VirtualBrainCloud

Personalized Recommendations for Neurodegenerative Disease

www.VirtualBrainCloud-2020.eu

Public deliverable report

D3.16: Data catalogue with harmonized annotations at project

Date	December 2020
Authors	Stephan Gebel; Sebastian Schaaf; Colin Birkenbihl; Stephan Springstubbe (Fraunhofer SCAI) Petra Ritter (CHARITÉ) © VirtualBrainCloud consortium
Dissemination level	**public**
Website	www.VirtualBrainCloud-2020.eu

This project has received funding from the **European Union´s Horizon 2020** research and innovation programme under **grant agreement No 826421**

Contents

1. Background

The overarching goal of The VirtualBrainCloud (TVB-Cloud) is personalized prevention and treatment of dementia. To achieve generalizable results that help individual patients, the VirtualBrainCloud integrates the data of large cohorts of patients and healthy controls through multi-scale brain simulation using The Virtual Brain (or TVB) simulator. There is a need for infrastructures for sharing and processing health data at a large scale that comply with the EU general data protection regulations (or GDPR). The VirtualBrainCloud consortium closes this gap, making health data actionable. Elaborated data protection concepts minimize the risks for data subjects and allow scientists to use sensitive data for research.

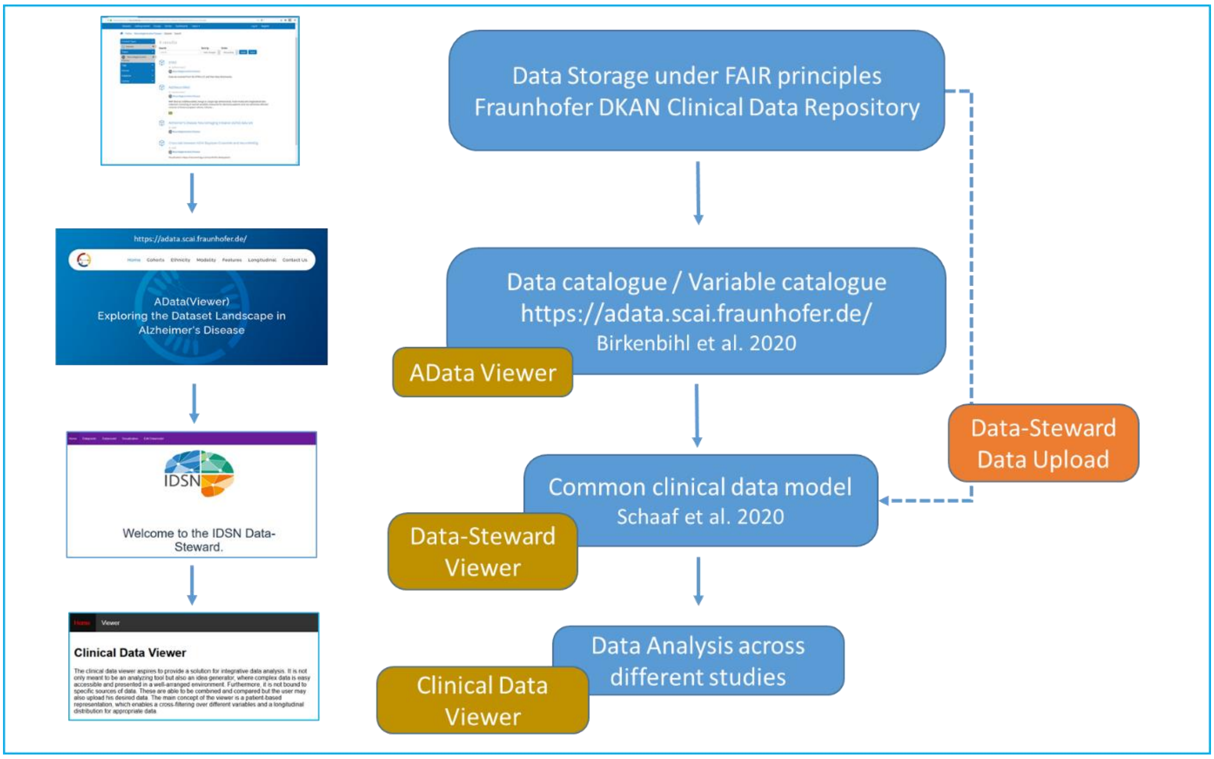
A key objective of our TVB-Cloud tasks “Workflows for clinical data curation and processing” and “A metadata framework for unified metadata annotations and Data Catalogues” under the overarching topic of “Data Processing, Standardization and Data-FAIRification” is to deliver a semantic framework for neurodegenerative diseases (NDD), that serves as a central resource for controlled vocabularies and shared ontologies to access and use them within different TVB-Cloud work packages.

