Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-549

Square² - A Web Application for Data Monitoring in Epidemiological and Clinical Studies

Carsten Oliver SCHMIDT[1], Christine KRABBE, Janka SCHÖSSOW, Martin ALBERS, Dörte RADKE, Jörg HENKE
Institute for Community Medicine SHIP-KEF, University Medicine of Greifswald, Walther Rathenau Str. 48, 17475 Greifswald

Abstract. Valid scientific inferences from epidemiological and clinical studies require high data quality. Data generating departments therefore aim to detect data irregularities as early as possible in order to guide quality management processes. In addition, after the completion of data collections the obtained data quality must be evaluated. This can be challenging in complex studies due to a wide scope of examinations, numerous study variables, multiple examiners, devices, and examination centers. This paper describes a Java EE web application used to monitor and evaluate data quality in institutions with complex and multiple studies, named Square². It uses the Java libraries Apache MyFaces 2, extended by BootsFaces for layout and style. RServe and REngine manage calls to R server processes. All study data and metadata are stored in PostgreSQL. R is the statistics backend and LaTeX is used for the generation of print ready PDF reports. A GUI manages the entire workflow. Square² covers all steps in the data monitoring workflow, including the setup of studies and their structure, the handling of metadata for data monitoring purposes, selection of variables, upload of data, statistical analyses, and the generation as well as inspection of quality reports. To take into account data protection issues, Square² comprises an extensive user rights and roles concept.

Keywords. data quality; data monitoring; web-application; epidemiology; statistical analyses

1. Introduction

To enable valid scientific inferences from epidemiological and clinical studies it is essential to obtain a high data quality. A broad scope of recommendations for study design, and quality assurance measures have been described to achieve this goal.[6; 10]. An indispensable aspect of quality assurance processes is a functioning data quality monitoring. Several data quality indicators have been described [1-3; 7; 8] for this purpose. They target data properties such as missing values, implausible values, extreme values, and measures of reliability and validity. For example, the "Guideline for the Adaptive Management of Data Quality in Cohort Studies and Registers" describes 51 quality indicators which are organized in the categories of plausibility (27 indicators), organization (16 indicators), correctness (6 indicators), and metadata (2 indicators).[5]

Data generating departments aim to detect data irregularities as early as possible to guide quality management processes. After the completion of data collections the quality

[1] Corresponding author, University Medicine Greifswald, Institute for Community Medicine, SHIP-KEF, Walther-Rathenau Str. 48. 17475 Greifswald; E-mail: Carsten.schmidt@uni-greifswald.de.

of this data must be rigorously evaluated before the start of scientific analyses.[4] In complex studies this can be a challenging task due to a wide scope of examinations with numerous study variables, as well as multiple examiners, devices, and examination centers. The Study of Health in Pomerania (SHIP) may serve as an example for this complexity.[9] SHIP studies the prevalence and incidence of risk factors, subclinical disorders, clinical diseases, and their associations. To date, almost 9000 adults have been examined up to five times. Examinations consisted of an extensive computer assisted personal interview, self-report questionnaires, the collection of biomaterials (blood, urine, faeces, saliva), imaging (e.g. ultrasound of the carotid artery, liver, thyroid, heart; full-body magnetic resonance imaging (MRI)), a dental and dermatological examination, and more. Thousands of variables, grouped into dozens of examination categories must be managed. The complexity is increased by changing examination teams, temporal examination centers and the conduct of numerous secondary data generation projects, which are for example related to the reading of MRI images (e.g. disc herniation). All of these secondary data collections may themselves be regarded as studies with an independent workflow. Other major cohort studies face comparable complexities.[11]

Under these circumstances, an efficient, manual data monitoring method seemed no longer feasible. Therefore, we aimed to standardize processes by developing appropriate IT tools. First, a partial standardization of the workflow was achieved by combining a manually controlled STATA analysis environment with a web frontend to generate PDF reports. This was implemented in the year 2010. Second, based on our experience with the first data quality analysis tool and our interactions with internal and external SHIP project partners, and team members we developed a web application to control the entire data monitoring process, Square², which is described in detail in this paper. The decision for a new development was made because (1) software solutions to monitor data quality were highly uncommon in major epidemiologic studies [3] despite their use in other fields of research, (2) existing solutions did not meet to a sufficient degree the requirements of large epidemiologic cohort studies. Important requirements were, among others, a standalone web-application allowing for a multi-study management with a differentiated rights- and roles concept to safeguard data protection issues, the possibility to automatically generate standard reports without statistical programming, the option to flexibly adapt reports to individual demands, a strong focus on measurement error related issues, flexible extension of statistical functionalities based on the integration of standard statistical packages (e.g. R), and non-commercial availability of all components to avoid additional costs for academic users.

