A Generic Method and Implementation to Evaluate and Improve Data Quality in Distributed Research Networks

D. Juárez[1,2] E.E. Schmidt[1,2] S. Stahl-Toyota[3] F. Ückert[3] M. Lablans[1,2]

[1] Federated Information Systems, German Cancer Research Center (DKFZ), Heidelberg, Germany
[2] German Cancer Consortium (DKTK), Heidelberg, Germany
[3] Medical Informatics in Translational Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany

Address for correspondence David Juárez, MCS, Federated Information Systems, German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany (e-mail: d.juarez@dkfz-heidelberg.de).

Methods Inf Med 2019;58:86–93.

Abstract

Background With the increasing personalization of clinical therapies, translational research is evermore dependent on multisite research cooperations to obtain sufficient data and biomaterial. Distributed research networks rely on the availability of high-quality data stored in local databases operated by their member institutions. However, reusing data documented by independent health providers for the purpose of care, rather than research ("secondary use"), reveal a high variability in terms of data formats, as well as poor data quality, across network sites.

Objectives The aim of this work is the provision of a process for the assessment of data quality with regard to completeness and syntactic accuracy across independently operated data warehouses using common definitions stored in a central (network-wide) metadata repository (MDR).

Methods For assessment of data quality across multiple sites, we employ a framework of so-called bridgeheads. These are federated data warehouses, which allow the sites to participate in a research network. A central MDR is used to store the definitions of the commonly agreed data elements and their permissible values.

Results We present the design for a generator of quality reports within a bridgehead, allowing the validation of data in the local data warehouse against a research network's central MDR. A standardized quality report can be produced at each network site, providing a means to compare data quality across sites, as well as to channel feedback to the local data source systems, and local documentation personnel. A reference implementation for this concept has been successfully utilized at 10 sites across the German Cancer Consortium.

Conclusions We have shown that comparable data quality assessment across different partners of a distributed research network is feasible when a central metadata repository is combined with locally installed assessment processes. To achieve this, we designed a quality report and the process for generating such a report. The final step was the implementation in a German research network.

Keywords
- medical informatics
- metadata
- data accuracy
- translational medical research
- health information interoperability

received
February 15, 2019
accepted after revision
June 7, 2019

DOI https://doi.org/10.1055/s-0039-1693685.
ISSN 0026-1270.

Introduction

Clinical research in a globalized world relies on the collaborative work of the scientific community. Especially in the context of promising new approaches to personalized medicine, which require broad access to biological samples,[1] research groups can no longer depend on data available at their home institutions alone.[2–4] The need for access to clinical data across multiple institutions is increasingly being addressed by the formation of distributed research networks (DRNs) and infrastructures like PCORnet,[5] BBMRI-ERIC,[6,7] or ELIXIR.[8,9] An important characteristic of a DRN is the local integration and storage of data while making it accessible for cross-site applications. DRNs should also apply FAIR principles (findable, accessible, interoperable, and reusable),[10] as these principles enhance the ability of a DRN to find and use data,[11] as exemplified in the Recommendations for Improving the Quality of Rare Disease Registries.[12]

An essential prerequisite for successful research is the availability of interconnected high-quality clinical data and well-annotated biobank samples at each DRN partner.[12,13] Data quality is a multifaceted challenge: in their classification framework for data quality dimensions, Batini et al[14] divide the concept into eight dimensions. In this work, we focus primarily on the evaluation of completeness and syntactic accuracy of (clinical) data, as these are relatively straightforward to assess in our context, in comparison with other data quality dimensions.

Many DRNs extract preexisting data from one or several distributed source systems and transform it guided by data definitions agreed across the network. A major challenge experienced by such DRNs is the fact that the data in question has been collected in the context of patient care rather than systematically for research ("secondary use").[15,16] The documentation of clinical data at different institutions and for different purposes results in considerable heterogeneity of data formats and quality.[17] Data integration and harmonization, accompanied by data quality assurance processes, are, therefore, essential prerequisites for data usability.[18] However, in DRNs, it is not an easy task to measure data quality centrally.

Objectives

We propose a method to validate the data quality regarding completeness and syntactic accuracy within multiple data warehouses (each operated at a consortium site) using common definitions stored in a central (consortium-wide) metadata repository (MDR).

Methods

►**Figure 1** shows the outline of a DRN, consisting of components installed at each site that connect to central components, which in turn provide applications to researchers. For example, DRNs might provide a central search application allowing scientists to query data throughout the network, or an analysis application to perform statistical calculations. In the following, site components as well as the central MDR are described in more detail.

Site Components: Data Warehouse and Connector

To enable cooperation on shared routine clinical data, the first requirement is a component that provides this data in a uniform manner. In the context of DRNs, this component runs locally at each network site and is called a (local) data warehouse. Several implementations exist for this purpose. For example, i2b2 introduces a "CRC cell"[19,20], PCORnet uses "DataMarts"[21], whereas BBMRI-ERIC,[7] and the German Biobank Alliance (GBA)[22,23] make use of a generic open-source data warehouse based on the "Samply" architecture.

