Article published online: 2024-05-13

Original Article

Europe's Largest Research Infrastructure for Curated Medical Data Models with Semantic Annotations

Sarah Riepenhausen[1] Max Blumenstock[2] Christian Niklas[2] Stefan Hegselmann[1] Philipp Neuhaus[1] Alexandra Meidt[1] Cornelia Püttmann[1] Michael Storck[1] Matthias Ganzinger[2] Julian Varghese[1] Martin Dugas[2,3]

[1] Institute of Medical Informatics, University of Münster, Münster, Nordrhein-Westfalen, Germany
[2] Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany
[3] European Research Center for Information Systems (ERCIS), Münster, Nordrhein-Westfalen, Germany

Address for correspondence Martin Dugas, MD, M.Sc., Institute of Medical Informatics, Heidelberg University Hospital, Im Neuenheimer Feld 130.3, D-69120 Heidelberg, Germany (e-mail: martin.dugas@med.uni-heidelberg.de).

Methods Inf Med

Abstract

Background Structural metadata from the majority of clinical studies and routine health care systems is currently not yet available to the scientific community.

Objective To provide an overview of available contents in the Portal of Medical Data Models (MDM Portal).

Methods The MDM Portal is a registered European information infrastructure for research and health care, and its contents are curated and semantically annotated by medical experts. It enables users to search, view, discuss, and download existing medical data models.

Results The most frequent keyword is "clinical trial" ($n = 18,777$), and the most frequent disease-specific keyword is "breast neoplasms" ($n = 1,943$). Most data items are available in English ($n = 545,749$) and German ($n = 109,267$). Manually curated semantic annotations are available for 805,308 elements (554,352 items, 58,101 item groups, and 192,855 code list items), which were derived from 25,257 data models. In total, 1,609,225 Unified Medical Language System (UMLS) codes have been assigned, with 66,373 unique UMLS codes.

Conclusion To our knowledge, the MDM Portal constitutes Europe's largest collection of medical data models with semantically annotated elements. As such, it can be used to increase compatibility of medical datasets and can be utilized as a large expert-annotated medical text corpus for natural language processing.

Keywords
- data model
- semantic annotation
- metadata
- repository
- information systems
- interoperability
- EDC
- EHR

Introduction

In a clinical study or in clinical information systems, the list of data items that appear on any form—including properties like item name, description, and data type—constitutes a data model. These models are crucial for assessing the compatibility of data from different sources, where data from compatible systems can be merged and directly compared with each other. To foster the sharing of such data models, in 2011 the Portal of Medical Data Models

received
October 21, 2021
accepted after revision
March 29, 2024

DOI https://doi.org/10.1055/s-0044-1786839.
ISSN 0026-1270.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

(MDM Portal; https://medical-data-models.org/) was established, and in 2016 we reported about this information infrastructure.[1]

The need for the MDM Portal is apparent when considering the limited transparency of data models from clinical research and health care. For example, ClinicalTrials.gov (https://clinicaltrials.gov/) reports >487,000 registered studies (as of March 2024) and provides eligibility criteria and study results. However, these eligibility criteria make up only about 1 to 2% of data items per study (on average: one to two pages of >100 pages per trial). At present, most of the information in case report forms (CRFs) is undisclosed: the scientific community does not have access to a precise description of collected data items.

The situation regarding information systems in routine health care is similar: almost every hospital or health care provider uses individually customized documentation forms that are not available to the public. Additionally, these data models are different regarding language, meaning that electronic health record (EHR) systems apply language-specific data elements (e.g., in German). Therefore, millions of non-standardized data elements do exist. Further, although considerable effort is dedicated to transforming and analyzing existing data, transformations performed after data collection have major limitations: typically, data need to be aggregated until compatibility is reached, resulting in a considerable loss of information. From an informatics point of view, the compatibility of medical data models should be addressed already at the design stage of information systems. In fact, this is a key aspect of the FAIR principles (making data findable, accessible, interoperable, reusable).[2]

The advantages of model sharing and open metadata have been described before.[3] Transparency is a mandatory prerequisite for better data models: consensus regarding data standards in medicine requires discussion between different stakeholders, and this discussion requires access to data models.

However, there are currently many similar, but different ways to model a given disease regarding medical history, findings, diagnosis, therapy, and outcomes. This is because medical terminology is so complex—for example, the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT, https://www.snomed.org/) contains more than 350,000 unique medical concepts—that even small changes in a data model, such as changing a pain scale from four to five grades, can lead to incompatible data.

Given that transparency in data models is important for fostering data standards in medicine and improving compatibility of data, the objective of the MDM Portal is to contribute to this transparency. It is a registered European information infrastructure that provides a multilingual platform for exchange and discussion of data models in medicine, both for medical research and routine health care. The MDM Portal's graphical user interface is available in eight languages (English, German, Spanish, Italian, Swedish, Portuguese, French, and Dutch).

In cooperation with Heidelberg University Library, digital object identifiers can be assigned to enable citation of data models. This is relevant because public funders of clinical research are increasingly demanding that researchers must publish CRFs to gain funding. Further, since 2016, the number of available models and the international user community has increased significantly. For example, the European LeukemiaNet (https://www.leukemia-net.org/) and the Study of Health in Pomerania (SHIP)[4] have contributed contents for the MDM Portal. Currently, the MDM Portal contains over 25,000 active data models, defined by the data model language Operational Data Model (ODM), which was developed by the CDISC organization (https://www.cdisc.org). Most data models encompass clinical trials, especially eligibility criteria. English and German are the most frequent languages of data elements.

