The Information Quality Triangle: A methodology to assess Clinical Information quality

Rémy Choquet[a], Samiha Qouiyd[a], David Ouagne[a], Emilie Pasche[b], Christel Daniel[a,c], Omar Boussaïd[d], Marie-Christine Jaulent[a]

[a]*INSERM, UMR_S 872, Eq. 20, Paris, F-75006 France; Université Pierre et Marie Curie, Paris, F-75006 France*
[b]*SIM, University Hospitals of Geneva, Geneva, Switzerland*
[c]*AP-HP, Hôpital George Pompidou, Département d'Informatique Hospitalière, Paris, F-75015 France*
[d]*Laboratory ERIC, Université de Lyon 2, Bron, F-69676 France*

Abstract

Building qualitative clinical decision support or monitoring based on information stored in clinical information (or EHR) systems cannot be done without assessing and controlling information quality. Numerous works have introduced methods and measures to qualify and enhance data, information models and terminologies quality. This paper introduces an approach based on an Information Quality Triangle that aims at providing a generic framework to help in characterizing quality measures and methods in the context of the integration of EHR data in a clinical datawarehouse. We have successfully experimented the proposed approach at the HEGP hospital in France, as part of the DebugIT EU FP7 project.

Keywords:

Data quality, Medical informatics computing, Hospital information systems, Total quality management, Information storage and retrieval, Database management systems

Introduction

Clinical Information Systems (CIS) are built to record operational data about the patient, to record her/his pathway through the health institute and everything that relates to her/his care. CIS includes Electronic Healthcare Records (EHR) data and other sources, such as laboratory systems data. Besides, Clinical Data Warehouses (CDW) are built to aggregate a great amount of data reliable for healthcare decision-making: decision support, monitoring, alert or data mining [1]. Today, building CDWs [2] is the right opportunity for health institutions to enhance the quality of their information systems.

CIS and CDW are different entities since the requirements of clinicians do not necessarily meet the ones of research physicians or organizations. In many health institutes, one issue is to feed automatically CDWs with heterogeneous data from CIS. Since CIS are not built to specifically structure and store clean aggregated data, quality is often not accurate enough in CDWs, which sometimes lead to wrong decisions [3].

We believe it is necessary to assess data quality within the CIS prior to its storage and use within a CDW. Therefore we propose a methodology to assess data quality of a CIS, then we introduce a framework and tools to control and enhance the information quality based on three dimensions: *concepts*, *terms* and *objects*.

This work takes part in the European DebugIT[1] project [4,5] which goal is to build a technical and semantic information technology platform able to share heterogeneous clinical data sets from different hospitals for the monitoring and the control of infectious diseases and antimicrobial resistances in Europe. The Georges Pompidou European Hospital in Paris (HEGP) is one of DebugIT's partners; a dedicated CDW named Transmed has been developed locally for this project. This CDW is fed with medical data including antibiotherapy and antibiogram data, recorded over the last decade within the HEGP EHR system.

In the following, we present an Information Quality Model and a methodology that we experimented to improve information quality in the context of importing HEGP operational EHR data into Transmed. Results are then presented and discussed.

Background

Data volumes have been growing massively along with inconsistencies and erroneous data. Data quality started to be studied in the late 60's by statisticians [6]. Then, at the beginning of the 90's, computer scientists considered the issue of defining, measuring and improving data quality. ISO defines quality as "the totality of features and characteristics of an entity that bears on its ability to satisfy stated and implied needs" (ISO 8402-1986, Quality Vocabulary). Wang [7] proposes to define a piece of data as of good quality if it matches the intended use in its context. However, that concept being broad enough to be an axiom, it is not narrow enough to be able to characterize precisely data quality. Redman [7] has proposed to specify that axiom into four dimensions: *accuracy*, *perfection*, *freshness*

---

[1] Detecting and Eliminating Bacteria UsinG Information Technology

and *uniformity*. Other quality factors were introduced in order to assess data quality criteria depending on semantic [8, 9], process [9] or goal [10, 9]. Wang [7] also introduces a Total Data Quality Management (TDQM) methodology based on Deming wheel[2] that defines a quality improvement process. Likewise, the data warehouse communities have introduced various methods to measure, enhance and monitor the quality of data in CDWs [11, 12].