Each network partner populates their data warehouse by means of an ETL (extract-transform-load)[24] process, employing materialized data integration[25,26] to overcome the heterogeneity of the site data sources. Ideally, each data warehouse would ensure that imported data conforms to the data definitions agreed within the DRN (stored in a central MDR, see section "Metadata Repository" below). However, this cannot be taken for granted: First, some implementations, such as i2b2's "CRC cell," do not explicitly validate incoming data. Second, the data warehouse may also use an internal schema that is different from the MDR which

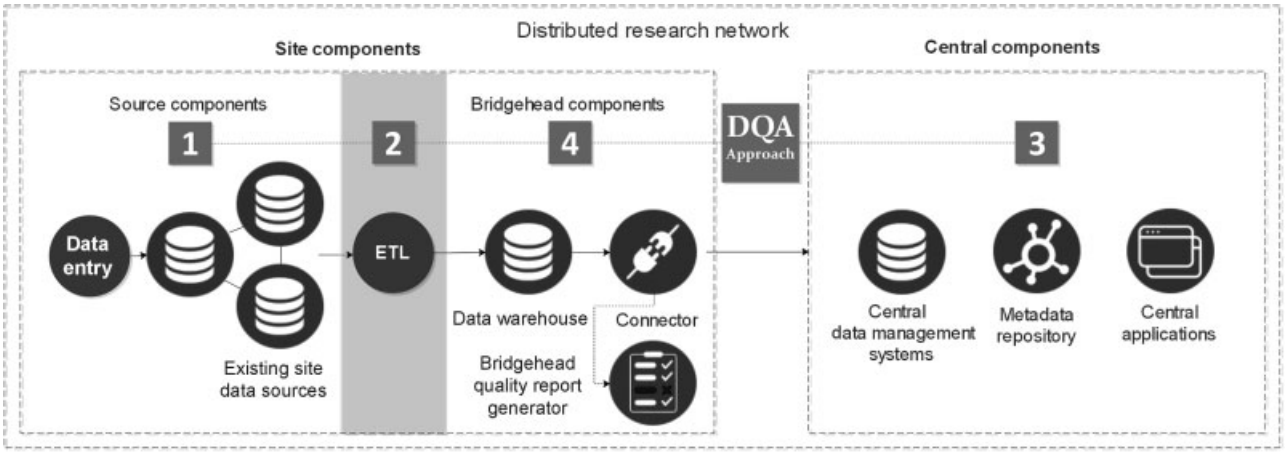

Fig. 1 Typical workflow for data integration in a DRN. The numbers refer to possible approaches to data quality assessment (DQA) discussed below in Section 5. (ETL, extract-transform-load)

necessitates the provision of the correct mapping for each corresponding data element (►Fig. 2A).

The second important component is the entry point of the sites to the DRN: the data warehouses contact central services through some kind of connector for which several implementations exist. For instance, i2b2's "aggregator" makes queries possible across several "CRC cells."[27,28] The equivalent in PCORnet is the "DataMart Client,"[21] whereas in BBMRI-ERIC, it is called "connector."[7,22] In the GBA, as well as the German Cancer Consortium (DKTK), data warehouse and connector ("Samply.Share Client") are distributed to each partner site as part of a bridgehead.[22,29]

	A	B	C	D	E	F	G	H	I	J	K
1	ID in MDR	(optional) ID in research network	Data element in MDR	Data element in data warehouse	Data type	Value (MDR match / mapped)	Value (MDR mismatch / pre-mapped)	Syntactic validation	No. of patients with this value	patients with this value ÷ patients with any value [%]	patients with this value ÷ total number of patients [%]
2	dktk:54:1	A-0	DKTK-ID	DKTK_GLOBAL	STRING			not found	0	0	0
133	dktk:49:4	B-4	Sampling date	SAMPLE_SAMPLINGDATE	DATE	23.03.2015		match	2	0,1	0
3467	dktk:28:1	K-3	Age at diagnosis	DIAG_AGE_OF_DIAG	INTEGER	88		match	108	0,7	0,7
3468	dktk:28:1	K-3	Age at diagnosis	DIAG_AGE_OF_DIAG	INTEGER	89		match	80	0,5	0,5
3539	dktk:28:1	K-3	Age at diagnosis	DIAG_AGE_OF_DIAG	INTEGER		-1	mismatch	2	0	0
3540	dktk:28:1	K-3	Age at diagnosis	DIAG_AGE_OF_DIAG	INTEGER	60		match	349	2,2	2,2
4058	dktk:4:2	K-6	Localisation	TLOC_LOCALISATIONCODE	STRING	C77.09		match	14	0,1	0,1
4257	dktk:4:2	K-6	Localisation	TLOC_LOCALISATIONCODE	STRING		C61.	mismatch	1	0	0
4258	dktk:4:2	K-6	Localisation	TLOC_LOCALISATIONCODE	STRING	C53.01		match	1	0	0
5166	dktk:20:3	K-24	Residual tumor classification	SURGERY_RCLASSIFIC	enumerated	R0		match	3948	87	24,9
5167	dktk:20:3	K-24	Residual tumor classification	SURGERY_RCLASSIFIC	enumerated	R1		match	1003	22,1	6,3
5168	dktk:20:3	K-24	Residual tumor classification	SURGERY_RCLASSIFIC	enumerated	R2		match	284	6,3	1,8
5169	dktk:20:3	K-24	Residual tumor classification	SURGERY_RCLASSIFIC	enumerated		R2a	mismatch	54	1,2	0,3
5170	dktk:20:3	K-24	Residual tumor classification	SURGERY_RCLASSIFIC	enumerated		x	mismatch	190	4,2	1,2
6616	dktk:33:2	K-32	Surgery	PROGRESS_SURGERY	BOOLEAN		X	mismatch	6331	45,8	40
6617	dktk:33:2	K-32	Surgery	PROGRESS_SURGERY	BOOLEAN	true	J	match	10636	76,9	67,2
6618	dktk:33:2	K-32	Surgery	PROGRESS_SURGERY	BOOLEAN	false	N | A	match	7763	56,2	49
6657	dktk:71:2	K-44	Other therapy type		STRING			not mapped	0	0	0