The MDM Portal can be used to search and optimize existing forms, and to design new forms based on existing contents, such as by reusing data elements or creating core datasets.[5–7] Further, because it contains over 800,000 manually curated element annotations, the portal can serve as a pragmatic tool for coding medical data elements. By providing coding principles, the MDM Portal can support consistent coding quality and, thus, data quality. For instance, codes from the MDM Portal are used by the metadata registry (MDR) Samply.MDR,[8] and annotations from the MDM Portal are used to enrich data dictionaries from SHIP cohort studies.[4]

The objective of this work is to provide an overview of MDM contents and available services for the scientific community.

Methods

IT Implementation

The architecture of the MDM Portal has been described previously.[1] In summary, this portal stores medical data models in CDISC ODM format in a PostgreSQL database. Since 2016, the MDM Portal's software platform and database have been completely re-designed and re-implemented; it is now based on the Spring MVC framework and is written in Java and R. Additional web services provide converters into several formats.[9] Apache Solr is applied for all search functionalities on the level of data models, data elements, and semantic annotations. Search functionalities are accessible through a public representational state transfer application programming interface (REST API) and, thereby, can be made available to external systems, such as a search and suggestion mechanism for semantic annotations.[10] Registered users can search, view, download, comment, edit, and upload data models.

The results from a search for data models can be filtered with three approaches that can be combined. First, a chapter or sub-chapter from the table of interest can be chosen, and only data models with keywords belonging to that chapter are displayed. Second, one or more keywords of interest can be chosen directly. Third, operators and other advanced search options can be used; for example, a search containing "-eligibility" will exclude results with this term.

Data Models

The MDM Portal's data models (with one ODM file defining one data model, possibly containing more than one form; for

an ODM example, see Dugas[11]) are mostly created on the basis of existing PDFs, tabular data files, data entry forms, or similar documents that are freely available or can be published with consent of the original author. Typical examples are eligibility criteria, CRFs, routine documentation forms from EHRs, or questionnaires such as patient-reported outcome measures. These documents are manually transformed into ODM files with web-based tools such as ODMedit.[12] Whenever possible, tabular data or data in other formats are transformed with the help of custom R scripts or converters (for an example in Java, see Hegselmann et al[13]). Uploads from external users are reviewed by an administrator or a moderator before public release. Some models were created based on selected MDM documents with the CDEGenerator[14] as core datasets.[5–7]

The data models are stored as ODM files in the above-mentioned PostgreSQL database. In addition, these ODM files are decomposed into their components, which are stored in a second database serving as a MDR according to ISO/IEC 11179.[15] In the MDR, identical elements are only stored once. For example, data model A contains a form with item "Age" within an item group; data model B contains a different form with different item groups but has an identical item "Age." This item "Age" is automatically assigned a unique identifier, but is linked to all instances from both data models in the MDR.

Items and item groups from this MDR can be re-used for new ODM files. The MDR can be queried from external systems via REST API connections; items and related items (i.e., items, which co-occur frequently) can be searched and viewed, including frequency of occurrence.

Semantic Annotation

Expert-curated semantic annotation with Unified Medical Language System (UMLS) codes[16] is provided for the majority of data items (on several levels: item group, data item, and code list). The manual UMLS annotation is based on established coding principles and semi-automatic code suggestions.[17] These principles provide a systematic workflow to the coder for pre-coordination and post-coordination of medical terms (►**Fig. 1**, generalized from Varghese and Dugas[18]). The semi-automatic code suggestion function is integrated so that annotations that have been frequently selected by previous coders for similar terms can be reused.[18] In prior work, we showed that both of these mechanisms (coding principles combined with code suggestions) can improve different coders' inter-rater reliability and reduce coding time.[17] Since 2011, semantic annotations for the MDM Portal have been generated by a team of two full-time physicians assisted by approximately five clinical-phase medical students (four eyes principle). In addition to UMLS annotations, other terminology or classification codes can also be used to code data elements in MDM. Each data model was assigned medical subject headings (MeSH)[19] with a similar approach.

Analysis Approach

The MDM Portal provides version control; therefore, the most recent version of each data model was analyzed, specifically regarding data elements and their semantic annotations. Each data item contains a name (e.g., "body weight") with a description (frequently multi-lingual), a data type (e.g., "float"), and, if annotated, one or more UMLS codes (e.g., "C0005910"). Keywords for those models are based on MeSH, with a few custom extensions from a local dictionary (e.g., "routine documentation," "released standard"). Primary categories of data models are clinical trials, EHRs, registries, quality assurance, and other. Contents of the MDM Portal were analyzed with R scripts to generate descriptive statistics regarding frequency of data elements and UMLS codes.

Results

A total of 805,308 semantically annotated elements (554,352 items, 58,101 item groups, and 192,855 code list items) from 25,257 data models are available in the MDM database and can be downloaded.[20] Anyone can access and view the contents of the MDM portal. All data models are available under a creative commons license. License information is available for each data model.

Registered users can create, adapt, analyze, download, and reuse data models free of charge. Search functionalities are also available to unregistered users. The info button in the search bar provides examples on how to use search operators and how to filter specific fields. Data models of interest can be selected and downloaded or analyzed directly with ODMSummary[21] or CDEGenerator.[14] The FAQ/help section provides additional material to support MDM users (e.g., texts, videos).

