Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230242

Definition, Composition, and Harmonization of Core Datasets Within the German Center for Lung Research

Mark R. STÖHR [a,1], Andreas GÜNTHER [a] and Raphael W. MAJEED [a,b]
[a] *UGMLC, German Center for Lung Research (DZL), Justus-Liebig-University, Giessen, Germany*
[b] *Institute of Medical Informatics, Medical Faculty of RWTH Aachen, Aachen, Germany*

Abstract. Core datasets are the composition of essential data items for a certain research scope. As they state commonalities between heterogeneous data collections, they serve as a basis for cross-site and cross-disease research. Therefore, researchers at the national and international levels have addressed the problem of missing core datasets. The German Center for Lung Research (DZL) comprises five sites and eight disease areas and aims to gain further scientific knowledge by continuously promoting collaborations. In this study, we elaborated a methodology for defining core datasets in the field of lung health science. Additionally, through support of domain experts, we have utilized our method and compiled core datasets for each DZL disease area and a general core dataset for lung research. All included data items were annotated with metadata and where possible they were assigned references to international classification systems. Our findings will support future scientific collaborations and meaningful data collections.

Keywords. Data collection, datasets as topic, quality indicators, respiratory system, controlled vocabulary.

1. Introduction

One fundamental step towards successful cross-site and cross-disease research is the agreement on common fields of interest. These fields need to be identified and specified by naming and describing their essential data items.

The German Center for Lung Research (DZL) aims for high-impact translational research involving five participating sites as well as associated partners. Over 300 principal investigators contribute to eight disease areas: asthma and allergy, acute respiratory distress syndrome, cystic fibrosis, chronic obstructive pulmonary disease, diffuse parenchymal lung disease, end stage lung disease, lung cancer, and pulmonary hypertension. For advanced data science and collaborations, a central access point for data analysis and feasibility queries was established in 2016. Data from various heterogeneous sources is harmonized and integrated in a central data warehouse, currently containing over 60 individual data sources. Yet, the overlap of common parameters is low. In consequence, the DZL data warehouse does not sufficiently support cross-domain data analysis.

To achieve meaningful data for cross-site and cross-disease research, a common approach is to concentrate on parameters that promise to have the highest impact in terms

[1] Corresponding Author: M.R. Stöhr, E-mail: mark.stoehr@innere.med.uni-giessen.de.

of overarching availability and research relevance. We refer to them as "core parameters" that are collected and structured in "core datasets". Core datasets – by definition – contribute to the fulfilling of the FAIR principles [1], especially the reusability aspect of meeting domain-relevant community standards. The third FAIR principle "interoperability" can be achieved by referring to international terminologies like the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT).

Several national and international projects defined core datasets: the German Medical Informatics Initiative (MII) [2] has defined six modules for their core dataset, the German Interdisciplinary Association for Intensive Care and Emergency Medicine (DIVI) defined a core data set for emergency hospitalization [3], and, on an international scale, the BrainIT group gathered 25 experts from nine countries to compose a core dataset for patients with traumatic brain injury [4]. None of the core datasets developed so far meet the requirements for profound lung research.

Goal of this study is to compose nine core datasets, one for each of the eight DZL disease areas, and one disease area overarching DZL core dataset. Those core datasets need to be well annotated, e.g., by referring to standard terminologies.

2. Methods

2.1. Defining Criteria for Elements of Core Datasets

The DZL central data management proposed a definition on how to identify core parameters for lung health research. After revision by the DZL Board of Directors, the definition served as guideline for the following compositions of core datasets.

2.2. Compilation of core datasets for the DZL disease areas

Disease area coordinators were asked to name at least two domain experts per disease area to oversee defining the respective core datasets. The domain experts had to compose parameter lists fulfilling the criteria for core dataset elements. It was not specified how to find the parameters. Nevertheless, we proposed the methodology to take a large and well-established study as basis and filter the parameters that fulfill the requirements. The resulting parameter lists were discussed within each disease area for consensus.

In a second step, the lists from all eight disease areas were harmonized in uniform Excel sheets. List items were classified into three levels of semantic depth: "category", "parameter", or "characteristic" (e.g., "Biometric Data", "Gender", and "Female"). Within several feedback loops involving the domain experts, entries were complemented by generalization (e.g., "Gender" → "Biometric Data") and specification (e.g. "Specimen" → "Blood"). For unambiguity, we added columns for "data type" as well as a link to the entry in our metadata repository CoMetaR [5] if available.