1. info | 2. filtered elements | 3. all elements | 4. patient local ids | 5. data element stats

A

	A	B	C	D	E	F	G	H	I	J	K	L
1	ID in MDR	(optional) ID in research network	Data element in MDR	Data element in data warehouse	No. of patients with any value	patients with any value ÷ patients total [%]	No. of patients with only matching values	patients with only matching values ÷ patients with any value [%]	patients with only matching values ÷ patients total [%]	No. of patients with ≥1 mismatching values	patients with any mismatching value ÷ patients with any value [%]	patients with any mismatching values ÷ patients total [%]
2	dktk:54:1	A-0	DKTK-ID	DKTK_GLOBAL	0	0	0	0	0	0	0	0
8	dktk:49:4	B-4	Sampling date	SAMPLE_SAMPLINGDATE	2861	18,1	2861	100	18,1	0	0	0
12	dktk:28:1	K-3	Age at diagnosis	DIAG_AGE_OF_DIAG	15834	100	15832	100	100	2	0	0
15	dktk:4:2	K-6	Localisation	LOC_LOCALISATIONCODE	15114	95,4	15113	100	95,4	1	0	0
33	dktk:20:3	K-24	Residual tumor classification	SURGERY_RCLASSIFIC	4540	28,7	0	0	0	4540	100	28,7
37	dktk:33:2	K-32	Surgery	PROGRESS_SURGERY	13825	87,3	7494	54,2	47,3	6331	45,8	40
49	dktk:71:2	K-44	Other therapy type		0	0	0	0	0	0	0	0

1. info | 2. filtered elements | 3. all elements | 4. patient local ids | 5. data element stats

B

$$v = value$$
$$de = data\ element$$
$$\# = number\ of$$

$$D(de) = value\ domain\ of\ de\ (permissible\ values\ for\ de)$$

$$t(de) = \begin{cases} 1 \ / \ data\ type\ (de) & \in \{String, Date, Integer\} \\ 0 \ / \ data\ type\ (de) & \neg\in \{String, Date, Integer\} \end{cases}$$

$$P = \#patients\ (total\ number\ of\ unique\ patients)$$
$$p(de) = \#patients\ with\ at\ least\ one\ value\ for\ de$$
$$p\ (de, v) = \#patients\ with\ value\ v\ for\ de$$

Matches:

$$p_m(de) = \#patients\ with\ all\ values\ v_i\ for\ de\ /\ v_i \in D(de)$$

Mismatches:

$$p_{\overline{m}}(de) = \#patients\ with\ any\ value\ v\ for\ de\ /v \neg\in D(de)$$

Filtered:

$$p_f(de, v) = \begin{cases} p(de, v) \ / \ v \neg\in D(de) \ \vee v \in D(de) \ \wedge \ t(de) = 0 \\ p_m(de, v) \ / \qquad\qquad v \in D(de) \ \wedge \ t(de) = 1 \end{cases}$$

$X_Y = MS\ Excel\ column\ X, sheet\ Y$

$$I_2(de, v) = p_f(de, v) \qquad I_3(de, v) = p(de, v)$$
$$J_2(de, v) = \frac{p_f(de, v)}{p(de)} \qquad J_3(de, v) = \frac{p(de, v)}{p(de)}$$
$$K_2(de, v) = \frac{p_f(de, v)}{P} \qquad K_3(de, v) = \frac{p(de, v)}{P}$$

$$E_5(de) = p(de)$$
$$F_5(de) = \frac{p(de)}{P}$$

$$G_5(de) = p_m(de)$$
$$H_5(de) = \frac{p_m(de)}{p(de)}$$
$$I_5(de) = \frac{p_m(de)}{P}$$

$$J_5(de) = p_{\overline{m}}(de)$$
$$K_5(de) = \frac{p_{\overline{m}}(de)}{p(de)}$$
$$L_5(de) = \frac{p_{\overline{m}}(de)}{P}$$

C

Fig. 2 Example QR: sheets with relevant columns for the validation of syntactic accuracy (**A**) and completeness (**B**). Several columns are explained with formulas (**C**), for example, the formula corresponding to sheet 5 colomn I (I_5) is highlighted with a frame. For the sake of this publication, we have translated the text to English and obfuscated all numbers. MDR, metadata repository; QR, quality report.

