2015 IEEE 28th International Symposium on Computer-Based Medical Systems

Data Quality in HL7 Messages – A Real Case Analysis

Ricardo Ferreira and Manuel E. Correia
CRACS-INESC TEC
DCC, Faculty of Science
University of Porto, Portugal
rjtf@inescporto.pt;mcc@dcc.fc.up.pt

Francisco Rocha-Gonçalves
IPO Porto
Porto, Portugal
fnrgoncalves@ipoporto.min-saude.pt

Ricardo Cruz-Correia
CINTESIS
Faculty Of Medicine
University of Porto, Portugal
rcorreia@med.up.pt

***Abstract*—The development of eHealth technologies over the last few years has been pushing healthcare institutions to evolve their own infrastructures. Along with this evolution, critical systems now need to use communication standards such as HL7 or DICOM in order to exchange information in a more meaningful and efficient way. However, healthcare institutions often experience complications when different systems communicate directly even when using communication standards. We aim to assess the quality of the data present in HL7 messages exchanged between different critical systems in a large healthcare facility and therefore propose an integration infrastructure that allows a real time and centralized way to manage, route and monitor the integration flows between various systems.**

***Keywords*-Health Level Seven; Healthcare Integration; Data Quality;**

I. Introduction

Healthcare institutions have been increasingly challenged to reduce internal costs while still being able to maintain or even improve the level of care provided to each patient [1], [2]. At the same time, the development of eHealth products made by private companies and its establishment as a solid area of academic research, led healthcare providers to adopt many of these technologies as mature productions systems playing critical roles in the hospital daily activities.

Associated with the deployment of many different systems, healthcare facilities are often left with increasingly complex challenges due to the fact that each disparate system often needs to fetch and exchange information with other systems. This gains even more relevance if we take into account that many of the core systems being currently deployed at current healthcare institutions are actually developed and supported by many different software vendors [3].

This exchange of information often leads to the creation of fragmented and duplicated versions of the same information. Besides the obvious drawback and additional costs associated with having to store multiple pieces of information at different locations, such fragmentation also means that healthcare institutions will often have the same information stored under different formats in different systems, which incurs the extra complexity of maintaining all those versions up to date and consistent with each other. Another cause for data fragmentation within healthcare institutions is often the existence of multiple data entry points, where the format into which the data ends up being stored depends upon the systems where it is being collected. As such, core information such as the patients Electronic Health Records (EHRs) ends up being fragmented throughout multiple software applications, each one storing it using its own structures and formats [3], [4]. In order to tackle these challenges, healthcare facilities adopted the Health Level Seven (HL7) messaging standard as the means by which disparate systems can interconnect and exchange meaningful information.

To assess the extent to which the quality of data can be affected by Information Systems (ISs) vendors lack of transparency we have developed and installed an integration infrastructure called Integrated Routing Audit for HL7 (IRA) that allows us, by directly collecting data from the network, to centrally manage and assess in near real time the quality of the data within the HL7 messages that are being exchanged between different ISs in a large healthcare facility. During November 2014 we have collected and analysed a total of 1,207,519 HL7 messages, 94% of which had some type of syntax/semantic problem. The most common problems consist on deprecated fields or data outside the correct field. However the most serious problems encountered were non-documented custom structured values defined by software providers which endangers future system integrations and promotes further vendor lock-ins. Although the existing system integrations found at the healthcare facility fulfill their intended goal and allow data to be exchanged among ISs, system managers were unaware that their systems rely on the use of incomplete or non standardize HL7 messages. This decreases the institution's ability to promote further system integrations and even makes it increasingly difficult to swap end systems in a transparent way.

In this paper, we present an analysis of the quality of the data present in HL7 messages being exchanged within a large healthcare institution in Porto, Portugal. Based on our findings we propose a framework architecture capable of empowering healthcare facilities with the means to analyze in depth the data quality of HL7 messages being exchanged within their critical systems integrations. Based on that same architecture, our system is also capable of providing the means by which one can develop a real time alert system

2372-9198/15 $3.00 © 2015 IEEE
DOI 10.1109/CBMS.2015.30

based on the actual contents of the HL7 messages being exchanged.