The source code of MDM portal and associated web services are available at https://imigitlab.uni-muenster.de/published/mdm/.

Use Cases of MDM

Target users of MDM are designers of study databases (such as data managers) and clinical information systems in health-care (e.g., EHR analysts). Developers of medical data standards (like physicians) and medical statisticians can also directly use MDM. To build study databases, MDM contents can be downloaded in CDISC ODM, REDCap, OpenClinica, and MACRO format. Designers of clinical IT systems can use HL7 Fast Healthcare Interoperability Resources (FHIR) formats (XML, JSON, RDF), HL7 CDA as well as openEHR ADL format to implement data models from MDM in the local system. App developers can do model-driven software development with MDM contents in Research Kit (Android), Research Kit Swift (iOS), and Open Data Kit format. Data analysts can download item catalogs in SPSS, R, and Excel format to prepare statistical analysis of local datasets. Importantly, developers of medical data standards can use CDEGenerator[14] to identify the semantic core of different data models (even if provided in different languages) and use the MDM platform to reach consensus with an expert group about a data standard for a specific medical domain.

Data Elements

In total, available contents include 610,813 data items, 87,202 item groups, and 748,653 code list items. A data

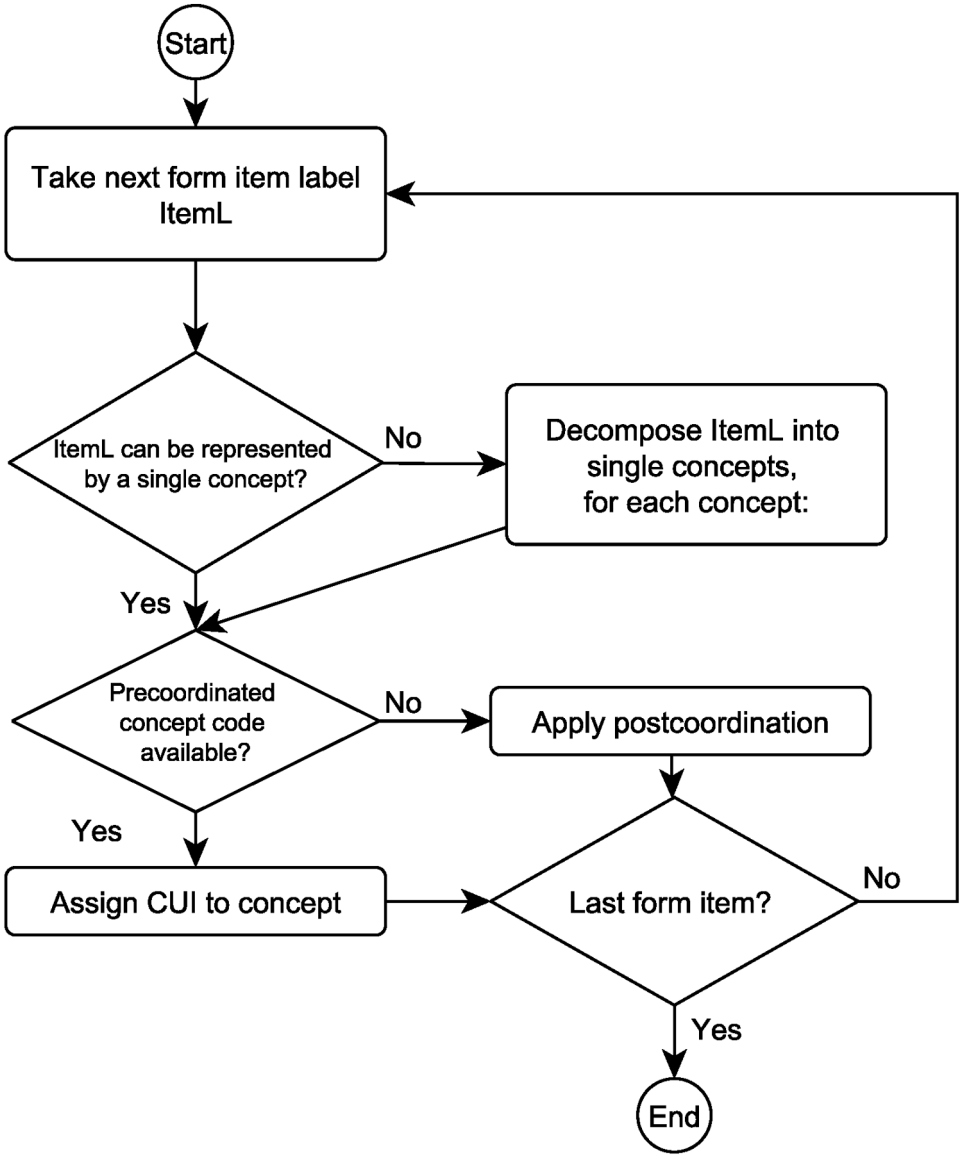

Fig. 1 Flow diagram of the annotation process of items from a medical form. Item labels are analyzed regarding medical concepts. Each concept is annotated semi-automatically: the system suggests a set of already used label-coding-combinations based on the search terms and sorted by fit and frequency. The user chooses the best fitting option or manually enters a different code. Pre-coordinated concept codes are given preference. If there are no suitable pre-coordinated options, two or more codes can be combined in post-coordination. CUI, concept unique identifier.

model consists of a median of 15 items (range 1–1,964; interquartile range [IQR]: 10–24). Most data items are available in English ($n = 545{,}749$) and German ($n = 109{,}267$). In total, 70 different languages are used, including language variants like American, Australian, or British English (53 languages, if variants are grouped). ►**Table 1** presents the 20 most frequent languages of items.