2. Introduction

Within the Virtual Brain Cloud (TVB-Cloud) project, we are building a reference infrastructure for sharing and processing Health and Biomedical research data, specifically data in the domain of NDD.

One milestone on the way to succeed is described in the present report D3.16 “Data catalogue with harmonized annotations at project”. Our goal is to implement FAIR data handling principles (Findable, Accessible, Interoperable, Reproducible) through homogeneous data annotations based on shared semantics and through a data catalogue that is accessible to all project partners. The data catalogue contains harmonized annotations of all data sets used in this project and can be searched in an integrative viewer. A core deliverable of TVB-Cloud is the uniform harmonization across all clinical data sets using homogeneous annotations provided by shared semantics such as the Clinical Trials Ontology (CTO) by the end of year 2.

Fraunhofer SCAI developed an infrastructure and software tools that are able to fulfill this milestone. The system contains four main pillars: the DKAN Clinical Data Repository, AData Viewer, the Common Clinical Data Model (CCDM) and the Clinical Data Viewer (Fig. 1).


https://adata.scai.fraunhofer.de/
Home Cohorts Ethnicity Modality Features Longitudinal Contact Us
AData(Viewer)
Exploring the Dataset Landscape in Alzheimer's Disease
IDSN
Welcome to the IDSN Data-Steward.
Home
Viewer
Data Storage under FAIR principles
Fraunhofer DKAN Clinical Data Repository
Data catalogue / Variable catalogue
https://adata.scai.fraunhofer.de/
Birkenbihl et al. 2020
AData Viewer
Data-Steward
Data Upload
Common clinical data model
Schaaf et al. 2020
Data-Steward
Viewer
Data Analysis across
different studies
Clinical Data
Viewer

Clinical Data Viewer

The clinical data viewer aspires to provide a solution for integrative data analysis. It is not only meant to be an analyzing tool but also an idea generator, where complex data is easy accessible and presented in a well-arranged environment. Furthermore, it is not bound to specific sources of data. These are able to be combined and compared but the user may also upload his desired data. The main concept of the viewer is a patient-based representation, which enables a cross-filtering over different variables and a longitudinal distribution for appropriate data.

Figure 1: Infrastructure for Clinical Data Management

The **Fraunhofer SCAI DKAN** system enables the storage of clinical study data in a safe environment in compliance with FAIR data handling principles. The **AData Viewer**[1] captures information on study related variables (attributes) with focus on a (mostly) semantic mapping of specific parameter ranging from ethnicity to diagnostics. The system enables a systematic view on the data and variables and comparisons across different studies (Birkenbihl et al., 2020[2]). The **Common Clinical Data Model** (CCDM) enables the mapping and alignment of attributes (see Table 5) on the level of individual patient data and the subsequent uploading of the individual patient data based on the mapping. Here alignment and mapping of the attributes is most complex as parameter often contains study specific measurements and value ranges. Based on the attribute alignment, patient level data could be uploaded in the CCDM. Due to the harmonized data inside the CCDM, this data base could be used for cross study analysis of metadata using the **Clinical Data Viewer** that is developed in the IDSN[3] project by project partners from University Hospital Bonn.

Here we describe the setup of the whole system with focus on two major Alzheimer's Disease (AD) related studies i.e., ADNI (Alzheimer's Disease Neuroimaging Initiative) and ANM (AddNeuroMed).