Besides the evaluation of the quality of data, many frameworks were introduced for evaluating information models quality [13, 14]. However, it remains difficult to assess quality of information models only with metrics [15]. Moody *and al.* proposed eight quality factors as metrics candidates. They also introduced subjective measurements of the information model quality by grading from 0 to 5 each of the 7 factors known as: *correctness, implementability, completeness, understandability, integration, flexibility and simplicity.*

In healthcare, data quality works have also been studied. A recognized need is growing for a secondary use of qualitative electronic health records for research (epidemiologic, public health). Nevertheless, it remains challenging to rely on good quality data [16]. Causes of insufficient data quality in medical records have been classified between systematic or random errors types [17] at different stages of the recording process. Kerr [18] introduced a framework to measure data quality based on the CIHI[3] recommendations. It led to setup a framework composed of 69 quality criteria grouped into 24 quality characteristics that groups into 6 quality dimensions: *accuracy*, *timeliness*, *comparability*, *usability*, *relevance* and *privacy* & *security*. These works propose methods and measures to assess data quality, yet we believe they have been essentially focused on data accuracy.

Beside this, the healthcare domain has built over years reference knowledge sources that could help in data quality assessment, namely, the standardized information models and domain terminologies or ontologies [19]. Reference terminologies have gained general acceptance over the research community [20], as they are key resources to interoperability and decision support. In addition to reference terminologies, "interface terminologies" are defined as "systematic collections of clinically oriented phrases aggregated to support clinician's entry of patient information in computer programs". Methods to compare the quality of interface and reference terminologies were introduced in [21].

Materials and Methods

Information Quality model

The clinical information contained within the CIS can be defined through 3 main dimensions: 1) data, or instances of real world objects, are physically stored information into CIS data stores, 2) information models are representations of concepts or relationships (among other properties) used to organize and structure information and 3) terminologies store referential data in various forms: terminology (list of terms), thesaurus (index and synonyms), classification (with generic relationships) or vocabulary (with definitions). We propose to classify the quality measures proposed in the literature, given those three dimensions defined as *objects*, *concepts*, and *terms*.
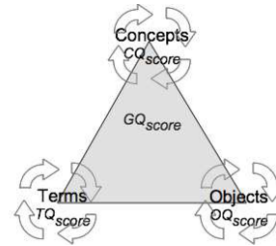

Figure 1- The Information Quality Triangle

The three information quality dimensions are the vertexes of our Information Quality Triangle (IQT) depicted in figure 1. For each vertex, there are methods and scores to measure information quality. Each score is then aggregated into a global score that would be defined as an information source score prior to data integration.

Material

The main data source in our experiment is the EHR system of HEGP. It stores 10 years of EHR data in various domains. We have focused our work on data domains that concerns infectious diseases, and particularly the laboratory results related to antibiotic resistance tests and antibiotic prescriptions.

The EHR dataset is composed of a volume of 1 200 000 patients, 1 600 000 admissions, 3 200 000 antibiograms, and 24 000 drugs/substances.

Most of the *object* quality criteria are measured using Talend© Open Studio[4] open source software as well as developed stored procedures in SQL[5].

Our domain is modelized with the help of the 6 following HL7 information models of the January 2009 version of the HL7 ballot:

- A_Encounter universal (COCT_RM010000UV01),
- Result Event (POLB_RM004000UV01),
- Composite Order (POOR_RM200999UV),
- Common Observation (POOB_RM410000UV),
- Adverse Reaction (REPC_RM000022UV),
- BillableClinicalService Encounter (COCT_RM290004UV06).

This covers the conceptual scope of the DebugIT project. The information model includes 61 classes and 262 properties.

To standardize the EHR vocabulary, we have first focused on two consensual resources in use within DebugIT: 1) ATC: The WHO drugs and substances international classification and 2) NEWT: organisms taxonomy database.