Metadata Repository

The second requirement for cooperation on shared data is its availability in a common format. We assume that there is a commonly agreed dataset to validate against, stored in a machine-readable format. For example, the International Organization for Standardization/International Electrotechnical Commission's (ISO/IEC) 11179 standard describes data elements arranged into data element groups. This standard includes designations, definitions, and value domains. Other initiatives, such as "openEHR," tackle this objective by developing various models and specifications, facilitating the sharing of health records by clinicians and other users.[30]

Our approach does not require a specific MDR implementation. We assume, however, that for each data element, the MDR provides information to validate the values deposited in the data warehouse, for example, value ranges (for numerical data elements), regular expressions (for strings) or a list of permissible values. In this article, we focus on the "Samply.MDR" implementation which has been developed in the German Cancer Consortium[29] as a server application[31] derived from ISO 11179–3 and has seen since wide use in several DRNs.[32,33] It is accessible through both a REST-based API[34] (Representational State Transfer/ Application Programming Interface; ►Fig. 1) and a web-based user interface for browsing and editing of metadata.

Results

We propose (1) a generic method to validate data stored in a DRN's local data warehouse against a central MDR by means of a locally installed "quality report generator" and (2) a reference implementation which has successfully been installed at ten sites of a translational DRN.

Quality Report Generator

The quality report generator (QR-generator) works in five steps, depicted in ►Fig. 3:

- The QR-generator is initialized with the identifiers (IDs) of the MDR data elements to be validated. These IDs are preconfigured by the bridgehead's administrator within a web administration interface or predefined by the DRN provisioning the bridgehead.
- With the resulting list of data elements, the QR-generator queries the data warehouse's REST API for all patient datasets containing an entry for at least one of those data elements.
- The QR-generator reads the values of the requested data elements for each patient and stores the patient IDs for each data element–value pair, allowing to assess data completeness[14,35] in subsequent analyses.
- The QR-generator validates each value syntactically against the permissible value definitions retrieved from the MDR. In the example of ►Fig. 3, the value domain of the data element "evaluation residual tumor" consists of the valid data values "R0," "R1," "R2," etc., and would invalidate any diverging entries such as "R2a."
- The results are saved in a comma-separated values (CSV) file. Relevant information regarding the QR's version is saved in a metafile. Finally, an MS Excel spreadsheet is

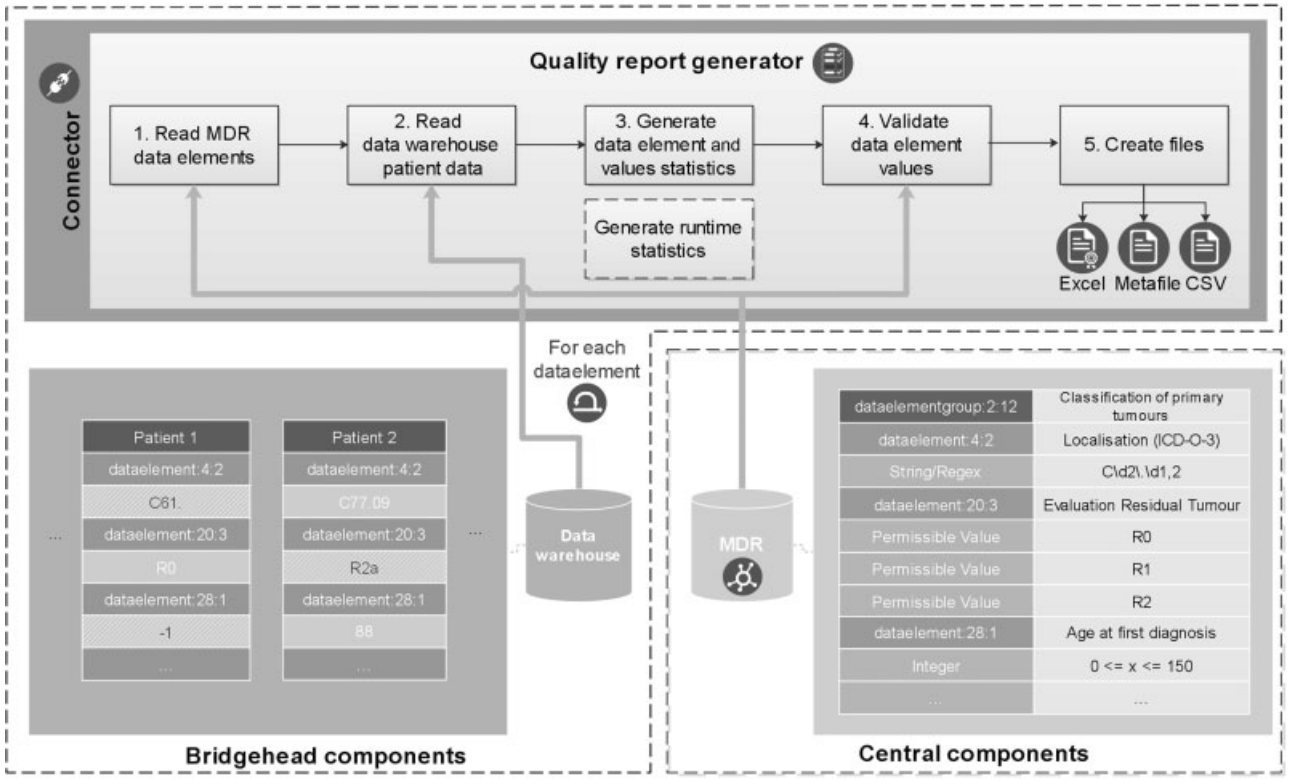

Fig. 3 Process for the generation of a quality report (QR) and the system components involved. Located in the Bridgehead's connector, the QR-generator retrieves data elements from the central MDR, validates corresponding values found in the data warehouse and compiles a spreadsheet-based QR from incorrect or incomplete values. In this example, shaded values are not among the permissible values and are therefore marked as a mismatch in the QR. MDR, metadata repository; QR, quality report.

created, in our implementation with the aid of the Java library Apache POI,[36] to facilitate evaluation by domain experts.