Keywords

Each data model is tagged with one or more keywords. ►**Table 2** presents the 20 most frequent keywords. ►**Table 3** reports frequency of keywords by MeSH disease category; it indicates that a wide range of diseases is covered. Clearly, most data models of the MDM Portal are derived from clinical studies. Most frequent disease-associated keywords are "breast neoplasms" (1,943) and "diabetes mellitus, type 2" (1,100). Most data models belong to oncological diseases (7,675 with at least one MeSH keyword from the diseases category C04; 9,779 MeSH terms from C04 used in total), followed next by cardiovascular diseases (3,122 models with at least one keyword from C14). A total of 1,298 models are derived from routine documentation. ►**Fig. 2** reports frequent combinations of the 10 most common keywords per data model.

Semantic Annotation

Manually curated semantic annotations with UMLS are available for 805,308 elements (554,352 data items

Table 1 The 20 most frequent languages of data items ("translated texts" of ODM "questions")

Language	Frequency
English (including variants from United States, United Kingdom, and Australia)[a]	545,749
German (including variants from Switzerland and Austria)[a]	109,267
Swedish	3,694
Italian	1,302
French (including variant from Canada)[a]	1,266
Spanish (including variants from Argentina, Chile, Spain, Mexico, and Paraguay)[a]	827
Portuguese (including variant from Brazil)[a]	789
Norwegian	636
Dutch (including variants)[a]	633
Arabian (including variant from Syria)[a]	491
Polish	345
Greek	311
Turkish	286
Danish	237
Finnish	209
Russian	206
Korean	191
Hungarian	189
Chinese (including mainland China variant)[a]	181
Bulgarian	148

Abbreviation: ODM, Operational Data Model.
[a]Language variants grouped.

(90.8%), 58,101 item groups (66.6%), and 192,855 code list items (25.8%)). A key use case for the MDM Portal is reuse of data items; therefore, most efforts regarding manual semantic annotation are spent on those items.

Overall, 1,609,225 UMLS concept codes are assigned, of which 66,373 are unique. ►**Table 4** presents the 20 most frequent UMLS codes. The median number of occurrences per UMLS code is only two, but there is a wide range (1–19,940, IQR: 1–7). This demonstrates the semantic richness of data elements: there is a long tail of UMLS codes that are used infrequently. The median number of UMLS codes per UMLS coded item is 2 (range: 1–227, IQR: 1–3). In total, 236,811 data items are assigned only one UMLS code (pre-coordination) and 317,541 items are annotated with two or more codes (post-coordination). The median number of UMLS-coded items per data model is 15 (range: 1–1,964, IQR: 10–23).

Usage Characteristics

►**Fig. 3** presents the total number of data models between 2011 and 2024. Between 2016 and March 2024, the total number of models increased from 4,387 to 25,257. The median number of versions per data model is 1 (range: 1–20, IQR: 1–2). MDM has 14,251 registered users worldwide (as of March 2024). ►**Fig. 4** presents an MDM screenshot with exemplary search results.

Table 2 The 20 most frequent keywords on the MDM Portal

Keyword	Frequency
Clinical trial	18,777
Eligibility determination	11,050
Breast neoplasms	1,943
Cardiology	1,682
Routine documentation	1,298
Neurology	1,185
Endocrinology	1,131
Hematology	1,111
Diabetes mellitus, type 2	1,100
Gynecology	1,082
Laboratories	1,077
Adverse event	960
Clinical trial, phase III	950
Gastroenterology	876
Vital signs	825
Psychiatry	816
Physical examination	801
Follow-up studies	737
Treatment form	732
Released standard	711

Abbreviation: MDM, Medical Data Model.

Discussion

The MDM Portal has been regularly presented at scientific congresses of the European Federation for Medical Informatics (https://efmi.org/), German Association for Medical Informatics, Biometry, and Epidemiology (https://www.gmds.de/), Meetings of the German CDISC User Network (https://www.cdisc.org/), and others. In cooperation with the technology and method platform for networked medical research (TMF e.V., https://www.tmf-ev.de/), workshops have been held annually and over 100 users have been trained, and continuous feedback on community requirements for MDM contents has been obtained. Further, an external team (University of Applied Sciences Bern, Switzerland) performed a usability study of the MDM Portal.[22] This study addressed technical aspects (e.g., test with Web site tool Nibbler) and offered an assessment by 10 users from two clinical trial units. In addition, 80% of the users agreed or strongly agreed that MDM provides relevant and reliable information.

As described above, the MDM Portal's dataset provides semantic annotation with UMLS codes, especially for items. Another important semantic coding system is SNOMED, which is widely used in Europe and beyond. However,

Table 3 Frequency of data models with at least one MeSH term from the respective MeSH disease category and overall frequency of MeSH terms from these categories