Perl routines have been developed at the University Hospitals of Geneva (HUG) to map free text terms to terminologies' entities. The drug names mapping is performed in several steps, until a successful candidate is found, consisting mainly in removing letters (the French version of the drug names often adds a final "e") or part of the term. For instance, the term *ac.fus* is not found by exact matching. When searching only for the second part (*fus*), the system returns several possible answers, but only one is an antibiotic: *fusidic acid*. The bacteria names mapping follows a similar approach. Moreover, when no candidate is found for a species, we attribute the parent taxon, the gender, to this term.

Quality method

We applied the TDQM 4 steps approach to score quality of each vertex [7]. They can be grouped into two functions, assessment (*audit* and *qualification*) on the one hand and enhancement (*standardization* and *surveillance*) on the other hand.

The *audit* consists in scoring the source for each vertex using defined criteria. Table 1 reports the criteria used at each vertex to build the overall score of the IQT. Each object quality score is composed of various criteria scores measured using computerized algorithms. Concept quality score is based on the subjective rating of information model quality proposed in [16]. Terminology score measurement is a statistical distance between the reference and the interface terminologies.

Table 1 – Criteria and Methods for auditing the data at each vertex

Vertex	Criteria	Method
Concept	Domain	Subjective Rating of Data Model Quality
Objects	Completeness	Total number of records filled compared to total number of records
Objects	Accuracy	Data format
Objects	Uniqueness	An algorithm that searches for uniqueness
Objects	Consistency	An algorithm to check consistency, for example check if Start Date of Prescription < End Date
Terms	Consistency	Distance to reference terminologies

The criteria are measured according to a reference for each of those dimensions. From the *object* dimension level, the reference is a set of rules (for example the date of death is after the date of birth). From the *concept* dimension, we state as a reference the HL7 version 3 information models specialized from the reference information model (RIM)[6], which proposes a conceptual representation for electronic health records. As for the *terms*, we use as reference the NEWT and WHO-ATC standards in order to normalize the terms of our domain of study.

The *qualifying* process aims at scoring the source information based on the IQT and the measures of the *audit* phase. Criteria scores are precise and accurate, though the qualifying of a source of information can be quite subjective. We use grades to score each IQT vertex of our information source, varying from A to D. For each quality domain, we made the average percentage of every criterion and split them every 25% to the corresponding grade. The meaning of the global scores can be qualified as following:

- A: The information quality is excellent. The information source carries enough semantics and organization to be queried without needs to be adapted.
- B: The information quality is good though it requires some work to improve one of the vertexes of the IQT.
- C: The information source quality is narrow. It could be usable only after some consequent work to improve information quality.
- D: The information quality of the source is low. The time and/or effort to improve it would be too high to consider this information source as a potential source for a CDW project.

Then, we enhance the quality within the *standardization* phase and we make sure the quality is controlled over time in the *surveillance* phase that helps to ensure the information is controlled over time within the destination CDW.

Results

For each step of the methodology, we present the results of our experimentation for the three vertexes of IQT.

Audit

For the *objects*, Table 2 shows result samples for three criteria on a limited number of objects.

Table 2 – Object Criteria Scores

Object	Criterion	Score	Comments
Discharge Date	Completeness	69,9%	Helps calculating the length of stay
DrugUCD	Consistency	75,6%	The UCD is a French classification code
Patient ID	Uniqueness	100%	The patient ID has unique values

As for the *terms*, Table 3 shows a sample result obtained against NEWT taxonomy for the Bacteria names, and against a local referential for sample location and type built by a medi-

---

[6] The RIM is a pictorial representation of the HL7 clinical data domains. See (http://www.hl7.org/implement/standards/rim.cfm)

cal expert. The measured distance is a percentage of correct terms within the EHR system compared to the standardized terminology.

