German Medical Data Sciences 2023 — Science. Close to People.
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI230690

Reusing Biomedical Data as Agreed - Towards Structured Metadata for Data Use Agreements

Caroline BÖNISCH[a1], Sabine HANẞ[a,b], Nicolai SPICHER[a,b,c], Ulrich SAX[a,c], and Dagmar KREFTING[a,b,c]

[a] *Department of Medical Informatics, University Medical Center Göttingen, Germany*
[b] *German Centre for Cardiovascular Research, Partner Site Göttingen, Germany*
[c] *Campus Institute Data Science, Georg-August-University Göttingen, Germany*

Abstract. Introduction With increasing availability of reusable biomedical data – from cohort studies to clinical routine data, data re-users face the problem to manage transferred data according to the heterogeneous data use agreements. While structured metadata is addressed in many contexts including informed consent, contracts are to date still unstructured text documents. In particular within collaborative and active working groups the actual usage agreement*'s regulations* are highly relevant for the daily practice – can I share the data with colleagues from the same university or the same research network, can they be stored on a PHD student's laptop, can I store the data for further approved data usage requests? **Methods** In this article, we inspect and review seven different data usage agreements. We focus on digital data that is copied and transferred to the requester's environment. **Results** We identified 24 metadata items in the four main categories *data usage, storage*, and *sharing*, as well as *publication of results*. **Discussion** While the topics are largely overlap in the data use agreements, the actual regulations of the topics are diverse. Although we do not explicitly investigate trusted research environments, where data is offered within an analytics platform, we consider them a as subgroup, where most of the practical questions from the data scientist's perspective also arise. **Conclusion** With a limited set of structured metadata items, data scientists could have information about the data use agreement at hand along with the transferred data in an easily accessible way.

Keywords. data reuse, metadata, data usage agreement, contract

1. Introduction

The FAIR Guiding Principles for research data stewardship — demanding findability, availability, interoperability and reusability of research data [1] — are increasingly implemented into data generating biomedical research projects.

While cohorts and registries have the main aim to be used continuously by researchers for new research questions, re-use of other biomedical data collections, such as clinical trials or healthcare data in the inpatient or even outpatient context, are also of high interest for data scientists, as they provide complementary data relevant e.g. for

[1] Corresponding Author, Caroline Bönisch, Department of Medical Informatics, University Medical Center Göttinge, Robert-Koch-Straße 40, 37075 Göttingen, Germany; E-mail: caroline.boenisch@med.uni-goettingen.de

bias-free machine learning models. Today there are several tools available that support FAIRification of different biomedical data sources, for example in the case of clinical routine data [2], or for patient reported outcome measures in pandemic apps [3].

As a consequence of the reproducibility crisis, prediction and classification methods typically need to be validated with different data sets from different locations to ensure results reproducibility as discussed in [4-6]. For example, the United States National Institute of Health released their Policy for Data Management and Sharing at the beginning of this year emphasizing the need for agreements on the topic of sharing biomedical data [7].

Table 1: Selected metadata items of NFDI4Health metadata Version 3.0 regarding licensing, data access and data use. Items solely defining semantic subgroups are not shown.