The remainder of this paper is organized as follows. Section II presents a background review on healthcare interoperability and the HL7 standard as well as recent work on data quality assessment in healthcare integration scenarios. Section III includes an analysis on the collected HL7 Section IV presents and discusses a system infrastructure to monitor and assess healthcare integrations present at a given institution. Lastly Section V concludes the paper with some final considerations on the presented findings and potential solutions for healthcare integration monitoring and data quality assessments.

II. Background

A. Messaging Standard

The main goal of the HL7 standard is to allow healthcare institutions to interconnect several disparate systems and therefore share meaningful medical information between heterogeneous systems [5] which, to this day, still remains one of most popular and widely used approaches [6].

However, one major drawback when using the HL7 standard is related to its need for custom parsing and handling tools, which further contributes to an increase in complexity of otherwise simple applications. For this reason in 2011 Fast Healthcare Interoperability Resources (FHIR) [7] began being planned. The main goal in developing this standard is to take advantage of typical Internet standards and protocols like Hypertext Transfer Protocol (HTTP) and Representational State Transfer (RESTful) in order to provide eHealth systems with the capability of having healthcare resources and processes described in a much simpler and standardized way by using data structures such as Extensible Markup Language (XML) or JSON.

B. Data quality assessment and monitoring

The quality assessment of the data being produced and managed by healthcare information systems is a recurring topic in academic research. Many studies focus their attention on the quality of data present in the patients' EHRs and try to evaluate it under different dimensions such as *data accuracy*, *timeliness*, *comparability*, *usability* and *relevance*. In order to evaluate and compare data quality under these different dimensions, a set of framework guidelines have been developed and proposed to consistently report and compare data quality assessment findings [8], [9].

Mphatswe et al. in [10] studied the data quality present in routine health information stored in the South Africa district health information system and concluded that the data quality improvements observed in such repositories were directly related with the training of healthcare workers, monthly data reviews and regular data audits. Botsis et al. in [11] analysed EHRs from the Columbia University Medical Center and claim that lack of data completeness can often happen mainly for two reasons (a) *data fragmentation* caused by patient treatments being made at different healthcare facilities and (b) *lack of contextual information* caused by poor medical documentation.

Related to data monitoring tools, the authors developed a monitoring system called IRA [12] capable of passively collecting from the network, the HL7 messages exchanged between various critical systems present in an healthcare facility. With IRA, the authors have been able to produce monitoring graphical dashboards, filled with meaningful healthcare production metrics and statistics such as the historical number of laboratory orders, appointments or patient discharges, that are proving to be useful elements supporting high level decision making for the Hospital administrators.

III. HL7 Data Quality

A. Methods

We collected HL7 messages exchanged between critical production systems present at a large oncological healthcare facility in Porto, Portugal, where a daily average of 61.500 HL7 messages are being produced.

Our study focuses on the quality of the data present on a set of messages collected during November 2014. During the collection phase, a total number of 1.207.519 HL7 messages were exchanged between different systems. The data used in this study was collected using the author's previously developed IRA monitoring tool which allowed us to passively collect all the HL7 messages produced and exchanged within the institution.

Among the collected HL7 messages, our analysis focused on assessing the following messages aspects as well as identify potential future integration challenges that can arise due to standard violations or lack of data quality:

- **Standard Inconsistencies:** Messages that do not respect the official HL7 standard;
- **Content Issues:** Faulty data such as corrupted or badly encoded characters;
- **System Differences:** Messages with the same HL7 type and trigger event that have a different format depending on the source application;

B. Results Discussion

Table I presents some of the main issues found in the HL7 messages collected by IRA during November 2014. The table aggregates the information based on the source, destination and type of the HL7 messages, stating for each entry the total number of messages received during the month. We then show for each HL7 segment the problems we found during our analysis and associate each issue with a level of occurrence stating how often the respective problems occurs in the overall messages.