2. Methods

Square² was designed as a Java EE web application, deployed in Tomcat 8. All data is stored in a PostgreSQL database. The statistics backend is R because of its free availability, wide acceptance in the statistical community, fast growing scope of packages, and the option to run it as a server process. LaTeX is used for the generation of print ready PDF reports. An overview of the components is provided in Figure 1. A GUI manages the entire workflow.

Square² uses the Java libraries Apache MyFaces 2, extended by BootsFaces for layout and style. RServe and REngine manage calls to R server processes. Additionally, we use unit testing libraries (JUnit). The following in-house developed libraries are used:

ShipDBM, a data persistence library, providing access of the web application to PostgreSQL, and Pwencrypt, a library for password encryption.

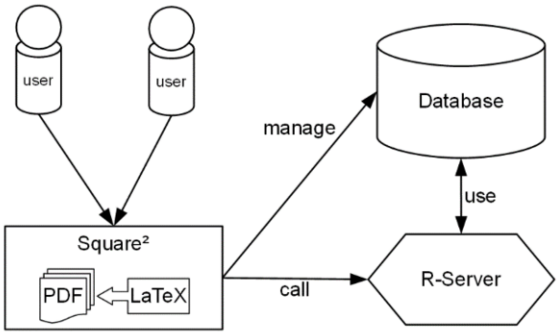

Figure 1. Square² Components.

R performs statistical analyses, and results are stored in the database. Analyses can be processed asynchronously to improve computational speed. The in-house developed R package squareControl manages all R server processes. It uses the package „rpg“ for persistence, "futile.logger" to log process information, and "parallel" to parallelize tasks.

Study data as well as all metadata and analyses results are stored in PostgreSQL. Square² starts calculations on the R server and disconnects thereafter, leaving all subsequent processes to the R Server. Thus all computational load remains on the R Server, including data base operations. All analysis results are stored in PostgreSQL. Graphical elements are embedded in LaTeX documents as base64 encodings.

3. Results

Square² is currently used within SHIP [9] but has been designed to meet the needs and requirements of different studies. The workflow consists of the following steps:

1. Study management: First, a new study needs to be defined in Square² for data monitoring purposes by providing descriptive information such as study name, study description, and if possible begin and end dates.
2. Study structure: In this step, the study structure and all necessary metadata for data monitoring are defined. Hierarchical elements of a study may consist of groups of examinations (e.g. medical examinations), examinations (e.g. hand grip, anthropometric measurements, blood pressure), and variables (e.g. the first measurement of hand grip strength in the left hand in kg). Next, metadata associated with a study and single variables are added. Metadata fields include, for example, the variable type (e.g. categorical, continuous, count), plausibility limits for continuous or count variables, reference categories for categorical variables, missing value indicators, observer, device or center indicators, and measurement times and dates. Elements of the study structure may either be added manually by using the GUI or by an import of available metadata.
3. Variable sets. Sets of variables are subsequently defined for data monitoring purposes. These sets may consist of variables from different studies. Quality

officers may only create reports for variables from variable sets to which they have been assigned.

4. Data management. This functionality controls the upload of study data for statistical analyses.
5. Statistics. The statistics module serves to enter R scripts into Square². For each R script, input and output parameters must be defined to properly link R scripts with variable and study metadata and to enable the web-application to integrate statistical output into reports. R-scripts are assigned to predefined report categories (e.g. descriptive statistics, missing values, extreme values, observer or device variability). Statisticians may add statistical functions without in depth knowledge of the web application.
6. Templates. The reporting of data quality within studies often follows standard requirements and reports should be highly comparable. For this purpose, templates can be created to structure reports (using elements such as headers, sub-headers, text blocks, tables, statistical output, page breaks).
7. Quality reports. The preceding steps provide the necessary background information to now create specific quality reports. First, a new report is defined by assigning a name, templates, and a variable set. Second, an analysis matrix is created to link variables from the variable set with statistical routines. Third, additional information may be entered into the predefined text fields. Fourth, the report is generated, and may subsequently be inspected.

Once all background information is entered, users mainly work in step seven. Square² includes an extensive user rights and roles concept. This includes tailored access to study data and reports by assigning personnel to specific studies, or even to specific sets of variables within studies to protect data safety. Multiple roles can be managed such as principal investigators (e.g. for the definition or deletion of studies), quality officers (e.g. to add and modify study metadata and to create reports), statisticians (e.g. to add R scripts), and examiners (e.g. to only read reports related to their own examinations).

4. Conclusion

Square² was primarily designed to support the monitoring of data quality in institutions running multiple complex studies. It allows for an efficient and timely generation of standard data quality reports. Efficiency is essential because funding for extensive data quality control activities is often limited. The design of Square² draws strongly from the concept of automatization of monitoring processes. However, there are limits. Square² may not be the most appropriate tool to address highly specific data quality aspects. This limitation is mainly related to logistic considerations. While the modular design of Square² allows for a strong degree of individualization, the integration of such highly specific analyses might be realized more quickly outside the Square² framework. The strengths of Square² are most apparent when there is a need for a repetitive reporting with a centralized demand to oversee and control these reporting activities.