MeSH Tree-number	MeSH disease category/term	Frequency of data models	Overall frequency of MeSH terms
C04	Neoplasms	7,675	9,779
C14	Cardiovascular diseases	3,122	4,606
C20	Immune system diseases	3,074	3,645
C10	Nervous system diseases	2,570	5,960
C17	Skin and connective tissue diseases	2,497	2,591
C19	Endocrine system diseases	2,224	2,463
C06	Digestive system diseases	2,076	5,378
F03	Mental disorders	2,056	2,738
C18	Nutritional/metabolic diseases	2,055	2,300
C15	Hemic and lymphatic diseases	1,824	2,594
C08	Respiratory tract diseases	1,520	3,218
C23	Pathological conditions, signs and symptoms	1,501	1,860
C12	Male urogenital diseases	1,348	3,001
C13	Female urogenital diseases and pregnancy complications	1,074	1,902
C02	Virus diseases	1,056	2,203
C05	Musculoskeletal diseases	730	1,491
C11	Eye diseases	573	727
C01	Bacterial infections and mycoses	299	537
C16	Congenital, hereditary and neonatal diseases, and abnormalities	249	612
C09	Otorhinolaryngologic diseases	234	349
C25	Chemically induced disorders	188	199
C07	Stomatognathic diseases	124	285
C26	Wounds and injuries	85	145
C03	Parasitic diseases	84	87
C24	Occupational diseases	7	7
C22	Animal diseases	1	1

Abbreviation: MeSH, medical subject heading.

although several countries do not have a national SNOMED license yet, users with a SNOMED license can make use of existing cross-mappings between UMLS and SNOMED. Further, semantically annotated data models can be compared with tools like ODMSummary[21] or CDEGenerator,[14] for example, to develop common data elements for information systems. This has already been done, e.g., for acute coronary syndrome,[5] acute myeloid leukemias,[6] and the neuroinflammatory disease multiple sclerosis.[7] The latter was developed for use in neurological units of two university hospitals.

The Institute of Community Medicine of the University of Greifswald extracts the UMLS annotation provided by the MDM Portal via a restful API connection for use in their own data dictionary of SHIP,[4,15] and the Medical Informatics Group of the University Hospital of Frankfurt has integrated the code suggestion function of the MDM Portal to profit from the large set of annotated elements for annotation in Samply.MDR.[8]

In total, 805,308 elements with semantic annotations are available from 25,257 data models; thus the contextual meaning of diverse medical text segments is machine readable, including synonyms and complex clinically relevant concept relations. Thus, this dataset can be applied as a unique knowledge base for various natural language processing[23] systems in the clinical domain. However, there is still a long road ahead, as medicine is so complex that even having >25,000 data models only represents a starting point: the International Classification of Diseases version 10[24] lists in its German version more than 13,000 diagnoses and is a coarse-grained system. Each diagnosis is associated with a different set of data items for medical history, clinical examination, therapeutic interventions, and follow-up. In our setting, semantic annotation proved to be a difficult task: several approaches for fully

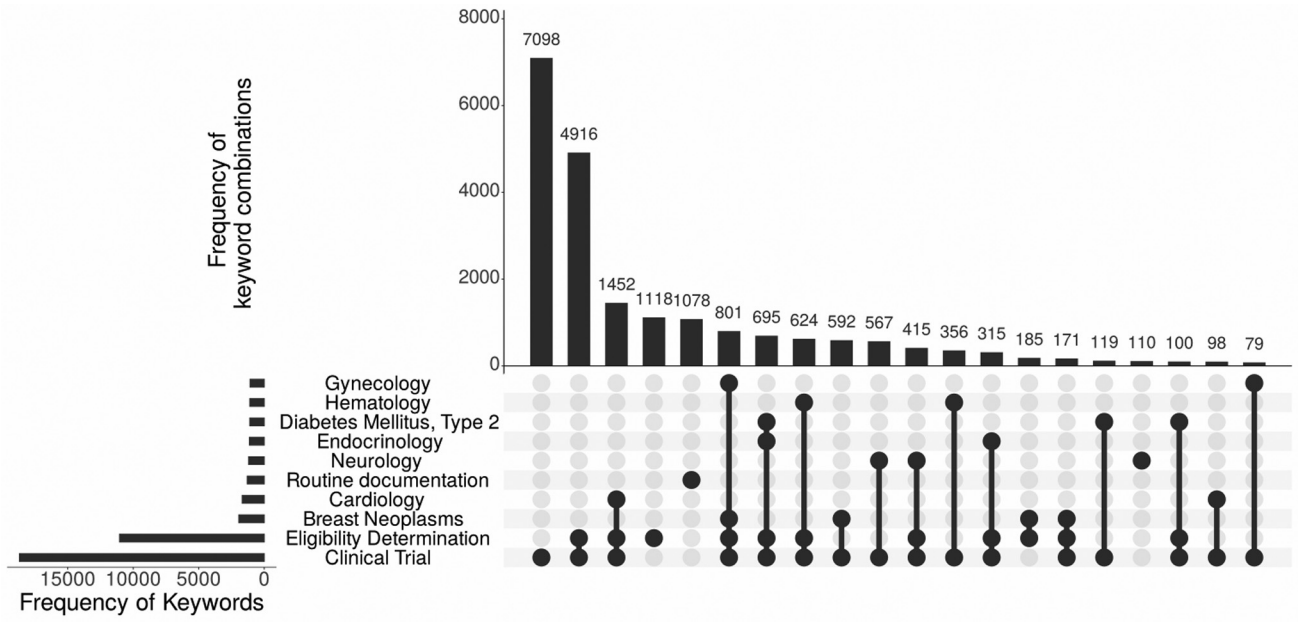

Fig. 2 UpSet plot of the top 10 keywords assigned to data models. It indicates the most frequent combinations of keywords. For example, there are 1,452 models regarding eligibility determination in clinical trials dealing with cardiology. "Clinical trial" and "clinical trial" plus "eligibility determination" occur very frequently because of combinations with many different less-common keywords.