By looking at Table I we can observe that for each HL7 message type, independently of its origin and destination, there is at least one segment that has some type of problem

Table I: HL7 Analysis Results

HL7 Message Source	HL7 Message Destination	HL7 Message Type	Monthly Total Messages	HL7 Message Segment	Issue	Level of Occurrence
HIS	Radiology Radioteraphy	ADT	15,964	PID	Use of deprecated PID-4 field (alternate ID)	Always
					Values outside correct field (patient middle name)	Mostly
					Values outside correct sub-field (patient address)	Always
					Custom structure used in PID-23 (birth place)	Always
					Special characters unescaped	Rarely
					Unknown character encoding	Rarely
	Laboratories	OML	227,336	PV1	Values outside correct sub-field (visit location)	Always
					Unknown character encoding	Rarely
				ORC	Values outside correct field (ordering provider suffix)	Always
				OBR	Values outside correct field (ordering provider suffix)	Always
Radioterapy	HIS	SIU	118,891	SCH	Values outside correct field (attending doctor)	Mostly
					Use of wrong field to carry dates (SCH-2 filler appointment ID)	Mostly
					Use of custom unspecified field separator	Always
					Appointment ID exceeds size limits	Always
Laboratory	HIS	OML	392,698	OBX	Field with invalid values (OBX-2 value type)	Always
					Incorrect method to encapsulate PDF file	Always
				OBR	Unknown character encoding	Rarely
				PID	Values outside correct field (patient middle name)	Mostly
		ORU	160,256	PID	Values outside correct field (patient middle name)	Mostly
		ORL	147,468	ORC	Values outside correct field (ordering provider suffix)	Always

that always occurs. According to our analysis, we can state that of all the HL7 messages exchanged during November 2014, a total number of 1.137.414 messages possessed some type of problem, which represents a total of 94.2% of all the HL7 traffic generated by the healthcare institution.

We found out that there are systems that often resort to deprecated HL7 fields in order to place information that should be specified in different fields or formats. Also related to the placing of information in the correct fields, HL7 sub-fields are not correctly used, which leads to data being incorrectly structured.

Finally, in our analyses we often found data being exchanged in custom structures and formats outside of the HL7 standard. Examples of such issues were detected in HL7 PV1 segments where dates were being aggregated with the visit identification number or ADT messages containing numerical codes to define the birth place of a given patient that should otherwise be represented as a string of characters.

IV. Monitoring System Architecture

A. Integration Monitoring and Data Analysis

In order to create a monitoring mechanism capable of monitoring the issues presented back in Section III, we propose taking advantage of the author's previously developed IRA system and extend its current functionalities to create an automated tool capable of analysing in real-time the quality of the data present in the HL7 messages. Since HL7 messages have a well defined standard, we can use the integration engine at the core of the IRA infrastructure to create a set of channels capable of analysing the contents of each message and trigger alerts if the message standard isn't being correctly used.

By feeding the IRA system with messages exchanged between integrated systems we allow the creation of much more accurate and useful monitoring mechanisms to the Information Technlogies (IT) services which can be based on the HL7 messages that carry some type of control or error codes used by applications to assess if a certain procedure was actually successful or well interpreted by the end system. This particular alert system based on the contents of each HL7 message presents a great advantage to healthcare IT teams since it would allow them to be informed in a real-time manner of existing problems in critical systems which are otherwise considered as black boxes that only software providers can access. One of the main benefits from using the IRA infrastructure to monitor system integrations is related to its ability to gather the required data. Since HL7 messages are collected directly from Transmission Control Protocol (TCP) packets traversing the network, healthcare facilities are lifted from the burden of having to contact software vendors in order to obtain the exchanged integration messages. The gathered data, besides being used to produce useful data quality statistics could also be used to debug critical systems that may experience communication challenges. Also, from a knowledge gain point of view, the usage of the IRA infrastructure as a central hub and its ability to log each collected message also allows IT services to gain a more meaningful knowledge of any integration problems that may occur, therefore allowing IT employees to discuss with the software vendors potential integration fixes on production systems in a more informed way.

V. Conclusion

Healthcare facilities are nowadays composed by a multitude of different information systems each one playing

a critical role in the daily hospital activities. Due to being developed by different software providers, those same information systems are subject to severe communication challenges. The HL7 standard was created precisely in order to respond to these challenges and provide the means to which different systems can exchange meaningful healthcare information.

We analysed HL7 messages exchanged in an healthcare institution and presented a set of issues related to the contents of the messages and the application of the standard itself. Related to the actual extraction of HL7 messages directly from the network by sniffing TCP traffic, one might argue on the overall security state of the institutions' communications. However, while in fact there's no actual communication encryption being applied, all core network traffic is segmented in its own Virtual Local Area Network (VLAN) which in turn can only be accessed on certain network switches stored in secure data centers.