ID	Metadata item	Description
1.10.3.1	License	License defining the rights to (re-)use the [RESOURCE].
1.10.3.2.1	Do you confirm that the authors have authority to license the [RESOURCE]?	Confirmation that the authors have authority to license the resource.
1.10.3.2.2	Do you confirm that the authors have read and understood the terms of the chosen license?	Confirmation that the authors have read and understand the terms of the chosen license.
1.10.3.2.3	Do you confirm that the authors understand that Creative Commons (CC) licensing is irrevocable?	Confirmation that the authors understand that Creative Commons (CC) licensing is not revocable.
1.10.3.2.4	Do you confirm that the authors allow NFDI4Health to license the document?	Confirmation that NFDI4Health is allowed to license the document, i.e. to mark the resource with the license information.
1.10.3.3	Additional information about use rights	Any additional descriptive information explaining terms and conditions to (re-)use the resource.
1.17.35.1	Is it planned to share the data?	Indication whether there is a plan to make data collected in the study available. In case of studies with patients or other individuals, this refers to individual participant data (IPD).
1.17.35.3	When and for how long will the data be available?	Indication when the data and, if applicable, supporting documents will become available and for how long.
1.17.35.4	Criteria for the data access	Indication by what access criteria data will be shared, including a) with whom, b) for what types of analyses, and c) by what mechanism.
1.17.35.5	Additional information about data sharing	Additional descriptive information providing more details about the data sharing, e.g. indication what data in particular will be shared or why the data will not be shared or why it is not yet decided.
1.17.35.7	Web page with additional information about data sharing	If existing, a link to the web page where additional information about data sharing can be found.

Options for access of individual patient data have increased for data scientists in the last decade. However, managing transferred data sets from different providers at the re-user's responsibility gets more and more complex while complying with heterogeneous use and publication regulations. To make it easier to assess the different provider's regulations on data handling we see the need to provide these regulations in a machine-readable and actionable way. This would make it possible to implement automatic rules for internal data access, to remind the user automatically to send results and publications back to the data provider and delete data on time if mandatory.

Metadata items on data sharing policies and in general data sharing schemas such as DataCite [8] or Dublin Core Metadata initiative (DCMI) [9] are typically coarse-grained, for example the metadata “data access” is limited to open or restricted access and not further subdivisible.

Different initiatives have taken the task to define and use metadata items describing data use agreements (DUA). For example, the National Research Data Infrastructure for Personal Health Data (NFDI4Health) has defined a metadata schema for cohort studies with multiple metadata elements as part of its task objectives [10]. This schema comprises 184 elements in the current version 3, created in collaboration with various participants and University Medical Centers in Germany. It encompasses also metadata referring to licensing and use rights, c.f. Table 1. However, DUA that are not covered by a Creative Commons license are currently envisioned to be described in a single text field (ID 1.10.3.3).

On the other hand, the Medical Data Integration Center of the University Medical Center Göttingen (UMG-MeDIC) recently published metadata, identified as relevant in order to operate data re-use and transfer of clinical care data in an efficient manner, c.f. Table 2. The metadata schema version 1.0 within the UMG-MeDIC encompasses provenance information like referenced data or information about the source system of the derived data, as well as information about licensing, consent and data usage [11].

Table 2. UMG-MeDIC metadata version 1.0 of clinical care data regarding licensing, data access and data usage.

Metadata item	Description
UsageLicense/Copyright	License, respective copyright of the data
UsageContext	Context in which the data can be used
VestingPeriod	Period, in which the data is locked due to study regulations
ConsentType	Type of consent for the data
ConsentValidation	Validity of the Consent
ConsentVerification	Date of when the consent became verified
ConsentModule	Specific area/question the patient consented to

Furthermore, these metadata sets have been defined along the requirements of the data providers who need to assess for example compliance of data sharing with informed consent or embargo phases according to study policies. In this paper, we explicitly take the perspective of the data scientist who needs to handle the diverse DUAs of a growing number of data providers. A recent study on data sharing policies in neuroimaging data repositories analyses DUAs of seven data providers and categorize them by the presence of aspects in the DUA that refer to obligations of the data user: (a) Prohibition on re-identifying subject, (b) Limitations on further disclosure or use of data, (c) Security measures in place, (d) Acknowledgment of data repository as data source, (e) Report research use of data upon request, (f) Report of violation [12].

However, to our knowledge no overarching ontology or data model yet exists for structured representation of DUA. Therefore, the objective of this work is to deduce a set of metadata items describing the aspects of DUA in a way that it helps data users to comply with them.