Table 4 The 20 most frequent UMLS concept unique identifiers used in MDM Portal

UMLS concept unique identifier	Concept	Frequency
C0011008	Date in time	19,940
C0680251	Exclusion criteria	12,109
C1512693	Inclusion	11,581
C0205394	Other	11,402
C0001779	Age	9,680
C2348585	Clinical trial subject unique identifier	9,166
C0021430	Informed consent	8,407
C0040223	Time	8,207
C0030705	Patients	7,507
C0013227	Pharmaceutical preparations	6,991
C1298908	No	6,898
C0332307	Type—attribute	6,599
C0877248	Adverse event	6,592
C1274040	Result	6,538
C0439673	Unknown	6,495
C0518766	Vital signs	6,477
C0022885	Laboratory procedures	6,388
C0332197	Absent	6,316
C0087111	Therapeutic procedure	6,306
C0032961	Pregnancy	5,994

Abbreviation: MDM, Medical Data Model; UMLS, Unified Medical Language System.

automated annotation did not yet yield an acceptable coding quality; therefore, we applied manual expert curation. National and international collaboration is needed to further develop contents according to the needs of the scientific community.

MDM is providing data models, not ontologies, which is a different setting:

An ontology encompasses a representation, formal naming and definition of categories, properties and relations between concepts, data, and entities. In contrast, a data model (for example derived from a specific clinical study) corresponds to real existing datasets. Aside from the MDM Portal, related approaches toward publishing data models do exist. For example, REDCap,[25] an electronic data collection (EDC) system from Vanderbilt University, provides a CRF library with 2,434 data collection instruments and forms (as of March 2024) but without semantic coding. Another external REDCap library is the PhenX toolkit (https://www.phenxtoolkit.org/), which was funded by the U.S. National Human Genome Research Institute and contains 984 protocols (as of March 2024). A further CRF library is provided by the EDC system OpenClinica, comprising a collection of 56 CDISC CDASH-compliant eCRFs (as of March 2024).

There are public collections of data elements: for instance, the Cancer Data Standards Registry and Repository (https://cadsr.cancer.gov/) from the National Cancer Institute publishes data elements, common data elements, and CRFs. Common data elements are also published by the National Institute of Neurological Disorders and Stroke.[26] Furthermore, there are also published data models in the context of EHR systems, which are coordinated by the HL7 organization (https://www.hl7.org/). One example is the XML-based Clinical Document Architecture (CDA) for the exchange of documents. In recent years, various organizations around the

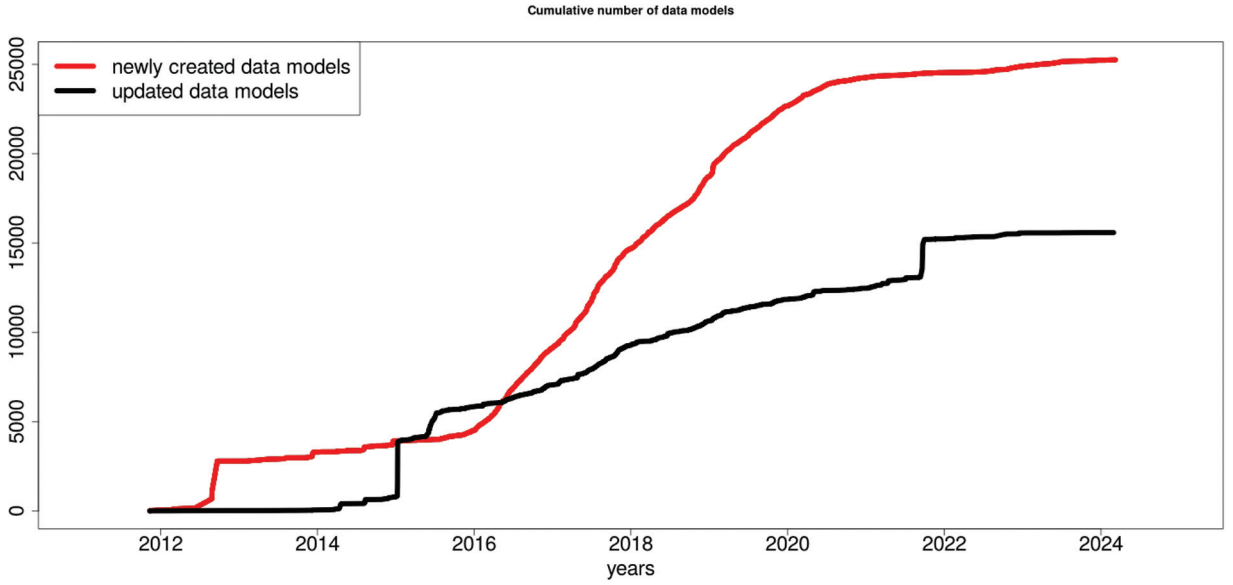

Fig. 3 Time course of developing the MDM contents. MDM, Medical Data Model.