Practical Application within a Translational DRN

To evaluate our method, we developed a reference implementation within the German Cancer Consortium, a joint initiative involving leading academic research institutions and university hospitals.[37] Within the consortium, bridgeheads were installed at 10 sites and populated with clinical data.[29,38] We extended the bridgehead connectors with the QR-generator ("Samply.QA") which can now be initiated by the click of a button by the local staff. An example of a QR is shown in ►Fig. 2.

The process generates a QR in MS Excel format consisting of five sheets. The sheet "info" contains instructions for use, clarifications of columns and any general information or alerts important to the user. The sheet "all elements" contains the principal information of the report: Columns A to H (►Fig. 2A) provide a comparison of the actual data element values in the data warehouse against the data definitions retrieved from the MDR, allowing an assessment of the syntactic accuracy of the local data. Columns I to K provide statistical information. The table provides a separate row for each distinct value identified in the data warehouse for each data element included in the report. Table rows containing values with syntactic errors (i.e., values invalidated against the MDR definition) are shaded. The sheet "filtered elements" is a condensed version of "all elements." Clicking on the corresponding field "no. of patients …" redirects to the sheet "patient local ids" which contains information for identifying the dataset in question for manual correction in the source system(s) or in the ETL process. Lastly, the sheet "data elements stats" (►Fig. 2B) contains a further analysis of completeness and syntactic accuracy at data element level.

Discussion

Related Work

Within the data integration workflow of a DRN, the data quality assessment could be performed within several different components, as designated by numbers in ►Fig. 1, as follows:

- Many electronic data capture systems make it possible to define a wide range of checks for case report forms, which identify unlikely or implausible values before data are even stored[39–41]. For example, Fortier et al initially designed DataSHaPER[18] to provide standardized questionnaires for prospective harmonization. However, in our use case, this approach cannot be applied as we are not involved in the data entry process at all, but rather make use of data previously collected in primary systems outside of our control ("secondary use"). Similarly, Fortier et al found that such an a priori standardization "would be of limited applicability to retrospective harmonization"[18] and extended their platform with functionalities for retrospective harmonization.
- Data quality assurance is also possible during the ETL process.[42] Data integration applications, such as Talend Open Studio[43] or IBM Cognos Data Manager[44,45] allow validating data against some kind of external dictionary or metadata repository.[46,47] This approach, however, requires each individual partner site to implement their own solution compatible with the chosen data integration solution and the given infrastructure on site.
- Another approach consists of checking incoming data in the DRN's central applications. For example, a central database could reject uploads not compliant with the definitions deposited in the MDR. This, however, would require each central application to perform such a quality check individually, as opposed to a quality check undertaken only once in the bridgehead (see below). As soon as there are several central databases or the processes are designed without uploads to a central database, data quality checks at the central component level become impractical. In addition, doing the data quality assurance at this level would have to take place after pseudonymization or anonymization. This would make it infeasible to generate the sheet "patient local ids." As a result, the site would not be able to correct their ETL processes, mappings or data in the source systems.
- Lastly, data quality could be assessed after the ETL process within a bridgehead. There are several advantages with this approach as follows: (1) Data integration will "fail early"[48] at the first point at which the data are expected to conform to the commonly agreed data definitions; therefore, constituting a natural checkpoint for actual conformity. (2) The data in the bridgehead remains under local control, facilitating the handling of assessment reports, and the correction of errors, while circumventing data protection issues that might arise with uploading data to central resources or third parties. (3) Since data are expected to be loaded into the bridgehead in a harmonized manner, only one data quality assessment process needs to be implemented, as opposed to individual processes for individual source systems, or multiple processes for multiple subsequent analysis tools.

Within the bridgehead, there are two options as to where to perform the data quality assessment, in the data warehouse or within the connector. An example for the former option is PCORnet, while Achilles[46] and our proposed QR-generator implement the latter approach. Achilles follows the standard OMOP-CDM v4[46] and provides a well-established set of validation rules for data stored in the OMOP data model. As expected for a defined data model, this approach is very robust and successful, while the downside is less flexible. By contrast, validation against an MDR, as performed by the QR-generator, allows the evaluation of arbitrary data models, as long as they are modeled in the MDR.

Placing the QR-generator within the connector also gives the ability to choose among different data warehouse implementations. This is particularly important for the extensibility of the DRN through the bridgehead. This allowed the German Cancer Consortium, for example, to link 10 hospitals

with different biobank and tumor documentation systems as data sources (such as GTDS,[49,50] CREDOS,[51] CARAT,[52] CentraXX[53]) to a DRN. The option to choose the data warehouse technology to be deployed provides considerable flexibility. In addition, locating the QR-generator within the connector constitutes a convenient solution that is usable by all partners of a DRN, independent of the data infrastructure implemented on site.