Table 3 – Terminology Distance Scores

Terminology (Standard Referential)	Distance to Standard Referential
Bacteria Names (NEWT)	85,53%
Sample Location (local)	93,11%
Sample Type (local)	36,77%

For the *concepts*, we obtained the scores shown on Figure 2 with the help of a domain expert and a data modeler. Each axis represents the empirical evaluation of the source information model rating from 1 (poor) to 5 (excellent). Many examples illustrate the weaknesses of the EHR information model compared to HL7 version information models. In the field of lab results, the HL7 version 3 model "Result Event" (POLB_RM004000UV01) proposes a specific class to manage clusters of tests to be made on a given derived specimen (aliquote), which allows better management of the microbiological lab test results compared to the implementation done in the HEGP CIS. In the field of medication order, from a vocabulary perspective, a table named C_SPECIALITE records in fact drug prescriptions. A strength of the EHR information model is that it integrates a repository of shared data elements that are linked to national and local referential terminologies. This helps for integrating the EHR into the CIS.
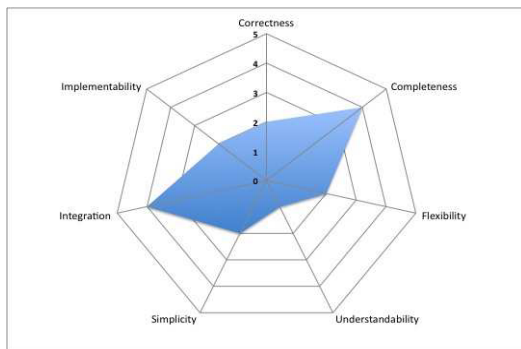

Figure 2- The EHR IM Quality subjective rating

Qualifying

The overall Information Quality Triangle of our information source for the restricted domain chosen is shown in Table 4.

This qualification phase can be considered as the validation method to our IQT. Each score is calculated given the scores found at each vertex. For example, the average of scores of the *concept* vertex being 2.42, the score is C. The *global* score of our source EHR system reflects its overall quality in the context of our evaluation.

Table 4 – Source EHR system scores

Vertex	Score
Object	C
Term	B
Concept	C
Global	**C**

Standardization

The standardization process was made during the loading of our clinical data warehouse for research. It was made at two different levels. At a first stage, we built an information quality firewall between the EHR and the CDW. Using the open source software Talend© Open Studio, we setup procedures to standardize the objects using a CTS[7] based information model. We built ETL scripts to correct objects based on a dictionary of 'dirty' and preferred terms stored into the CTS information model. At a second stage, we implemented two PERL scripts in order to map the local terminology preferred term with the standardized term. Up to 76% of the drug names present in the CDW were normalized successfully. Concerning pathogen normalization using NEWT, 99% of the pathogen names identified by an antibiogram were mapped to a NEWT term, which was manually validated.

The CDW physical information model was standardized using HL7 version 3 information models. The OMDF[8] platform was used to specialize the HL7-based conceptual models and generate the MySQL scripts in order to implement the CDW HL7-based information model corresponding to our domain.

Monitoring

We setup the necessary scripts to monitor the future load of data using Talend© Open Studio. We also setup the necessary alert routines that notifies of any unknown new concept loaded into the CDW in order to keep controlled clinical data warehouse information.

Discussion and Conclusion

The contribution of this work is three fold. First, this work enabled the HEGP hospital to provide a first attempt of measuring the distance between standardized information models and reference terminologies against its CIS after 10 years of production (it could also result in enhancing the information quality of their CIS). Secondly, it enhanced the quality of information within the CDW, which allowed building pertinent and coherent monitoring trends illustrating antibiotic resistance profiles over 10 years of data. Finally, it enabled the HEGP hospital to interoperate with other health institutes within the DebugIT project. Controlled vocabularies are a necessity to share data across Europe.

---

[7] Common Terminology Services: HL7 specifications for managing terminologies within clinical information systems
[8] Open Medical Development Framework: https://gforge.spim.jussieu.fr/projects/omdf-hl7/

We have introduced another method based on the literature work to assess information quality of electronic health records that takes into account a 'formal' framework to measure information quality, given the health care domain specificities: the IQT. We think this methodology abstract enough to be generalized. For instance, the use of an ontology could be preferred to the use of an information model to conceptualize a domain. In that case the conceptual vertex of the triangle could be an ontology.