2.3. Compilation of a Core Dataset for the DZL

The entries of all disease area core datasets were compared and overlapping entries identified. The DZL core dataset was then defined as all parameters covered by at least two disease areas. To assess the feasibility of future collection of these parameters from all disease areas, regardless of whether they defined it for their own core dataset, a meeting with domain experts from all disease areas was called. All attendees were asked to which DZL core parameters their domain can contribute.

2.4. Integration of Core Dataset in the DZL Metadata Repository

All parameters from all core datasets were added to the DZL metadata repository CoMetaR if not yet included. Besides attributes like label, description, and data type, whenever applicable, a reference to international classifications like SNOMET-CT, ICD-10, and LOINC was added. We recorded how many parameters had to be added to CoMetaR and to what extent core datasets were covered by international classifications.

3. Results

3.1. Core Dataset Definition

The DZL leadership agreed on the following definition of a core dataset: "A disease area-specific dataset definition is required to enable cross-dataset evaluations within the disease area as well as across disease areas and to ensure good data quality. The dataset definition MUST contain all information required for a reliable diagnosis. These are parameters and criteria that are essential for correct phenotyping according to current guidelines. The dataset definition SHOULD include all relevant information on symptom burden, prognosis, quality of life (e.g. EQ5D), inclusion criteria as well as longitudinal course and extent of treatment."

3.2. Core Dataset Size and Terminology Coverage

We composed one DZL and eight plus one core datasets for all disease areas. The additional core dataset was created during the composition within the ARDS disease area. The disease "pneumonia", included in the ARDS disease area, turned out to have significantly distinct parameters. Table 1 shows the sizes of all core datasets, to what extent they were already included in the DZL metadata catalogue, and how well the core datasets are covered by international classifications. The number of parameters ranges from 16 to 159, the ratio of new parameters per core dataset ranges from three percent to 65 percent, and the coverage of international classifications ranges from 48 percent to 100 percent. The terminologies used for standardized code annotations were SNOMED-CT, International Classification of Diseases (ICD) German modification, Logical Observation Identifiers Names and Codes (LOINC), Anatomical Therapeutic Chemical (ATC) Classification System, and Operation and Procedure (OPS).

For quantitative view on the DZL core dataset, in the following paragraph, ARDS parameter coverage counts as true if both the ARDS and pneumonia core dataset cover it. Only two parameters were seen as core parameters by all eight disease areas, three by seven, two by six, nine by five, four by four, six by three, and 15 by two disease areas. The feasibility meeting resulted in nine parameters that are available for eight disease areas, nine parameters by seven disease areas, five by six, four by five, two by four, none by three, and eleven parameters by two disease areas.

The DZL core dataset is an intersection of the 9 disease area specific core datasets. The parameter "comorbidities" is an exception, as it is essential for all disease areas, but the process of generalization and specification of core parameters yielded in different lists of comorbidities due to different research scope of the disease areas. Other than parameters like “smoking status” with definite characteristics, comorbidities are not immediately comparable across datasets.

Table 1. Shown are the core dataset domains, the number of parameters, the ratio of newly added parameters and the ratio of parameters that are listed in international classifications. DZL stands for German Center for Lung Research, AA for Asthma and Allergy, ARDS for Acute Respiratory Distress Syndrome, CF for Cystic Fibrosis, COPD for Chronic Obstructive Pulmonary Disease, DPLD for Diffuse Parenchymal Lung Disease, ELD for End stage Lung Disease, LC for Lung Cancer, and PH for Pulmonary Hypertension.