[1] https://adata.scai.fraunhofer.de/

[2] Birkenbihl C. et al. Evaluating the Alzheimer's disease data landscape. Alzheimer's Dement. 2020, Volume 6; Issue 1 https://doi.org/10.1002/trc2.12102

[3] IDSN (Integrative Data Semantics for Neurodegenerative research) project is a German Federal Ministry of Education and Research (BMBF) funded project. More details are available at https://www.idsn.info/en/

3. Partners involved

FAIR-data principles implementation is led by Fraunhofer SCAI. They are responsible for the harmonization of data using semantic frameworks and the implementation of the FAIR principles in TVB-Cloud. Partner UNIVIE provides ethical and legal considerations to be applied in these processes, which are in the responsibility of the respective data controllers.

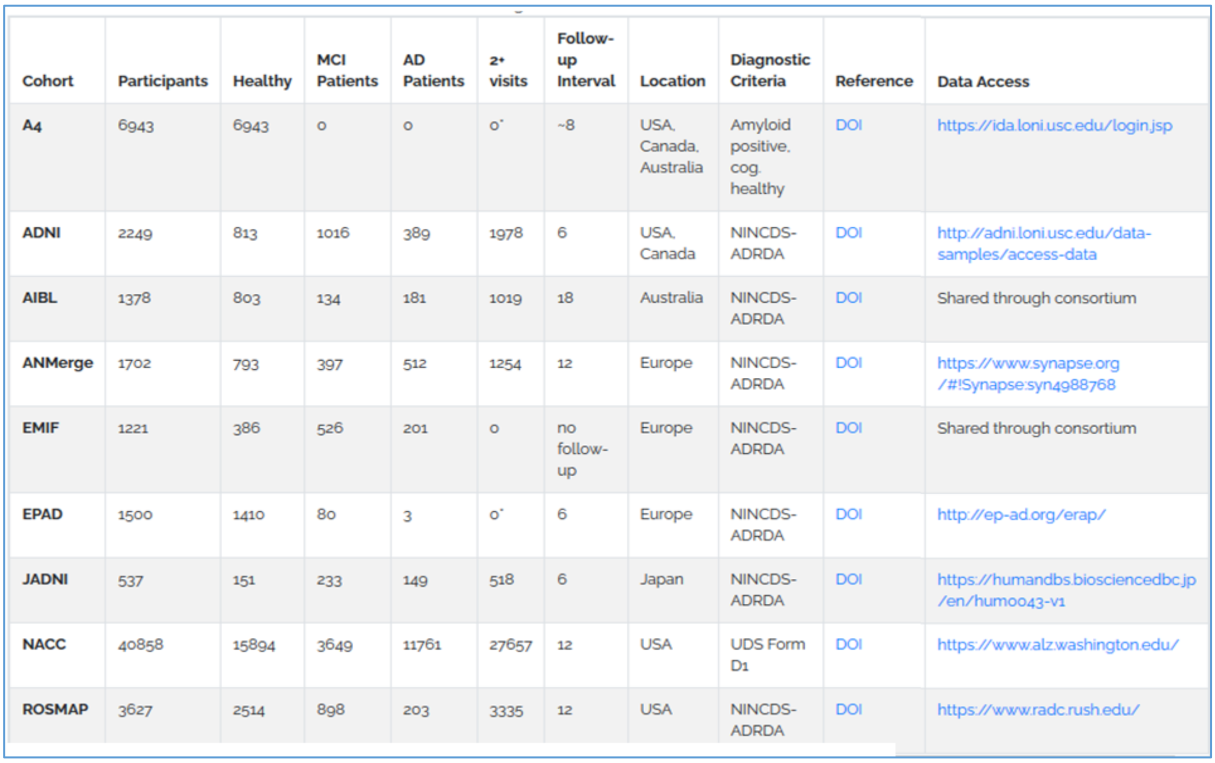
4. Description of work performed

4.1. Collecting relevant clinical data from AD studies (data acquisition)

We acquired major AD cohort studies (Table 1) to set up a data catalogue that represents a dataset landscape of AD related studies.

The first challenge was gaining access to a sufficient number of cohort datasets. Data access usually requires completing an application procedure with numerous legal and ethical requirements. If access is granted, manual curation, harmonization and investigation of data follows. Although difficult to establish, a comprehensive data-driven view on the AD data landscape is crucial.