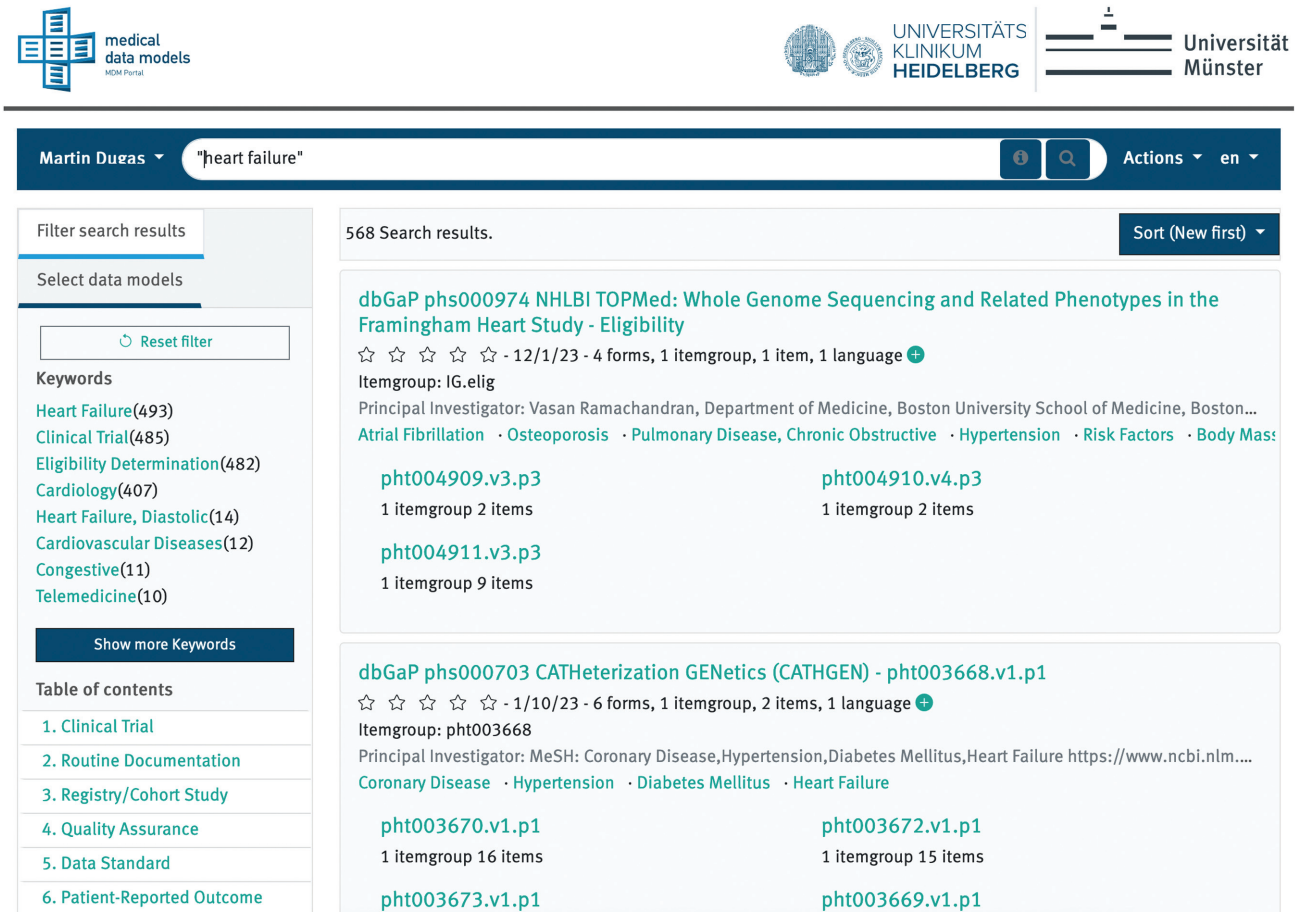

Fig. 4 Screenshot from MDM portal. Search results for “heart failure” are displayed. MDM, Medical Data Model.

world have developed unified medical documentation based on CDA. For this purpose, the structure of these documents is essential, and it is specified by the CDA model and can be modified for different use cases. HL7 also hosts material of the Clinical Information Modeling Initiative.[27] At present, the FHIR standard from HL7 is the most important evolving standard for health care data exchange. OpenEHR is another international initiative to standardize and publish medical

data structures. The OpenEHR Clinical Knowledge Manager[28] provides 180 active templates (as of March 2024).

Compared with all those other systems, a unique characteristic of MDM is support for 20 different technical formats to address key stakeholders of medical data models: data managers of clinical studies (ODM and other EDC formats), EHR analysts (HL7 formats), physicians (office formats) as well as statisticians (R, SPSS). The MDM Portal provides expert-curated semantic annotations for existing, real-world data models, i.e., data structures that have been used to collect patient data. To our knowledge, it constitutes Europe's largest collection of medical data models with semantically annotated elements. It reflects the reality of medical data collection, with all its benefits and shortcomings. Information system designers can use this resource to learn from the past and to implement more compatible systems in the future.

Authors' Contribution

S.R.: manuscript writing, statistics, revision, (supervision of) data model creation and annotation, research of available data models. M.B.: software development, revision, export of metadata and code. C.N.: revision, supervision of data model creation and annotation, research of available data models. S.H.: software architecture and development. P.N.: software development and supervision thereof, revision, export of metadata and code. A.M.: project management, dissemination concept, writing, and revision. C.P.: data model creation and annotation, research of available data models. M.S.: software development and supervision thereof. M.G.: software development and supervision thereof. J.V.: software development, writing and revision, (supervision of) data model creation and annotation, research of available data models. M.D.: Principal Investigator of MDM portal, conceptualization, selection of data models, supervision of software development, manuscript writing.

Funding

This work was supported by German Research Foundation (Deutsche Forschungsgemeinschaft, DFG grants DU 352/11-1, DU 352/11-2, DU 352/14-4).