The Advantages of Excel-Based Spreadsheets

There are several advantages to provide data quality reports as an MS Excel-based spreadsheet rather than a web-based interface within the connector. First, it cannot be assumed that personnel involved in improving data quality has access to the connector or data warehouse, as it contains sensitive patient-related data and may thus be located in a protected network. Second, persons unfamiliar with the interface may have trouble finding and understanding the information displayed. Third, connecting these people from different backgrounds and disciplines requires reports to be easily shareable and editable before passing them on, for example, to remove columns for reasons of data protection. In summary, we consider storing QRs in a widely-used, versatile format, such as MS Excel, the most straightforward way to support the various parties involved in improving data quality. In addition, the QR-generator provides a supplementary CSV file with the core data of the Excel-QR, which can easily be used for further computation with statistical analysis tools such as R.

Advantages of Open Source Software

Some providers of commercial data integration solutions also offer competitive data quality services: Gartner's "magic quadrant for data quality tools" lists the 15 most important ones.[54] Some of them, like Talend Open Studio,[43] are partially open-source, but reserve several services to their paying customers. For instance, "Talend Open Studio for data quality" does not make their QR available free of charge.[55]

Also, platform dependencies, as well as licensing issues, may impede integration of these features into DRNs building on open-source solutions. In the end, DRNs need to carefully consider the advantages of existing commercial tools against the benefits from consortium-wide, open-source mechanisms for data quality assurance.

Limitations and Outlook

Data quality assurance is a very broad topic and this work only scratches the surface. While we applied those metrics helpful to our specific DRN, there are many more to be considered.[14,56] The approach could be extended to allow running R-scripts from a secured script repository, allowing advanced users to evaluate any relevant metric. This can then be called upon by tools such as the QR-generator, incorporating the results of the scripts in an automated fashion. For example, consistency could be addressed by an implementation of the rules suggested by the European Network of Cancer Registries.[57] In addition, visual analytics techniques such as glyph-based variants[58] could help in identifying outliers and data anomalies even without a disease-specific ruleset.

It should also be noted that our approach focuses on evaluating harmonized data exported by the network partners into a bridgehead. It is not designed to assess data quality in the original source systems directly. But obviously, errors in the source system passed on to the bridgehead are identified through the QR and, as mentioned above, can prompt correction at source, contributing to better data quality overall.

Currently, the QRs of each of our DRN sites are collected regularly by a centrally coordinated team which analyzes each QR manually and returns a list of recommendations to each site, detecting and considering issues common to all sites. However, this process could be automated, as is done at other DRNs like PECARN,[59] saving time and avoiding human errors.

While the MDR provides a flexible way to define data elements at the atomic level, fast health care interoperability resources (FHIR)[60] goes further in structuring and linking data elements to form complex entities (resources like "patient," "procedure," and "observation") and "business objects" that reflect a particular clinical or biomedical reality. By specifying contextual rules and constraints, FHIR enables plausibility checks to be performed to improve data quality. The standard, although relatively young, shows potential, inter alia for structuring data and improving data quality. Therefore, we are now conducting feasibility studies and developing prototypes to evaluate possibilities for using and integrating FHIR into the workflow of data validation in data sharing.[61]

Conclusions

High-data quality is essential for using primary clinical data in secondary use research efforts but cannot be taken for granted given the different purposes for which the data were originally collected. Effective assessment of data quality is the first step toward improvement. In the context of a DRN, it can be addressed by a combination of integrated tools situated both centrally and at each partner site: a central metadata repository holds common data elements and value definitions which are used to validate the content of data warehouses operated at each site. This way, the consortium can work with standardized reports on data quality, while preserving the autonomy of each partner site. We have shown that data quality assessment performed within the bridgehead framework not only satisfies these requirements but also enables the individual sites to improve data in their local source systems.

Conflicts of Interest

None declared.

Acknowledgments

The authors would like to acknowledge the valuable collaboration with their partners in the German Cancer Consortium (DKTK), in particular Barbara Uhl and Kristina Ihrig of the Office of the Clinical Communication Platform (CCP Office) and all consortium sites for their helpful feedback. They also thank Christian Koch for the development of the MDR-Client, and Melanie Forche and Janine Al-Hmad for their support during the graphics design process. They are grateful to David Croft for checking grammar and language.

References

1 Zatloukal K, Stumptner C, Kungl P, Mueller H. Biobanks in personalized medicine. Expert Rev Precis Med Drug Dev 2018;3(04): 265–273

2 Ginsburg GS, Phillips KA. Precision medicine: from science to value. Health Aff (Millwood) 2018;37(05):694–701

3 Tsimberidou AM, Ringborg U, Schilsky RL. Strategies to overcome clinical, regulatory, and financial challenges in the implementation of personalized medicine. Am Soc Clin Oncol Educ Book 2013:118–125

4 Abrahams E, Ginsburg GS, Silver M. The Personalized Medicine Coalition: goals and strategies. Am J Pharmacogenomics 2005;5 (06):345–355