2. Methods

Starting with the categories defined by [12], we analyse the statements found in DUAs and related documents of seven existing biomedical data providers and map them to common metadata items and value sets. We focus on non-public data sources, where data is transferred to the data user upon the acceptance of DUAs, publication policies or other contracts. We cover different data creation contexts – cohorts, clinical trials, and health care data. We furthermore focus on digital data, as these would be in the main interest of data scientists. Therefore, we did not investigate compliance rules that are exclusively relevant for material transfer of biospecimen. We selected the following research data infrastructures, ordered by the year of founding and examined the latest versions of the respective DUA or further documents for this research project, indicated in parentheses.

The *SHIP study* of Health in Pomerania encompasses 3 population based cohorts where the first cohort started in 1997, and subsequent cohorts were recruited in 2008 and 2021, respectively [13]. Each cohort encompasses about 4000 participants living in Pomerania with comprehensive clinical phenotyping including medical imaging, dental examination and biospecimen. DUAs (Version 03.07.2012) are available in German on the website.

The *UK Biobank* is - according to their own website - "the most detailed, long-term prospective health research study in the world" [14, 15]. Starting in 2006, it now provides comprehensive digital data such as clinical data and imaging along with biospecimen from about 500.000 participants from the UK. On their website, detailed information is found on the study design and the usage agreements (Version 1.2). UK biobank offers both data to be transferred and data to be processed within their trusted research environment.

The *DZHK Heartbank* of the German Centre for Cardiovascular Research (DZHK) was founded in 2012. Since then, it provides a so-call clinical study platform, where the clinical data, biospecimen and imaging data of all clinical studies - both clinical cohorts and clinical trials - that are fully funded by the DZHK, are provided for reuse after a certain period of exclusive access by the study PIs [16]. Both clinical data items and biospecimen collection are harmonized throughout all studies, and imaging data is centrally quality assured. The so-called DZHK base dataset - comprising 42 clinical data items - are available for all quality assured data without embargo and can be searched in a publicly available feasibility explorer [17]. DUA (Version 03-2021) as well as publication guidelines (Version 11-2022) are available in German and English on the DZHK website [18].

ClinicalStudyDataRequest.com is — according to the website — "a consortium of clinical study Sponsors*"*. To date, 3046 Studies are listed for data sharing [19]. 12 data sponsors are listed, including Bayer, Novartis and Teva, among others. Research Proposals undergo a stepwise review process, with an Independent Review Panel organized by the Welcome Trust as the last step after review by the study Sponsors. DUAs (Version 04/15/15) are available from the website.

The *Yale University Open Data Access* (YODA) Project has "facilitated access to clinical trial data since 2013" [20]. The website lists currently 45 clinical trials from three data partners [21]. The data can be searched according to the substance, the health condition and further filters such as mean age, number of enrollment, among others. The project not only provides the DUA (Version February 2019) online, but offers a self-training on their data usage agreement describing scenarios that need to be rated as compliant or violating.

The German *Medical Informatics Initiative* has been founded in 2016 and aims at making health care data created in the university hospitals available for research. Each university hospital is currently building a so-called data integration center, where data from the different primary hospital information systems is collected and provided as an interoperable data set. Data can be requested through the research data portal for health [22]. Currently only members of participating institutions can request data, but it is envisioned to open it for external researchers after the current test phase. The DUA (Version 1.1) is available in German from the website. [23]

The *NAPKON* German National Pandemic COVID-19 Cohort Network has been founded 2020 as part of the Network University Medicine and conducts three cohorts of COVID-19 patients from different base populations and with different levels of phenotyping [24]. To date, over 97.000 visits including imaging and 36.000 primary biosamples from about 7.000 patients are currently available for reuse. The DUA (Version 3.0) is available in German on the website [25].

3. Results

In Table 3, we describe the metadata items of the main topics that we found in the DUAs and accompanying policies across the seven data platforms and infrastructures. These topics comprise mainly the areas of data usage (dus) - requirements and regulations on how data can be utilized, data storage (dst) - information on how to archive the data, data sharing (dsh) - regulations on how the data can be exchanged and publication of results (pur) - requirements for the publication of results with references or publication policies to be considered.

