8 laboratory parameters (Calcium, C-reactive protein, Haemoglobin, Lactate dehydrogenase, Creatinine, Thrombocytes, Albumin, Beta-2 microglobulin) was taken into account in 4 events of the disease (diagnosis, chemotherapy, transplantation and after 100 days of transplantation) from 162 patients belonging to the scientific database. This data was selected because it was already verified through a manual source data verification (SDV) process, defined as gold standard in this paper. The current manual process of transcription and the new automatic process were evaluated.

The criteria to measure data quality were correctness and completeness [9] (Equations 1 and 2). Results were dichotomized (0 1) as data with error and without error respectively, where ‘0’ represents a field ‘incomplete’ or ‘incorrect’ and ‘1’ represents ‘complete’ or ‘correct’ respectively, w_1 represents an attribute value in the information system and w_r the respective attribute value in the real world.

$$Q_{\text{Completeness}}(w) := \begin{cases} 0 & \text{if } w = \text{Null or w NULL (semantic) equivalent} \\ 1 & \text{else} \end{cases} \quad (1)$$

$$d_1(w_1, w_r) := \begin{cases} 0 & \text{if } w_1 \neq w_r \\ 1 & \text{else} \end{cases} \quad (2)$$

To carry out the assessment in the manual process of transcription, laboratory results from data pre-SDV with the gold standard were compared. All patients where the date of the event did not match with the date of the event in the gold standard were excluded from the sample.

Next, a script was developed in Talend Open Studio software to perform the entire ETL process collecting data from comma-separated value (CSV) files from the HIS. This script was developed to work with the structure provided by the HIS of i.s.h.med and with a specific MySQL database as destination.

Statistical analysis was performed using a chi square test as 2x2 table to analyze if processes are dependent of the error rate found. Test of given proportion was used to evaluate the statistical significance in the difference of error rate found in both processes of collecting data. The significance statistical level was fixed to 0.05.

3. Results

The error rate in the manual process was 7.1%. The incompleteness error was the most common type of error present in overall with 81.6% in comparison with the incorrectness error (18.4%). The errors found by event did not vary much with error rates from 5.4% to 8.9%. The parameter that presented highest error rate was C-reactive protein (CRP) with 38.8% (from this percentage 36.8% were due to incompleteness and only 2% due to incorrectness error). For the rest of the parameters the rates of errors were similar, with a rate of error from 1% (Creatinine) to 5.3% (Beta-2 microglobulin). The results change if the sample is analyzed without taking into consideration the CRP parameter. In this case the general error rate decreases to 2.6% (87/3399) being the difference statistically significant with a $p<10E-3$.

Second, the ETL process was implemented allowing an automatic import into the database, identifying incorrect and incomplete data at the point of entry, and also expanded the range of laboratory parameters captured possible to be analyzed (from 8 laboratory parameters in the manual process to 809 in the automatic process).

The general error rate in the automatic process was 12.1%, being also the incompleteness error (94.9%) the main cause. If data missing in the HIS, e.g. physical documents, they were excluded and the general error rate in the automatic process decreases to 1.9%.

To compare manual and automatic process of collecting data, the CRP parameter and data not present in the HIS were excluded from the analysis. The comparison shows that the manual process presented higher error rate (1.8%) than automatic process (0.18%). The evaluation of chi square test shows that the error rate is dependent of the type of process, e.g. the manual process influences in the higher error rate found ($p<10E-3$) and the given proportion test shows that the automatic process presented significant less error rate in comparison with manual process ($p<10E-3$).

4. Discussion

The general error rate in manual process of transcription of the data was relatively high with a rate of 7.1% of errors when compared to the average of 2.3% described in the literature for double entry method in a clinical research database [8].

The total of errors per event were similar varying in rates between 5.4% and 8.9%, being the most common incompleteness in all events. The parameter, which presented considerable most error, was CRP, with 38.8%. In the analysis of this parameter, the highest rate was due to incompleteness data. This could be explained by the fact that there was no consensus and no clear protocol when the CRP value is below the detection limit. In this case, the detection limit in Heidelberg laboratory is 1mg/L. Some training specialists entered the limit value of 1 when it is below the detection limit and others used the rule if CRP is below the detection limit, they leave the item blank ('missing value'). When the CRP parameter is excluded from the analysis, the general error rate decreases until 2.6% in the manual process of collecting data, which is rather similar to the rate described in the literature before.