5 Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014;21(04):578–582

6 Litton J-E. Launch of an Infrastructure for Health Research: BBMRI-ERIC. Biopreserv Biobank 2018 (e-pub ahead of print). Doi: 10.1089/bio.2018.0027

7 Biobanking And Biomolecular Resources Research Infrastructure —European Research Infrastructure Consortium. Available at: http://www.bbmri-eric.eu/. Accessed July 11, 2019

8 Durinx C, McEntyre J, Appel R, et al. Identifying ELIXIR core data resources. F1000 Res 2016;5(ELIXIR):2422–2439

9 ELIXIR. A distributed infrastructure for life-science information. Available at: https://www.elixir-europe.org. Accessed July 11, 2019

10 Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018

11 Daniel C, Choquet R. Clinical research informatics: contributions from 2016. Yearb Med Inform 2017;26(01):209–213

12 Kodra Y, Weinbach J, Posada-de-la-Paz M, et al. Recommendations for improving the quality of rare disease registries. Int J Environ Res Public Health 2018;15(08):E1644

13 Hewitt RE. Biobanking: the foundation of personalized medicine. Curr Opin Oncol 2011;23(01):112–119

14 Batini C, Scannapieco M. Data and Information Quality: Dimensions, Principles and Techniques. Switzerland: Springer International Publishing; 2016

15 Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: An American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007;14(01):1–9

16 Dugas M, Neuhaus P, Meidt A, et al. Portal of medical data models: Information infrastructure for medical research and healthcare. Database (Oxford) 2016;2016;. Doi: 10.1093/database/bav121

17 Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(01):1244

18 Fortier I, Burton PR, Robson PJ, et al. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. Int J Epidemiol 2010;39(05):1383–1393

19 Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(02):124–130

20 i2b2 Cell Messaging: Data Repository (CRC) Cell. Available at: https://www.i2b2.org/software/files/PDF/current/CRC_Messaging.pdf. Accessed October 22, 2018

21 Timbie J, Rudin R, Towe V, et al. National Patient-Centered Clinical Research Network (PCORnet) Phase I: Final Evaluation Report. Santa Monica, CA: RAND Corporation; 2015

22 Lablans M, Kadioglu D, Mate S, Leb I, Prokosch H-U, Ückert F. [Strategies for biobank networks. Classification of different approaches for locating samples and an outlook on the future within the BBMRI-ERIC]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2016;59(03):373–378

23 Mate S, Kadioglu D, Majeed RW, et al. Proof-of-concept integration of heterogeneous biobank it infrastructures into a hybrid biobanking network. Stud Health Technol Inform 2017;243: 100–104

24 Kimball R, Caserta J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Indianapolis, IN: Wiley Publishing Inc.; 2009

25 Embley DW, Thalheim B, Eds. Handbook of Conceptual Modeling: Theory, Practice, and Research Challenges. Berlin,Germany: Springer-Verlag; 2011

26 Leser U, Naumann F. Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen. 1st ed. Heidelberg: dpunkt-Verlag; 2007

27 Weber GM, Murphy SN, McMurry AJ, et al. The shared health research information network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc 2009;16(05):624–630

28 i2b2: Informatics for Integrating Biology & Bedside. Data Sharing Network (SHRINE). Available at: https://www.i2b2.org/work/shrine.html. Accessed October 22, 2018

29 Lablans M, Kadioglu D, Muscholl M, Ückert F. Exploiting Distributed, Heterogeneous and Sensitive Data Stocks while Maintaining the Owner's Data Sovereignty Methods. Inf Med 2015;54(04): 346–352

30 Garde S, Knaup P, Hovenga E, Heard S. Towards semantic interoperability for electronic health records. Methods Inf Med 2007; 46(03):332–343

31 Kadioglu D. Institutionsübergreifende Nutzung Verteilter Metadata Repositories. [Master Thesis]. Dortmund: Fachhochschule Dortmund; 2013

32 Kadioglu D, Breil B, Knell C, et al. Samply.MDR - a metadata repository and its application in various research networks. Stud Health Technol Inform 2018;253:50–54

33 Kadioglu D, Weingardt P, Ückert F, Wagner T. Samply.MDR–Ein Open-Source-Metadaten-Repository. HEC 2016: Health—Exploring Complexity 2016 Joint Conference of GMDS, DGEpi, IEA-EEF, EFMI. 2016

34 Ulrich H, Kock A-K, Duhm-Harbeck P, Habermann JK, Ingenerf J. Metadata repository for improved data sharing and reuse based on HL7 FHIR. Stud Health Technol Inform 2016;228:162–166

35 Guenther A, Nowak I, Pertz J, Sirman G Qualitätsbewertung von Routinedaten zur Sekundärdatenanalyse in der medizinischen Forschung mdi Forum der Medizin_Dokumentation und Medizin_Informatik 2016;Heft 2

36 Apache POI. Apache POI–the Java API for Microsoft Documents. Available at: https://poi.apache.org/. Accessed October 22, 2018

37 German Cancer Consortium. Available at: https://dktk.dkfz.de/en/home. Accessed July 11, 2019

38 Lablans M, Schmidt EE, Ückert F. An architecture for translational cancer research as exemplified by the German Cancer Consortium. JCO Clin Cancer Inform 2017;(01):1–8