Domain	No. of Parameters	No. of New Parameters	No. of Parameters with intern. Codes
DZL	40	12 (30%)	34 (85%)
AA	51	30 (59%)	47 (92%)
ARDS	159	103 (65%)	125 (78%)
CF	62	2 (3%)	43 (69%)
COPD	80	44 (55%)	69 (86%)
DPLD	94	34 (36%)	75 (80%)
LC	80	20 (25%)	38 (48%)
ELD	16	9 (56%)	10 (63%)
PH	51	3 (6%)	51 (100%)
Pneumonia	95	54 (57%)	88 (93%)

4. Discussion

With this study, we were able to define ten core datasets with over 700 essential parameters. During meetings with disease area domain experts, we constantly faced the problem that different participants had different understandings of what "core dataset" means. The participants had different background and were often "lead principal investigators" responsible for studies or registers with narrow research focus. Additionally, previous contact to the scientists who performed this study was often related to technical data integration into the DZL central data warehouse. We encountered three different understandings of the concept "core dataset": (1) with focus on the disease area: list of parameters that are the most important for clinical research in the given domain, (2) with focus on a certain study or register: minimal list of parameters that have to be recorded in each single dataset, and (3) with focus on data integration: exclusive list of parameters that have to be delivered to the central data warehouse.

Besides the definition of which data items are to be considered as core dataset parameters, we did not specify how the domain experts should compose those datasets. We found this a reasonable choice since every disease area has its own peculiarities. In retrospect, it allowed us to adapt to variances. For example, one disease area had to rethink the structure of their diagnostics. Another disease area turned out to require two distinct core datasets for two different diseases, which are both included in the same disease area. For one disease area, a core dataset practically already existed. Additionally, it was a common question to ask the study performers about what parameters other disease areas had already picked for their core dataset. Consequently, one disease area adopted the whole "lung function parameters" section.

We found that the methodology of filtering the most relevant parameters from a given study or register is very efficient. Exemplarily, one disease area named four domain experts who each rated the collected data items of a big study with scores from zero to four. Items with big gaps (e.g., 4-4-1-4) were discussed in plenary and finally, all items with a sum of 15 or 16 were included in the core dataset.

We defined the DZL Core Dataset by picking all parameters that are included in at least two disease area core datasets. It is arguable whether this is the best definition. Alternatives would be a higher threshold, e.g., at least half of all specific datasets, or a separate survey. The presented dataset gives an impression of all parameters that are of interest for cross-disease area lung research. However, the included parameters may be falsely assumed as to be available in every single study or register within the DZL.

The amount of more than 200 parameters that are considered as essential by domain experts, but were previously missing in the metadata repository, is one aspect confirming the importance of this study. Additionally, the compiled datasets were already used as basis for further data integration of studies/registers into the DZL central data warehouse.

The fact that all disease areas consider the parameter "comorbidities" to be important, but that the recorded characteristics vary considerably, leads to the need for further research. Only when a parameter is recorded in a comparable manner, it can be used to gain meaningful cross-disease research results.

5. Conclusions

Finding the common denominator for the most important clinical parameters in a health research area is challenging, given the diverse perspectives and research interests of dozens of stakeholders. To facilitate this process, we developed a profound definition of what to include in a health research core dataset. We exemplarily defined core datasets within the German Center for Lung Research and share our experiences on what to consider during the composition and harmonization process. Our work resulted in ten core datasets that were fully implemented in our metadata repository CoMetaR. For purpose of interoperability, international classifications were referenced where possible. The composition of core datasets is the next step towards more meaningful health data collections and contributes to future cross-site and cross-disease area research projects.

References

[1] Mark D. Wilkinson. The FAIR Guiding Principles for scientific data management and stewardship.

[2] Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018;57(S 01):e50-e56. PMID:30016818

[3] Kulla M, Baacke M, Schöpke T, Walcher F, Ballaschk A, Röhrig R, Ahlbrandt J, Helm M, Lampl L, Bernhard M, Brammen D. Kerndatensatz „Notaufnahme" der DIVI. Notfall Rettungsmed 2014;17(8):671-681. doi:10.1007/s10049-014-1860-9

[4] Piper I, Citerio G, Chambers I, Contant C, Enblad P, Fiddes H, Howells T, Kiening K, Nilsson P, Yau YH. The BrainIT group: concept and core dataset definition. Acta Neurochir (Wien) 2003;145(8):615-28; discussion 628-9. PMID:14520540

[5] Stöhr MR, Helm G, Majeed RW, Günther A. CoMetaR: A Collaborative Metadata Repository for Biomedical Research Networks. Stud Health Technol Inform 2017;245:1337. PMID:29295418