MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

A Web-based Service for the Management of Data Quality

Jürgen Stausberg, Reinhard Mauer, Michael Nonnemacher

Institute for Medical Informatics, Biometry and Epidemiology, University Duisburg-Essen, Germany

Abstract

We developed a guideline for the adaptive management of data quality in registers and cohort studies. The guideline is to focus on the adaptation of source data verification (SDV) and feedback to the actual level of data quality observed. 24 indicators for data quality are combined into one ordinal quality score, each with its individual relative weight. Sample sizes for source data verification are calculated so that the sample size will be larger for sites with bad data quality and smaller for those with good quality. The observed data quality is reported back to each site, optionally with open or anonymous benchmark data of the other sites. The recommendations as well as the workflow model for its use were implemented in a three-tier-architecture as web-based service. Through a web-browser or by upload of XML-files the input data are recorded. After execution of the guideline the recommendations are available in XML-format, as PDF-files or through the web-interface. Data quality becomes a crucial issue in the fast growing use of health-related data. The guideline and its implementation as web-based service is a major step towards a systematic and evidence-based approach covering respective challenges in data management.

Keywords:

data quality, feedback, quality indicator, source data verification, web-based service

Introduction

It is time to replace a skeptical position to mass data in health care with a systematical approach for the control of data quality. That should include the measurement of data quality, the interpretation of data quality parameters, and the linkage between data quality and data management procedures. As member of a meta-organization for networked medical research (called in German "Telematikplattform für Medizinische Forschungsnetze e.V.", cf. http://www.tmf-ev.de/) we conducted a project dealing with data management of registers and cohort studies.

Focusing on requirements of research databases we developed a guideline for an adaptive management of data quality and implemented a web-based service for a computer mediated application of its recommendations. The recommendations took also into account studies, concepts, and experiences from routine care, clinical studies, and other industries. To some extend the results are therefore applicable to non-research databases.

Adaptive management of data quality

Based on a detailed literature review and expert interviews the guideline identifies 24 quality indicators in the categories plausibility, organization, and trueness. A threshold differentiating between normal and problematic rates and a relative weight is given for each indicator.

A score of data quality is calculated in the following way. The rate measured by a project is compared with the threshold, either with its raw value or with a confidence interval. The weights of indicators with normal results are summed up. The sum is divided by the optimal sum of weights of the included indicators and multiplied with 100. That gives a score between 0 and 100 which is independent from the number of included (i.e. for the project at hand calculable) indicators. We defined 5 grades of data quality: 0 to less 20 points - very poor, 20 to less 40 points - poor, 40 to less 60 points - moderate, 60 to less 80 points - good, and 80 to 100 points - very good.

Within the scope of the guideline, the data quality grade is used twofold. First it is used to adapt source data verification, i.e. the comparison of registered data with source data as the paper-based patient record, to the data quality in such a manner, that resources are concentrated to sites with worse data quality. Secondly, the information on data quality is included in regular reports. Both actions are regarded as operations for quality control to reach a predefined level of data quality.

Source data verification (SDV)

SDV is an established procedure for the assessment of completeness and correctness of data quality. Its application could also raise data quality through the interaction with the responsible sites. The guideline offers recommendations for the volume (number of observation units e. g. patients), depth (number of data items), and frequency of SDV in a given interval. It takes into account the data quality grade, rates of the two indicators "agreement with source data referring to data items" and "agreement with source data referring to observation units" from the last analysis period, a precision parameter and the estimated total number of data items and observation units in the next period.

Sample size calculation of the volume is based on an estimation of the number of observation units with at least one wrong item. We assume a binomial distribution which is approximated by a normal distribution. The adaptation to data quality is controlled by the absolute value of the confidence interval's half width, which is smaller in case of poor data quality and larger in case of good data quality. We propose 0.01 for grade very poor, and an increment of 0.01 until 0.05 for grade very good. Sample size calculation of the depth is done identically. Frequency is calculated by spreading the necessary number of visits for the next analysis period regularly.

Feedback

Feedback improves data quality. In the guideline's recommendations three feedback strategies are offered: for a specific site no further data included, for a specific site with additional anonymous benchmark data for other sites, and for a specific site with additional open benchmark data of other sites.

The feedback consists of reports with information on data quality score, grade, and results of each included quality indicator with thresholds, weight, and rate. The reports are combined for the central data management as well as divided for each participating site.

Application

From the beginning it was planned to offer the guideline paper-based as well as mediated through a computer application. In consequence, the role of the guideline within data management was modeled using the Unified Modeling Language (UML) even in the paper-based form. UML was then chosen as methodology for software design as well. We complemented the activity diagrams with class and sequence diagrams. In addition to the requirements formulated in the guideline, the application should offer multi-subscriber capability, a history of recommendations, an interactive web-interface, input-/output-interfaces for data in eXtensible Markup Language (XML) and an output in Portable Document Format (PDF).

For the WWW-based service a three-tier architecture was chosen making special use of the Model-View-Controller (MVC) design pattern. On the data-layer (Model) the data are stored in XML-files. XML-files exist for the following data-subsets: general information on the project (input), numerators and denominator of each quality indicator (input), constants used for statistical calculations (input), summary covering quality indicators etc. (output), and results of the sample size calculation (output).

According to the web-based approach a standard WWW-Bowser is sufficient as user interface (View). The link between the user interface and the data is realized by Java Server Faces, a special extension of Java Server Pages enabling a simple definition of dynamic HTML-pages (Controller). The use of a secure protocol (HTTPS) insures the confidentiality of the transmitted data.

Further work

Focusing on research data bases as registers and cohort studies, the implemented guideline represents the utmost evidence in the field of management of data quality. Many thoughts will be relevant for data bases in routine health care as electronic health records as well. The delivery of the guideline through a web-based service offers a mean to overcome the obstacle of guideline implementation in daily practice. But some tasks to solve remain for the future.

Workload is still needed to apply the guideline, even by using the web-based service. In particular, the determination of numerator and denominator of each quality indicator by the central data management is labor-intensive. A generic interface to research data bases in medicine would be desirable, but was outside the attainable objectives of our project. Some years will pass until interface standards are available covering the details necessary for the determination of quality indicators.

Still needed is an empirical evaluation of the guideline's impact on data quality. We are in preparation of a randomized, multicentric and open trial with about 120 sites from a couple of projects. In this trial we will have three arms, one performing a full SDV, one following the guidelines recommendations using the web-based service, and one without any SDV. The guidelines itself could not be blinded, because it was published by the TMF.

Acknowledgments

The work presented here is a project for cooperative medical research on behalf of the "Telematikplattform für Medizinische Forschungsnetze e.V." and was funded by the Federal Ministry of Education and Research. We thank our project partners from the Competence Networks on Congenital Heart Defects, Dementia, HIV/AIDS, and Parkinson's disease for cooperation.

References

[1] Stausberg J, Nonnemacher M, Weiland D, Antony G, Neuhäuser M. Management of Data Quality - Development of a Computer-Mediated Guideline. In: Hasman A, Haux R, van der Lei J, De Clercq E, Roger France FH, eds. Ubiquity: Technologies for Better Health in Aging Societies. Amsterdam: IOS, 2006: 477-82.

Address for correspondence

Priv.-Doz. Dr. med. Jürgen Stausberg
Institute for Medical Informatics, Biometry and Epidemiology
University of Duisburg-Essen,
Hufelandstr. 55, D-45122 Essen, Germany
Tel.: +49 201 723 4512, Fax: +49 201 723 5933
E-Mail: stausberg@ekmed.de