Conflict of Interests

None declared.

Acknowledgment

The permission of physicians and scientists to publish their data models in the MDM Portal is acknowledged. The work of many student assistants to process data models and develop the MDM Portal's software is acknowledged.

References

1 Dugas M, Neuhaus P, Meidt A, et al. Portal of medical data models: information infrastructure for medical research and healthcare. Database (Oxford) 2016;2016:bav121
2 Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018
3 Dugas M, Jöckel KH, Friede T, et al. Memorandum "Open Metadata". Open access to documentation forms and item catalogs in healthcare. Methods Inf Med 2015;54(04):376–378
4 Völzke H, Alte D, Schmidt CO, et al. Cohort profile: the study of health in Pomerania. Int J Epidemiol 2011;40(02):294–307
5 Kentgen M, Varghese J, Samol A, Waltenberger J, Dugas M. Common data elements for acute coronary syndrome: analysis based on the unified medical language system. JMIR Med Inform 2019;7(03):e14107
6 Holz C, Kessler T, Dugas M, Varghese J. Core data elements in acute myeloid leukemia: a unified medical language system-based semantic analysis and experts' review. JMIR Med Inform 2019;7(03):e13554
7 von Martial S, Brix TJ, Klotz L, et al. EMR-integrated minimal core dataset for routine health care and multiple research settings: a case study for neuroinflammatory demyelinating diseases. PLoS One 2019;14(10):e0223886
8 Vengadeswaran A, Neuhaus P, Hegselmann S, Storf H, Kadioglu D. Semantically Annotated Metadata: Interconnecting Samply.MDR and MDM-Portal. Stud Health Technol Inform 2019;267:86–92
9 Soto-Rey I, Neuhaus P, Bruland P, et al. Standardising the development of ODM converters: the ODMToolBox. Stud Health Technol Inform 2018;247:231–235
10 Hegselmann S, Storck M, Geßner S, Neuhaus P, Varghese J, Dugas M. A web service to suggest semantic codes based on the MDM-Portal. Stud Health Technol Inform 2018;253:35–39
11 Dugas M. ODM2CDA and CDA2ODM: tools to convert documentation forms between EDC and EHR systems. BMC Med Inform Decis Mak 2015;15:40
12 Dugas M, Meidt A, Neuhaus P, Storck M, Varghese J. ODMedit: uniform semantic annotation for data integration in medicine based on a public metadata repository. BMC Med Res Methodol 2016;16:65
13 Hegselmann S, Gessner S, Neuhaus P, Henke J, Schmidt CO, Dugas M. Automatic conversion of metadata from the study of health in Pomerania to ODM. Stud Health Technol Inform 2017;236:88–96
14 Varghese J, Fujarski M, Hegselmann S, Neuhaus P, Dugas M. CDE-Generator: an online platform to learn from existing data models to build model registries. Clin Epidemiol 2018;10:961–970
15 Hegselmann S, Storck M, Gessner S, et al. Pragmatic MDR: a metadata repository with bottom-up standardization of medical metadata through reuse. BMC Med Inform Decis Mak 2021;21(01):160
16 Amos L, Anderson D, Brody S, Ripple A, Humphreys BL. UMLS users and uses: a current overview. J Am Med Inform Assoc 2020;27(10):1606–1611
17 Varghese J, Sandmann S, Dugas M. Web-based information infrastructure increases the interrater reliability of medical coders: quasi-experimental study. J Med Internet Res 2018;20(10):e274
18 Varghese J, Dugas M. Frequency analysis of medical concepts in clinical trials and their coverage in MeSH and SNOMED-CT. Methods Inf Med 2015;54(01):83–92
19 National Center for Biotechnology Information & U.S. National Library of Medicine. Home - MeSH - NCBI. *MeSH – NCBI*. Accessed March 18, 2024 at: https://www.ncbi.nlm.nih.gov/mesh/
20 Dugas M. Medical data models. Mendeley Data 2020. Doi: 10.17632/wmwt7s2d8v.1
21 Storck M, Krumm R, Dugas M. ODMSummary: a tool for automatic structured comparison of multiple medical forms based on semantic annotation with the unified medical language system. PLoS One 2016;11(10):e0164569
22 Reichenpfader D, Glauser R, Dugas M, Denecke K. Assessing and improving the usability of the medical data models portal. Stud Health Technol Inform 2020;271:199–206

23 Deleger L, Li Q, Lingren T, et al. Building gold standard corpora for medical natural language processing tasks. AMIA Annu Symp Proc 2012;2012:144–153
24 World Health Organisation. ICD-10 Version: 2019. International Statistical Classification of Diseases and Related Health Problems 10th Revision Accessed March 26, 2024 at: https://icd.who.int/browse10/2019/en
25 Vanderbilt University. REDCap Shared Library. REDCap Accessed March 16, 2024 at: https://redcap.vanderbilt.edu/consortium/library/search.php
26 National Institute of Neurological Disorders and Stroke. NINDS Common Data Elements. Accessed March 16, 2024 at: https://www.commondataelements.ninds.nih.gov/
27 Clinical Information Modeling Initiative | HL7 International. Health Level Seven International. Accessed March 16, 2024 at: https://www.hl7.org/Special/Committees/cimi/
28 openEHR Foundation. Clinical Knowledge Manager. OpenEHR - Open industry specifications, models and software for e-health Accessed March 16, 2024 at: https://www.openehr.org/ckm/