Relevant AD cohort studies were chosen to allow for a thorough investigation of the data landscape in relation to measured parameters and used attributes as shown below (Table 1).



Cohort	Participants	Healthy	MCI Patients	AD Patients	2+ visits	Follow-up Interval	Location	Diagnostic Criteria	Reference	Data Access
A4	6943	6943	0	0	0*	~8	USA, Canada, Australia	Amyloid positive, cog. healthy	DOI	https://ida.loni.usc.edu/login.jsp
ADNI	2249	813	1016	389	1978	6	USA, Canada	NINCDS-ADRDA	DOI	http://adni.loni.usc.edu/data-samples/access-data
AIBL	1378	803	134	181	1019	18	Australia	NINCDS-ADRDA	DOI	Shared through consortium
ANMerge	1702	793	397	512	1254	12	Europe	NINCDS-ADRDA	DOI	https://www.synapse.org/#!Synapse:syn4988768
EMIF	1221	386	526	201	0	no follow-up	Europe	NINCDS-ADRDA	DOI	Shared through consortium
EPAD	1500	1410	80	3	0*	6	Europe	NINCDS-ADRDA	DOI	http://ep-ad.org/erap/
JADNI	537	151	233	149	518	6	Japan	NINCDS-ADRDA	DOI	https://humandbs.biosciencedbc.jp/en/hum0043-v1
NACC	40858	15894	3649	11761	27657	12	USA	UDS Form D1	DOI	https://www.alz.washington.edu/
ROSMAP	3627	2514	898	203	3335	12	USA	NINCDS-ADRDA	DOI	https://www.radc.rush.edu/

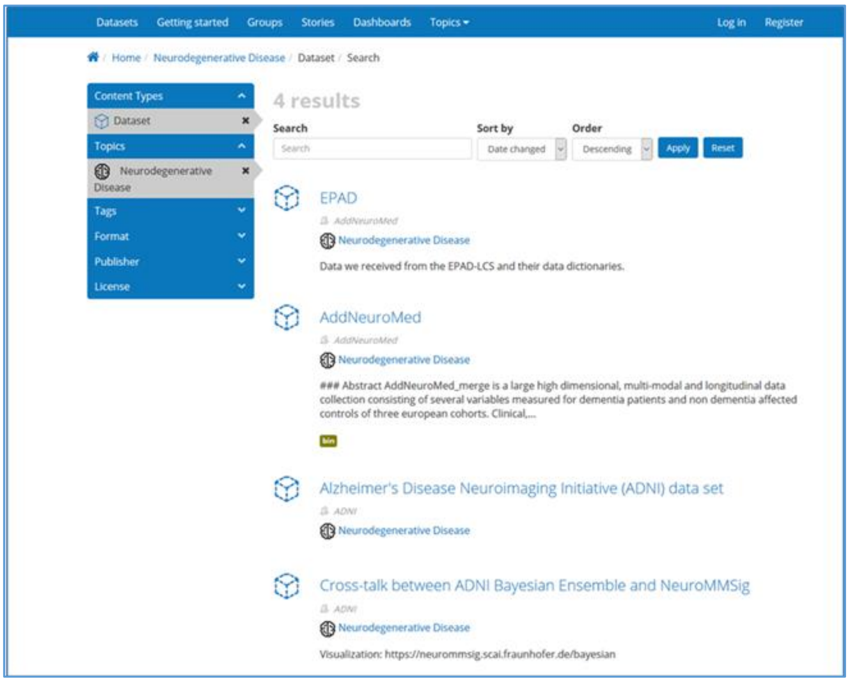
Table 1: AData cohort studies and related key information (table taken from https://adata.scai.fraunhofer.de/). Note: ANMerge which was used in the AData-Viewer is a new version of the AddNeuroMed dataset[4].

[4] Birkenbihl C. et al., ANMerge: A comprehensive and accessible Alzheimer's disease patient-level dataset. J Alzheimers Dis. 2020 Dec 1. doi:10.3233/JAD-200948.