39 Couchoud C, Lassalle M, Cornet R, Jager KJ. Renal replacement therapy registries–time for a structured data quality evaluation programme. Nephrol Dial Transplant 2013;28(09): 2215–2220

40 Baigent C, Harrell FE, Buyse M, Emberson JR, Altman DG. Ensuring trial validity by data quality assurance and diversification of monitoring methods. Clin Trials 2008;5(01):49–55

41 Venet D, Doffagne E, Burzykowski T, et al. A statistical approach to central monitoring of data quality in clinical trials. Clin Trials 2012;9(06):705–713

42 Berndt DJ, Fisher JW, Hevner AR, Studnicki J. Healthcare data warehousing and quality assurance. Computer 2001;34(12): 56–65. Available at: https://ieeexplore.ieee.org/document/970578

43 Talend Open Studio. Open source integration software. Available at: https://www.talend.com/products/talend-open-studio. Accessed October 22, 2018

44 Corp IBM. The Role of Data Quality in BI and Performance Management 2008. Available at: ftp://public.dhe.ibm.com/software/data/sw-library/cognos/pdfs/whitepapers/wp_the_role_of_data_quality_in_bi_and_performance_management.pdf. Accessed October 22, 2018

45 IBM Cognos Data Manager. Available at: https://www.ibm.com/support/knowledgecenter/en/SSEP7J_10.1.1/com.ibm.swg.ba.cognos.ug_ds.10.1.1.doc/c_introducingdecisionstream.html. Accessed July 11, 2019

46 Daniel C, Sinaci A, Ouagne D, et al. Standard-based EHR-enabled applications for clinical research and patient safety: CDISC - IHE QRPH - EHR4CR & SALUS collaboration. AMIA Jt Summits Transl Sci Proc 2014;2014:19–25

47 Choquet R, Qouiyd S, Ouagne D, et al. The information quality triangle: a methodology to assess clinical information quality. Stud Health Technol Inform 2010;160(Pt. 1):699–703

48 DiPiro JT, Chisholm-Burns MA. Fail fast. Am J Pharm Educ 2013;77 (08):159

49 Altmann U, Dudeck J. The Giessen tumor documentation system (GTDS)–review and perspectives. Methods Inf Med 2006;45(01): 108–115

50 GTDS. Gießener Tumordokumentationssystem. Available at: http://www.med.uni-giessen.de/akkk/gtds/gtdsna1d.htm. Accessed October 22, 2018

51 Universitäts Klinikum Ulm. CREDOS (Cancer Retrieval Evaluation and DOcumentation System). Available at: https://www.uniklinik-ulm.de/comprehensive-cancer-center-ulm-cccu/klinisches-krebsregister/software/credos-tumordokumentation.html. Accessed October 22, 2018

52 Universitäts Klinikum Freiburg. CARAT—die CCCF Anwendung zur Registrierung und Auswertung von Tumordaten. Available at: https://www.uniklinik-freiburg.de/cccf/aerzte-fachleute/krebs-register-it/carat-erfassungssystem.html. Accessed October 22, 2018

53 Kairos GmbH. Centraxx (Official Webseite). Available at: https://www.kairos.de/en/products/centraxx/. Accessed July 11, 2019

54 Selvage M, Judah S, Jain A. Magic quadrant for data quality tools. Available at: https://www.gartner.com/doc/3818863/magic-quadrant-data-quality-tools. Accessed October 22, 2018

55 Run jobs and publish/export results of analysis in Talend Open Studio. Available at: https://community.talend.com/t5/Data-Quality-Preparation-and/Run-jobs-and-publish-export-results-of-analysis-in-Talend-Open/m-p/7390#M38. Accessed October 22, 2018

56 Nonnemacher M, Nasseh D, Stausberg J. Datenqualität in der medizinischen Forschung: Leitlinie zum adaptiven Management von Datenqualität in Kohortenstudien und Registern. 2., aktualisierte und erweiterte Auflage 2014. Schriftenreihe der TMF–Technologie-und Methodenplattform für die vernetzte medizinische Forschung e.V

57 European Network of Cancer Registries. Recommendations issued by ENCR. Available at: https://www.encr.eu/working-groups-and-recommendations. Accessed October 22, 2018

58 Müller H, Reihs R, Zatloukal K, et al. State-of-the-Art and Future Challenges in the Integration of Biobank Catalogues. In: Holzinger A, Rocker C, Ziefle M, eds. State-of-the-Art and Future Challenges in the Integration of Biobank Catalogues: Open problems and future challenges. Vol. 8700. Cham: Springer; 2015:261–273

59 Deakyne Davies SJ, Grundmeier RW, Campos DA, et al; Pediatric Emergency Care Applied Research Network. The pediatric emergency care applied research network registry: a multicenter electronic health record registry of pediatric emergency care. Appl Clin Inform 2018;9(02):366–376

60 Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. In: Rodrigues PP, ed. IEEE 26th International Symposium on Computer-Based Medical Systems (CBMS), 2013, University of Porto, Portugal. Piscataway, NJ: IEEE 2013:326–331

61 Kern J, Boeker M, Brucker DP, et al. Engineering a data model for distributed research networks in Oncology based on FHIR 64. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS) 2019

Appendices

Source Code (JAVA): https://bitbucket.org/medicalinformatics/samply.share.client