MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190483

Automated Reports for Monitoring and Improving Data Quality in a Translational Research Network

Esther E Schmidt[a,b], Corinna Eichelser[a,b,c,d], Bernd Ahlborn[a,b,e], Dietmar Keune[f], Esmeralda Castaños-Vélez[g], David Juárez[a,b], Martin Lablans[a,b]

[a] *German Cancer Consortium (DKTK), Heidelberg, Germany,*
[b] *German Cancer Research Center (DKFZ), Heidelberg, Germany,*
[c] *Department of Internal Medicine III, Klinikum der Universität München (KUM), Germany,*
[d] *Department of Medicine III, Klinikum rechts der Isar (MRI), TUM, Germany,*
[e] *Institute for Medical Information Sciences, Biometry and Epidemiology (IBE), Ludwig-Maximilians-Universität München (LMU), München, Germany,*
[f] *Clinical Cancer Registry, Charité Comprehensive Cancer Center, Berlin, Germany,*
[g] *Experimental Neurology, Charité – Universitätsmedizin Berlin, Berlin, Germany*

Abstract

Standardised, automated quality reports were generated at three pilot locations of the decentralized translational research network DKTK with separated local data warehouses (LDW), for assessing syntactic conformity against common data element definitions deposited in a central metadata repository (MDR). Deviations in the LDW were categorised, and locally corrected. Comparisons of reports from two time points confirm a major improvement in data quality in terms of syntactic conformity, an essential prerequisite for network-wide data integration.

Keywords:

Common Data Elements, Metadata, Data Accuracy

Introduction

The German Consortium for Translational Cancer Research (DKTK) [1] is a network of more than 20 institutions at nine locations, pursuing the aim of efficiently transferring results from cancer research into the clinical practice. The DKTK implements a federated data storage concept, interconnecting LDW with so-called bridgeheads [2]. In order to allow comparisons between the partners' data on clinical cases and biomaterial, the network has agreed on a common data set (MDS) with defined data elements and their permitted values, which has been deposited in a central metadata repository (MDR) [3,4]. An essential prerequisite for the success of data integration processes within decentral research networks is high data quality [5,6]. For the assessment of data quality (here: syntactic conformity [7] with the MDR) we implemented a quality report generator (QR-generator), which can be installed at every network location, in order to perform an alignment of data in the LDW against the centrally defined data definitions in the MDR [8,9].

Methods

The QR-generator installed at a given network location produces quality reports in Excel format as described in [8,9]. Briefly, this report lists all values recorded for each data element in the LDW, highlighting each value that is not concordant with the respective definition specified in the MDR. In addition, the report provides statistical calculations for each data element with regard to completeness and syntactical correctness. Upon production of a report, all errors are manually categorised, compared with network partners' errors, and suggestions are given for solutions to be implemented by the site. Then a second report is produced and compared with the previous one, in order to assess changes in data quality. Though the reports provide detailed mismatch information on a per-patient basis to the local sites, the results described here are summarized on a per-data-element basis for data protection reasons.

Results

QR-generators were installed at each of the nine DKTK locations. In a pilot study, sets of quality reports have been produced at three network locations, allowing the sequential assessment of data quality over a period of 8-10 months.

At the Charité in Berlin, the LDW in the bridgehead serves as the primary tumour documentation system. Quality reports have been produced in March and December 2017 (104.000 and 137.000 patients in the LDW, respectively). The following error categories have been identified: deviations in the source system (including the usage of special codes that are not conformant with the German cancer coding system); minor syntactic errors such as differences in capitalisation; missing mappings of local values to the MDS; technical errors, e.g. with regard to date definitions. Error correction strategies included re-documentation in the source system under controlled conditions based on SOPs; adjustment of the technical interfaces to correct systematical deviations as well as mappings at the level of data elements or their values.

Out of 52 data elements assessed, 21 were syntactically correct from the beginning (Figure 1). Two data elements were fully corrected directly in the source system. Other data elements, showing capitalisation errors or additional text following a systematic pattern, could be corrected by adjustments to the technical interface, improving MDR conformity considerably, in six cases to more than 95%. For three data elements, an additional adjustment of the mapping at value level led to complete MDR conformity.

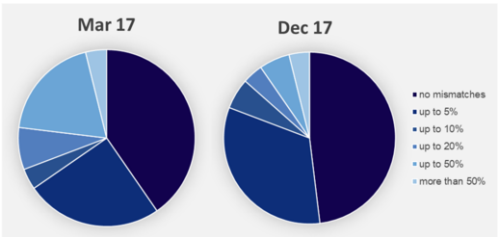

Figure 1 - Assessment of 52 data elements at the Charité, Berlin, at two time points

In Munich, there are two university medical centers hosting their own bridgeheads, namely the MRI, and the KUM. Both institutions provide an LDW in the bridgehead purely dedicated to making data available to the DKTK network. Data from the primary tumour documentation systems are transferred to the LDW via "extract-transform-load" (ETL) processes. ETL processes are well suited to handle systematic deviations. The major challenge here consists in the identification of the correct mappings between the primary systems and the LDW. The error categories identified were: missing data elements (not found); non-MDR-conformant value mappings (mismatch); correct element mapping with some remaining individual errors (partial match). Error correction strategies relied mainly on adjusting the local ETL processes, as well as centrally provided adjustments to the technical interface.