Most of the datasets we accessed were shared after going through an official data request process. Ultimately, Fraunhofer SCAI (mostly restricted to single persons and not accessible for the whole group) were granted access to nine distinct AD cohort datasets (Table 1). Datasets from other studies can be integrated in future. Datasets from other studies can be integrated in future.

4.2. Storage of study data in the DKAN Clinical Data Repository

The Fraunhofer DKAN platform contains clinical cohort data from several studies, currently ADNI (Alzheimer's Disease Neuroimaging Initiative), ANM (AddNeuroMed), and EPAD (European Prevention of Alzheimer's Dementia), as shown in Figure 2. Additional cohort datasets imported to the AData viewer (described in the section 4.1.) will be integrated into the DKAN platform in the near future.



4 results

EPAD
AddNeuroMed
Neurodegenerative Disease
Data we received from the EPAD-LCS and their data dictionaries.

AddNeuroMed
AddNeuroMed
Neurodegenerative Disease
Abstract AddNeuroMed_merge is a large high dimensional, multi-modal and longitudinal data collection consisting of several variables measured for dementia patients and non dementia affected controls of three european cohorts. Clinical,...

Alzheimer's Disease Neuroimaging Initiative (ADNI) data set
ADNI
Neurodegenerative Disease

Cross-talk between ADNI Bayesian Ensemble and NeuroMMSig
ADNI
Neurodegenerative Disease
Visualization: https://neurommsig.scai.fraunhofer.de/bayesian

Figure 2: Fraunhofer DKAN cohort platform

The Fraunhofer DKAN system has been proved as an appropriate, easy-to-use platform to provide study data including metadata annotations. Note: The Fraunhofer DKAN system in house system that can only be accessed from internal server presumably access authorization is given before.

For additional information on DKAN see supplementary.

4.3. Generation of an AD Data Catalog with harmonized annotations using the AData Viewer

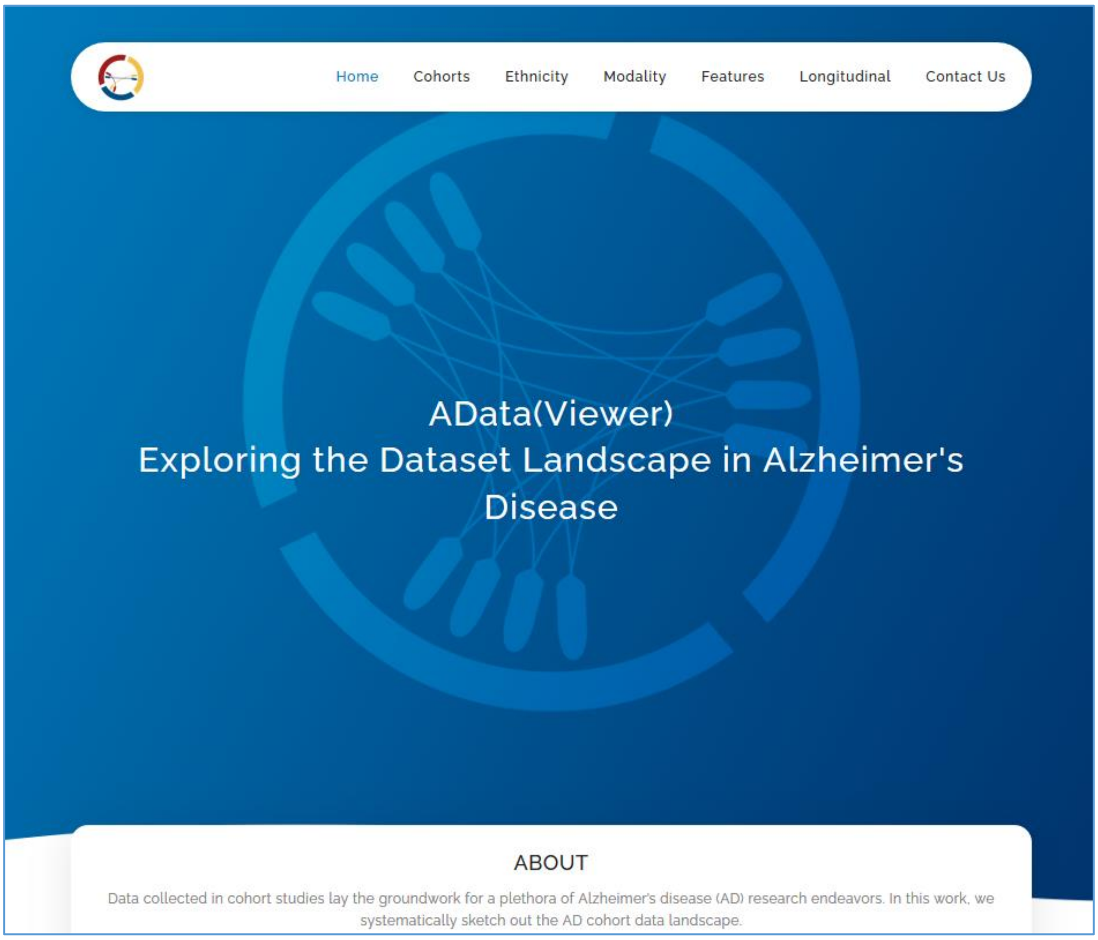
A comprehensive data-driven view on the AD data catalog (AData landscape) including collection and the categorization of annotations is the basis for the generation of a project specific data catalog with harmonized annotations.