We have successfully experimented this methodology in the context of the European project DebugIT. Not all of the experiment is automatized since it still requires the help of medical experts, but it does not invalidate the proposed methodology. The translational clinical data warehouse we built contains controlled objects, terminologies and concepts. We believe it is a first step to interoperability that cannot be avoided in the healthcare domain. It would be interesting to validate our approach on other EHR systems in order to better evaluate it. We also would like to investigate the use of our quality scores within a clinical decision support system.

Acknowledgments

References

[1] Gainer et al. Using the i2b2 Hive for Clinical Discovery: an Example. AMIA Annu Symp Proc (2007)

[2] Wisniewski MF, Kieszkowski P, Zagorski BM, Trick WE, Sommers and M Weinstein RA. Development of a clinical data warehouse for hospital infection control. JAMIA. 2003 vol. 10 (5) pp. 454-462.

[3] Hasan S and Padman R. Analyzing the Effect of Data Quality on the Accuracy of Clinical Decision Support Systems: A Computer Simulation Approach. AMIA Annual Symposium Proceedings, 2006 vol. 2006 pp. 324.

[4] Lovis C *et al*. DebugIT for patient safety - improving the treatment with antibiotics through multimedia data mining of heterogeneous clinical data. Stud Health Technol Inform. 136 (2008), 641-6

[5] Teodoro D *et al*. Biomedical Data Management: A Proposal Framework. Proceedings of the Medical Information for Europe congress. 2009.

[6] Fellegi I.P. and Sunter A.B.: A Theory for Record Linkage. J of the Am Statistical Association, vol. 64, 1969.

[7] Wang R.Y. A product Perspective on Total Data Quality Management. Communication of the ACM, vol. 41, no.2, 1998.

[8] Redman T.C. Data Quality for the Information Age. Artech House, 1996.

[9] Naumann F. and Rolker C. Assessment Methods for Information Quality Criteria. In Proc. Of the MIT Conf. on Information Quality, Cambidge, USA, 2000.

[10] Peralta V. Data quality evaluation in data integration systems. PhD Thesis, Université de Versailles (France) and Universidad de la República (Uruguay), 2006.

[11] Weikum G. Towards guaranteed quality and dependability of information systems. In Proc. of the Conf. Datenbanksysteme inB*uro, Technik und Wissenschaft, Freiburg, Germany, 1999.

[12] Berti-Equille L and Moussouni, F. Quality-aware integration and warehousing of genomic data. Proc. of the 10th Intl. Conf on Information Quality, MIT, Cambridge, U.S.A. (2005)

[13] Krogstie J. Lindland O.I. and Sindre, G. Towards a Deeper Understanding of Quality in Requirements Engineering. Proceedings of the 7th Int Conf on Advanced Information Systems Engineering, Jyvaskyla, Finland, June 1995.

[14] Moody D.L. and Shanks G.G. What Makes A Good Data Model? A Framework For Evaluating And Improving The Quality Of Entity Relationship Models"", Australian Computer Journal, August 1998.

[15] Moody D.L. Measuring the quality of data models: an empirical evaluation of the use of quality metrics in practice. Proceedings of the 11th Euro Conf on Information Systems, 2003.

[16] Goldberg SI, Niemierko A and Turchin A, Analysis of data errors in clinical research databases. AMIA Annual Symp Proc. 2008 Nov 6:242-6.

[17] Arts D., De Keizer N and Scheffer G.J. Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework. JAMIA 2002 vol. 9 (6) pp. 600-611.

[18] Kerr K, Norris A and Stockdale, R. Data Quality Information and Decision Making: A Healthcare Case Study. Proc. 18th Australasian Conf on Information Systems, 2007.

[19] Brown PJB and Sonksen P. Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model. JAMIA 2000 vol. 7 (4) pp. 392-403.

[20] Cornet *et al*. Overcoming Barriers to Evaluation of Terminological Systems. Stud Health Technol Inform (2004).

[21] Daniel C, Buemi A, Mazuel L, Ouagne D and Charlet J. Functional requirements of terminology services for coupling interface terminologies to reference terminologies. Stud Health Technol Inform. 2009;150:205-9.

Address for correspondence

Rémy Choquet
INSERM, UMR 872 Eq. 20
15 rue de l'école de médecine, 75006 Paris Cedex France
remy.choquet@spim.jussieu.fr