The table contains 24 metadata items describing DUAs as derived from the various agreements and regulations. Each of these metadata items is additionally marked with *Not Applicable* and *Unknown* to allow mapping of the metadata, even if no information is available. We found large overlap in the topics of the DUAs, and many of the regulations could be mapped to a few values. Most of them could be easily modeled as structured single or multiple answer items. However, we allowed in particular text fields for specific regulations. For example, in two of the DUA we found pre-formulated acknowledgments that must be used in publications. We included the items for textual parts of the DUA, where we expect them to be short and useful to have them at hand when using the data. In case of accidental re-identification or a security incident, which may induce certain stress to the researcher, it is important to have the policies and contact information quickly available.

Furthermore, the licensing of the data also plays an important role within the metadata items describing DUA. NFDI4Health, for example, provides all information under Creative Commons License in different subtypes [26] or as all righs reserved. In Table 3, the licensing is taken into account under data license to applicant and the license text in results license text.

4. Discussion

The aim of the presented work was to establish a common set of metadata items describing DUA relevant for researchers. The comparison of the DUAs shows that there are largely overlapping topics, but substantial differences in the policies, resulting in a small set of common metadata items describing DUAs with up to five value set items for each of the metadata item with a single answer option (e.g. withdrawal policies - data deletion upon notification; other; no; not applicable; unknown). Each of the providers has its own requirements for the implementation of data re-use, licensing, publication of their collected data and the analysis. However, it could be shown that a common denominator between these providers are possible and can be comprised into 24 metadata items.

Table 3. Derived metadata items of the comparative comparison of the seven biomedical research data providers with corresponding value sets. Grp = Semantic groups of main topics: dus - data usage, dst - data storage, dsh - data sharing, pur - publication of results. Ctg = Categories defined by Jwu et al. Due to space limits, default values such as not applicable or unknown are not mentioned.

Grp	Metadata item	Values	Notes	Ctg
dus	Purpose restriction	Permitted Purpose only, other Purposes	single answer	b
dus	Related Purpose	String containing the title of project from the data request application	string	b
dus	Timespan of data usage	Final date of usage	date	b
dst	Data deletion required	Yes with proof, Yes without proof, No	single answer	b
dst	Data deletion proof	Deletion policy text	string	b
dus	Data license to applicant	revocable, worldwide, exclusive, transferable, royalty-free	multiple answer	b
dst	Data copy restriction	Yes, No	single answer	b
dsh	Sub-sharing of data	individual applicant only, named collaborators, members of data user's institution, unrestricted	multiple answer	b
dus	Withdrawal policies	data deletion upon notification, other, no	single answer	b
dus	Withdrawal notification channel	email-address or other contact method agreed upon in the agreement	string	b
pur	License to data provider	Yes, No	single answer	e
pur	Results license text	string containing the license text	string	e
pur	Publication policies	acknowledgment, mandatory citation, requested citation, manuscript review, coauthors, notification, other, none	multiple answer	d
pur	Acknowledgment	Acknowledgment text	string	d
pur	Related citations	DOI of publication	DOI	d
pur	ORCID-Coauthors	ORCID	ORCID	d
pur	non-ORCID coauthors	Name, Affiliation	string	d
pur	Manuscript review contact	email-address or other contact method agreed upon in the agreement	string	e
dus	Re-identification	Explicitly prohibited, otherwise specified, Allowed	single answer	a
dus	Re-identify policy	Text of policies on re-identification	string	a
dus	Re-contacting	Explicitly prohibited, otherwise specified, Allowed	single answer	a
pur	Results vesting period	Final date of blocked data	date	e
dus	Security incidence policy	Yes, No	single answer	f
dus	Security incidence contact	email-address or other contact method agreed upon in the agreement	string	f

The findings are only at first sight solely of interest for researchers that re-use data from more than one data provider. The current draft of the European Health Data Space explicitly mandates all health data holders to provide their data for re-use. It is very likely, that small-sized data holders delegate this to larger institutions that in turn form larger data sharing networks that need to provide information about the available data to the national data provision node [27]. The same holds for multiscale networks of data providers require an abstracted and generic method of data access points and associated metadata. This information must be as broad as necessary, but also as granular and structured as possible to be automatically processed.