Regards to the general error rate in the automatic process through the ETL tool developed were 12.1%. The events that presented higher rate of incompleteness were diagnosis and chemotherapy (29.9% and 12.7% respectively). At the time when the patient is included into the registry, most of them have already laboratory results from a physical document generated externally. This result is not included in the automatic export of the HIS and leads to an incompleteness error in the automatic process. In addition, for the event of chemotherapy, some patients start with this phase of the treatment immediately after diagnosis. Therefore, the specialists may take the laboratory results from the diagnosis event and classify the same result as well for the diagnosis as for the chemotherapy event.

The error rate of 12.1% decreased to 1.9% if physical documentation not included in the hospital information system is excluded, which is a low rate in comparison with 8% described in the literature [11] for automated collection process of structured data. Moreover, the error rate of 1.9% decreases to 0% when the data in the automatic process are compared directly with the HIS and not with the gold standard.

In this study we found up to 10.2% of unstructured data that cannot be analyzed by an automatic process. When compared to the literature we found that Bae et al. described a severe limitation with clinical data, specifically laboratory data that are present in free-text or scanned documents, in Electronic Health Records (EHR). He found 2.4% of unstructured laboratory data that leads in less complete data to be analyzed compared to the manual EHR import process [10]. Also, Arts et al [11] shows in his study 4% of data not available in electronic format, which is low in comparison with the 10.2% found in this study.

When the sample is adjusted by keeping only data that follows a standardized data collection protocol and included in HIS, the error rate is higher in the manual process than in the automatic process (1.8% and 0.18% respectively; $p<10E-3$).

Transcription time was not measured in this study, however it can be expected that an automatic ETL process significantly reduce the processing.

The results presented in this study show that the ETL tool can be used to collect laboratory data if data in the HIS as well as physical documentation not included are identified previously and following a standardized protocol of collecting data.

Acknowledgements

This study was supported by grant from DAAD (Deutscher Akademischer Austausch Dienst) of the Heidelberg University, Germany. We gratefully acknowledge to the staff of the Institute of Medical Biometry and Informatics, Heidelberg University, Germany. MC was partially funded by FONDECYT (3140447), BNI (P09-015-F), and VISUAL D (ACT10712).

References

[1] Köpcke F, Kraus S, Scholler A, Nau C, Schüttler J, Prokosch H-U, et al. Secondary use of routinely collected patient data in a clinical trial: an evaluation of the effects on patient recruitment and data acquisition. *Int J Med Inform*. 2013 Mar;**82**(3):185–92.
[2] CPMP/ICH. Good Clinical Practice. European Community 1996.
[3] Krishnankutty B, Naveen Kumar B, Moodahadu L, Bellary S. Data management in clinical research: An overview. *Indian Journal of Pharmacology*; 2012: **44**(2):168.
[4] Wagner G, Leiner F, Gaus W, Haux R, Knaup-Gregori P. Medical Data Management: A Practical Guide Springer New York; 2006.
[5] Nahm ML,Pieper CF, Cunningham MM.Quantifying Data Quality for Clinical Trials Using Electronic Data Capture; 2015: 3(8).
[6] El Fadly A, Rance B, Lucas N, Mead C, Chatellier G, Lastic P-Y, et al. Integrating clinical research with the Healthcare Enterprise: From the RE-USE project to the EHR4CR platform. *Journal of Biomedical Informatics*; 2011. **44**:S94–S102.
[7] Benson T. Principles of Health Interoperability HL7 and SNOMED. Springer; 2012.
[8] Goldberg S, Niemierko A, Turchin A. Analysis of Data Errors in Clinical Research Databases. *AMIA . Annual Symposium proceedings* / AMIA Symposium AMIA Symposium;2008:242–6.
[9] Hildebrand K, Gebauer M, Hinrichs H, Mielke M. Daten- und Informationsqualität: Auf dem Weg zur Information Excellence. 3a Ed. Germany: Springer-Verlag; 2015.
[10] Bae CJ, Griffith S, Fan Y, Dunphy C, Thompson N, Urchek J, et al. The Challenges of Data Quality Evaluation in a Joint Data Warehouse. eGEMs (Generating Evidence & Methods to improve patient outcomes); 2015. l3:1-12.
[11] Arts DGT. Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework. *Journal of the American Medical Informatics Association.* 2002:**9**(6):600–11.