Another limitation is related to the focus on the properties of measured data. However, the obtained data quality in a study may only be fully appreciated by interpreting results in light of additional study meta-information, such as the study design.

The development of Square² is still ongoing. Access to report contents may for example be based on HTML pages rather than PDF reports. We intend to classify R-

routines based on data quality indicators as described by the "Guideline for the Adaptive Management of Data Quality in Cohort Studies and Registers".[5]

Square² may be accessed for academic use through scientific cooperation projects, please contact the first author for this purpose. Free access for academic users to a web-based instance of Square² is projected to be available as part of an ongoing network project on data quality indicators in cooperation with the TMF (http://www.tmf-ev.de), an umbrella organization for networked medical research in Germany.

5. Acknowledgements

This work was supported by the Ministry for Education, Science and Culture of the State of Mecklenburg-Vorpommern, the European Social Fund (Grant UG 11 035A), and by the German Research Foundation (DFG, SCHM 2744/3-1).

References

[1] D.G. Arts, N.F. De Keizer, and G.J. Scheffer, Defining and improving data quality in medical registries: a literature review, case study, and generic framework, *Journal of the American Medical Informatics Association* **9** (2002), 600-611.

[2] J.R. Brestoff and J. Van den Broeck, Reporting Data Quality, in: *Epidemiology Principles and Practical Guidelines,* J. Van den Broeck and J.R. Brestoff, eds., Springer, Dordrecht, 2013.

[3] O. Harel, E.F. Schisterman, A. Vexler, and M.D. Ruopp, Monitoring quality control: can we get better data?, *Epidemiology* **19** (2008), 621-627.

[4] M. Huebner, W. Vach, and S. le Cessie, A systematic approach to initial data analysis is good research practice, *Journal of Thoracic and Cardiovascular Surgery* **151** (2016), 25-27.

[5] M. Nonnemacher, D. Nasseh, and J. Stausberg, *Datenqualität in der medizinischen Forschung: Leitlinie zum Adaptiven Datenmanagement in Kohortenstudien und Registern*, TMF e.V., Berlin, 2014.

[6] C.O. Schmidt, Anwendungsempfehlungen für Kohorten in: *Datenqualität in der medizinischen Forschung: Leitlinie zum Adaptiven Datenmanagement in Kohortenstudien und Registern.,* M. Nonnemacher, N. D., and S. J., eds., TMF e.V., Berlin, 2014, pp. 117-127.

[7] J. Van den Broeck, S.A. Cunningham, R. Eeckels, and K. Herbst, Data cleaning: detecting, diagnosing, and editing data abnormalities, *PLoS Medicine* **2** (2005), e267.

[8] D. Venet, E. Doffagne, T. Burzykowski, F. Beckers, Y. Tellier, E. Genevois-Marlin, U. Becker, V. Bee, V. Wilson, C. Legrand, and M. Buyse, A statistical approach to central monitoring of data quality in clinical trials, *Clinical Trials* **9** (2012), 705-713.

[9] H. Volzke, D. Alte, C.O. Schmidt, D. Radke, R. Lorbeer, N. Friedrich, N. Aumann, K. Lau, M. Piontek, G. Born, C. Havemann, T. Ittermann, S. Schipf, R. Haring, S.E. Baumeister, H. Wallaschofski, M. Nauck, S. Frick, A. Arnold, M. Junger, J. Mayerle, M. Kraft, M.M. Lerch, M. Dorr, T. Reffelmann, K. Empen, S.B. Felix, A. Obst, B. Koch, S. Glaser, R. Ewert, I. Fietze, T. Penzel, M. Doren, W. Rathmann, J. Haerting, M. Hannemann, J. Ropcke, U. Schminke, C. Jurgens, F. Tost, R. Rettig, J.A. Kors, S. Ungerer, K. Hegenscheid, J.P. Kuhn, J. Kuhn, N. Hosten, R. Puls, J. Henke, O. Gloger, A. Teumer, G. Homuth, U. Volker, C. Schwahn, B. Holtfreter, I. Polzer, T. Kohlmann, H.J. Grabe, D. Rosskopf, H.K. Kroemer, T. Kocher, R. Biffar, U. John, and W. Hoffmann, Cohort profile: the study of health in Pomerania, *Int J Epidemiol* **40** (2011), 294-307.

[10] C.W. Whitney, B.K. Lind, and P.W. Wahl, Quality assurance and quality control in longitudinal studies, *Epidemiologic Reviews* **20** (1998), 71-80.

[11] H.E. Wichmann, R. Kaaks, W. Hoffmann, K.H. Jöckel, K.H. Greiser, and J. Linseisen, [The German National Cohort], *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* **55** (2012), 781-787.