Metadata items describing DUAs can have a positive impact on the implementation of FAIR Guiding Principles, DataCite or Dublin Core Metadata Initiative. One important aspect of making data accessible and reusable is to provide standardized information about the data usage agreements and restrictions. By harmonizing, data repositories and data providers can provide consistent information, thereby enhancing the findability and accessibility of the data.

The results obtained in the course of this work are subject to ongoing changes in requirements and adaptations. Limitations are the selection of the investigated infrastructures, as they might not been representative. Newly added data infrastructures and data provider, with their own data usage agreements, must be taken into account in the ongoing process. In addition, external changes — such as the European Health Data Space or the announced German regulations on health data use [28, 29] — must also be included. Next steps are to bring up this topic into the different initiatives and data providers and data re-users.

Acknowledgement

The authors declare that there is no conflict of interest.

This work was partially supported by the German Federal Ministry of Education and Research through DZHK (81X1300101), FAIRRMeDIC (01ZZ2012), HiGHmed (01ZZ1802B) and Netzwerk Universitätsmedizin (NUM – 01KX2121), German Research Foundation through NFDI4health COVID-19 (451265285) and NFDI4health (442326535), and Ministry for Science and Culture of Lower Saxony through the Junior Research Group Complex Biosignal Processing.

Author Contributions: Conceptualization (CB, SH, DK, US, NS), investigation (CB, DK), writing - original draft (CB, DK), writing - review and editing (CB, SH, DK, US, NS), supervision (DK), funding acquisition (SH, DK, US).

No ethical approval was needed for the execution of this research.

References

[1] Wilkinson M, Dumontier M, Aalbersberg I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, doi:10.1038/sdata.2016.18. 160018 (2016).

[2] Parciak M, Bender T, Sax U, Bauer CR. Applying FAIRness: Redesigning a Biomedical Informatics Research Data Management Pipeline. Methods Inf Med. 2019 Dec;58(6):229-234. doi: 10.1055/s-0040-1709158. Epub 2020 Apr 29.

[3] Muzoora MR, Schaarschmidt M, Krefting D, Oehm J, Riepenhausen S, Thun S. Towards FAIR Patient Reported Outcome: Application of the Interoperability Principle for Mobile Pandemic Apps. Stud Health Technol Inform. 2021 Nov 18;287:85–6. doi: 10.3233/SHTI210820

[4] Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Science translational medicine. 2016;8(341):341ps12-341ps12. doi: 10.1126/scitranslmed.aaf5027.

[5] Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten Simple Rules for Reproducible Computational Research. PLoS Comput Biol. 2013 9(10): e1003285. doi: 10.1371/journal.pcbi.1003285

[6] Stodden V. Reproducible Research: Tools and Strategies for Scientific Computing, Computing in Science & Engineering. 2012;14, 11-12. doi: 10.1109/MCSE.2012.82

[7] Kaiser J, Brainard J. Ready, set, share: Researchers brace for new data-sharing rules. Science. 2023;379(6630):322–5. doi: 10.1126/science.adg8142.

[8] DataCite Metadata Working Group. (2021). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.4. DataCite e.V. doi: 10.14454/3w3z-sa82

[9] Dublincore.org. dublincore.org [Internet]. [cited 2023 May 10]. Available from: https://www.dublincore.org/specifications/dublin-core/dces/

[10] Schmidt, CO, Fluck J, Golebiewski M. et al. COVID-19-Forschungsdaten leichter zugänglich machen – Aufbau einer bundesweiten Informationsinfrastruktur. Bundesgesundheitsbl 64, 1084–1092 (2021). doi: 10.1007/s00103-021-03386-x

[11] Bönisch C, Kesztyüs D, Kesztyüs T. Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata. Sci Data 9, 659 (2022). doi: 10.1038/s41597-022-01792-7

[12] Jwa AS, Poldrack RA. The spectrum of data sharing policies in neuroimaging data repositories. Hum Brain Mapp. 2022 Jun 1;43(8):2707-2721. doi: 10.1002/hbm.25803. Epub 2022 Feb 10.