We believe that the presented issues in Table I associated with point-to-point infrastructure greatly diminish the purpose of healthcare interoperability and may lead to complex integration challenges that can't easily be solved without the software providers intervention. In that scenario, having a good understanding of each different integration protocol present in the healthcare infrastructure along with an efficient integration monitoring tool can greatly enhance the institution's response to a given problem. According to a more detailed analysis of the of the HL7 message contents, we can conclude that the standard is not being correctly used. The usage of customized structures to transport data to which the standard already dictates a specific field adds some drawbacks and inconsistencies that could prove to be fatal for future integrations. Namely, it forces the end system to create additional parsing mechanisms for the received messages in order to extract information that otherwise should be readily available at a specific HL7 field. By using this non-standard structures that only few systems understand, we limit the scope of healthcare interoperability and therefore lose the ability to swap end systems in a more "plug-and-play" like way. By analysing the PID segments of the HL7 messages, we often found critical data such as contact information or patient addresses being placed in different sub-fields depending on the source system that generated the HL7 message. This type of flexibility provided by the HL7 standard can also contribute to limiting healthcare interoperability since it allows the same type of information to be exchanged in different correct ways.

We have described an architecture capable of passively monitor healthcare integrations without having to change critical healthcare production systems. Associated with the proposed infrastructure, the creation of specific mechanisms capable of analysing the contents of each HL7 message paves the way for the development of efficient alert systems for the interoperability infrastructure.

ACKNOWLEDGEMENTS

This work is financed by the FCT Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project UID/EEA/50014/2013.

REFERENCES

[1] G. L. Kreps and L. Neuhauser, "New directions in ehealth communication: Opportunities and challenges," *Patient Education and Counseling*, vol. 78, no. 3, pp. 329 – 336, 2010, changing Patient Education. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0738399110000224

[2] J. A. Blaya, H. S. Fraser, and B. Holt, "E-health technologies show promise in developing countries," *Health Affairs*, vol. 29, no. 2, pp. 244–251, 2010.

[3] R. Fox, R. Sahay, and M. Hauswirth, "Ppepr for enterprise healthcare integration," in *Electronic Healthcare*. Springer, 2009, pp. 130–137.

[4] S. Alshawi, F. Missi, and T. Eldabi, "Healthcare information management: the integration of patients' data," *Logistics Information Management*, vol. 16, no. 3/4, pp. 286–295, 2003.

[5] M. Eichelberg, T. Aden, J. Riesmeier, A. Dogac, and G. B. Laleci, "A survey and analysis of electronic healthcare record standards," *ACM Comput. Surv.*, vol. 37, no. 4, pp. 277–315, Dec. 2005. [Online]. Available: http://doi.acm.org/10.1145/1118890.1118891

[6] P. De Meo, G. Quattrone, and D. Ursino, "Integration of the hl7 standard in a multiagent system to support personalized access to e-health services," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, no. 8, pp. 1244–1260, Aug 2011.

[7] D. Bender and K. Sartipi, "Hl7 fhir: An agile and restful approach to healthcare information exchange," in *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*. IEEE, 2013, pp. 326–331.

[8] M. G. Kahn, M. A. Raebel, J. M. Glanz, K. Riedlinger, and J. F. Steiner, "A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research," *Medical care*, vol. 50, 2012.

[9] C. I. for Health Information, "The CIHI Data Quality Framework," http://www.cihi.ca/CIHI-ext-portal/pdf/internet/DATA_QUALITY_FRAMEWORK_2009_EN, 2009, [Online; accessed 2015/01/25].

[10] W. Mphatswe, K. Mate, B. Bennett, H. Ngidi, J. Reddy, P. Barker, and N. Rollins, "Improving public health information: a data quality intervention in kwazulu-natal, south africa," *Bulletin of the World Health Organization*, vol. 90, no. 3, pp. 176–182, 2012.

[11] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, "Secondary use of ehr: data quality issues and informatics opportunities," *AMIA summits on translational science proceedings*, vol. 2010, p. 1, 2010.

[12] R. Ferreira, M. E. Correia, F. Rocha-Gonçalves, and R. Cruz-Correia, "Visualization of passively extracted hl7 production metrics."