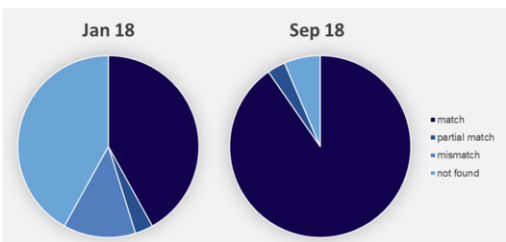

Figure 2 - Assessment of 62 data elements at the MRI at two time points

At the MRI, out of 62 data elements assessed in January 2018, 26 showed a perfect syntactic match (Figure 2). Another 26 data elements were not found in the LDW, while 10 data elements had mismatches or partial matches. At re-assessment in September 2018, the vast majority of data elements (90%) were fully MDR-conformant, with only 6 elements not being found or showing partial matches.

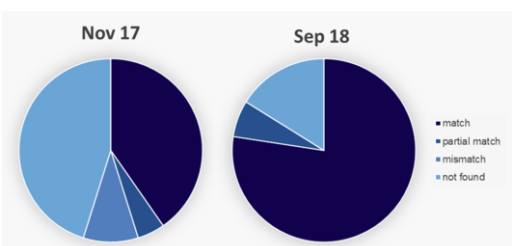

Figure 3 - Assessment of 62 data elements at the KUM at two time points

At the KUM, the data element status at the beginning (November 2017) was very similar to the MRI (Figure 3). Here, 48 out of 62 data elements (77%) showed full MDR-conformity at re-assessment in September 2018, while 14 elements were not found or had partial matches. Five of the data elements are missing due to the fact that the biobank could not yet be integrated into the LDW for organisational reasons.

Conclusions

Standardised quality reports represent suitable tools for the assessment and continuous improvement of data quality at DKTK locations. Moreover, they facilitate the exchange and consolidation of data within the network and help the local documentation systems in identifying and correcting errors. Given the dynamic of newly added data elements and growing patient numbers, the monitoring of data quality represents a continuous challenge.

Acknowledgements

This research is funded by the German Cancer Consortium (DKTK).

References

[1] Klinische Kommunikationsplattform (CCP) im DKTK, https://dktk.dkfz.de/de/klinische-plattformen/ueber-die-ccp/about-ccp [accessed on April 1, 2019].

[2] M. Lablans, E.E. Schmidt, and F. Ückert, An Architecture for Translational Cancer Research As Exemplified by the German Cancer Consortium, *JCO Clinical Cancer Informatics* **1** (2017), 1–8.

[3] A.A. Sinaci, and G.B. Laleci Erturkmen, A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains, *J Biomed Inform* **5** (2013), 784–794.

[4] Metadata Repository (MDR) des DKTK, https://mdr.ccp-it.dktk.dkfz.de [accessed on April 1, 2019].

[5] M.G. Kahn, T.J. Callahan, J. Barnard, A.E. Bauck, J. Brown, B.N. Davidson, H. Estiri, C. Goerg, E. Holve, S.G. Johnson, S.-T. Liaw, M. Hamilton-Lopez, D. Meeker, T.C. Ong, P. Ryan, N. Shang, N.G. Weiskopf, C. Weng, M.N. Zozus, and L. Schilling, A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data, *EGEMS (Washington, DC)* **1** (2016), 1244.

[6] I. Fortier, P.R. Burton, P.J. Robson, V. Ferretti, J. Little, F. L'Heureux, M. Deschênes, et al., Quality, quantity and harmony: The DataSHaPER approach to integrating data across bioclinical studies, *International journal of epidemiology* **5** (2010), 1383–1393.

[7] Batini C, Scannapieco M (2016) *Data and Information Quality: Dimensions, Principles and Techniques,* Springer, Cham.

[8] D. Juarez, E.E. Schmidt, M. Lablans, K. Senghas, and F. Ückert, Report generator for the external validation of conformity of local clinical data to a global metadata repository (MDR), *62. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS)* (2017).

[9] D. Juarez, E.E. Schmidt, S. Stahl-Toyota, F. Ückert, and M. Lablans, A generic method and implementation to evaluate and improve data quality in distributed research networks (submitted), (2019)

Address for correspondence

Dr. Esther Schmidt, DKFZ Heidelberg, Germany
e.schmidt@dkfz-heidelberg.de