[13] Völzke H. Study of Health in Pomerania (SHIP). Bundesgesundheitsbl. 55, 790–794 (2012). doi: 10.1007/s00103-012-1483-6

[14] Ukbiobank.ac.uk. Ukbiobank.ac.uk [Internet]. [cited 2023 February 1] Available from. https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us

[15] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 2015 Mar 31;12(3):e1001779. doi: 10.1371/journal.pmed.1001779

[16] Hoffmann J, Hanß S, Kraus M, Schaller J, Schäfer C, Stahl D et al. The DZHK research platform: maximisation of scientific value by enabling access to health data and biological samples collected in cardiovascular clinical studies. Clin Res Cardiol. 2023 Mar 8. doi: 10.1007/s00392-023-02177-5.

[17] Scheel H, Dathe H, Franke T, Scharfe T, Rottmann T. A Privacy Preserving Approach to Feasibility Analyses on Distributed Data Sources in Biomedical Research. Stud Health Technol Inform. 2019 Sep 3;267:254–61. doi: 10.3233/SHTI190835.

[18] DZHK.de. Dzhk.de [Internet]. [cited 2023 May 7] Available from: https://dzhk.de/en/

[19] ClinicalStudyDataRequest.com. ClinicalStudyDataRequest.com [Internet]. [cited 2023 March 31]. Available from: https://clinicalstudydatarequest.com/

[20] Ross JS, Waldstreicher J, Bamford S, Berlin JA, Childers K, Desai NR, et al. Overview and experience of the YODA Project with clinical trial data sharing after 5 years. Sci Data. 2018 Nov 27;5:180268.

[21] Yoda.yale.edu. Yoda.yale.edu [Internet]. [cited 2023 May 7] Available from: https://yoda.yale.edu/browsetrials/generic-name

[22] Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med. 2018 Jul;57(S 1):e50–6. doi: 10.3414/ME18-03-0003

[23] Geschäftsstelle TMF e.V., Forschungsdatenportal für Gesundheit [Internet]. [cited 2023 Mar 31]. Available from: https://www.forschen-fuer-gesundheit.de/index.php

[24] Schons M, Pilgram L, Reese JP, Stecher M, Anton G, Appel KS, et al. The German National Pandemic Cohort Network (NAPKON): rationale, study design and baseline characteristics. Eur J Epidemiol [Internet]. 2022 Jul 29; doi: 10.1007/s10654-022-00896-z

[25] Napkon.de. Napkon.de [Internet]. [cited 2023 May 7] Available from: https://napkon.de/use-and-access/

[26] Creativecommons.org. creativecommons.org [Internet]. [cited 2023 May 10] Available from: https://creativecommons.org/about/cclicenses/

[27] Gudi N, Kamath P, Chakraborty T, Jacob AG, Parsekar SS, Sarbadhikari SN, John O. Regulatory Frameworks for Clinical Trial Data Sharing: Scoping Review.J Med Internet Res 2022;24(5):e33591.doi: 10.2196/33591

[28] Proposal for a regulation - The European Health Data Space [Internet]. [cited 2023 March 31]. Available from: https://health.ec.europa.eu/publications/proposal-regulation-european-health-data-space_en

[29] Digitalisierungsstrategie vorgelegt [Internet]. [cited 2023 March 31]. Available from: https://www.bundesgesundheitsministerium.de/presse/pressemitteilungen/digitalisierungsstrategie-vorgelegt-09-03-2023.html