Major AD cohort studies were chosen to allow for a thorough investigation of the data landscape in relation to measured parameter and used attributes. Most of the datasets we accessed were shared after going through an official data request process. Ultimately, Fraunhofer SCAI got access to nine distinct AD cohort datasets (Table 1).

We investigated major AD cohort datasets with the aim of characterizing their underlying data, assessing the quantity and availability of data, and evaluating the interoperability across these distinct cohort datasets.

This approach captures information on study related variables (attributes) with focus on a (mostly) semantic mapping of specific parameter ranging from ethnicity to diagnostics. The system enables a systematic view on the data and variables and comparisons across different studies.

The web-application (**AData Viewer** https://adata.scai.fraunhofer.de/) allows for exploration of the AD data landscape and helps researchers to identify the most suitable datasets for their projects. It provides an overview of the AD data landscape by investigating clinical data from nine major clinical AD cohort studies (Fig. 3).



AData(Viewer)
Exploring the Dataset Landscape in Alzheimer's Disease

ABOUT

Data collected in cohort studies lay the groundwork for a plethora of Alzheimer's disease (AD) research endeavors. In this work, we systematically sketch out the AD cohort data landscape.

Figure 3 Screenshot of the AData Viewer (https://adata.scai.fraunhofer.de/)

Evaluating a data landscape involves organizing and comparing datasets in order to **qualitatively** assess their collected data modalities and variables, and **quantitatively** describe the demographics of the study population and distributions of measured variables. Such characterization provides a detailed overview of the data usability and accessibility. Finally, evaluating a data landscape inherently exposes already potential flaws with regard to the interoperability between existing datasets.

It is important to be aware that not all selected studies followed the same design nor goals. Each study enforced its own recruitment criteria and enrolled participants following distinct selection processes. While some aimed for a case-control setting and included a substantial amount of AD patients into their cohort, others deliberately excluded them to focus on early disease progression. Thereby, the cohort datasets are all subject to inherent biases.

For additional technical details see our public TVB-Cloud deliverable: “Paper manuscript describing the initial characterization of the studies, the equivalence of their variables the summary statistics”[5] (Birkenbihl et al., 2020).

By applying the described procedures, the addition of further data sets provided by TVB-Cloud project partners into the existing database could be easily performed, enabling the compilation and harmonization of study variables and comparison against the integrated AD data landscape.

This approach makes FAIR data principles actionable and enables comparison or aggregation of clinical studies. Here a common clinical data model was generated, the ground for interoperability and re-usability.

4.4. Generation of a common clinical data model for harmonisation of clinical data

One promising approach to increase dataset interoperability is the creation of a comprehensive, AD-specific common clinical data model. Such a data model supports the alignment and mapping of variables by providing easy-to-follow guidelines and a dedicated interface for retrospective data harmonization.

The Common Clinical Data Model (CCDM) described here, enables the mapping and alignment of attributes on the level of individual patient data and the upload of the individual patient data based on the mapping. One challenge in harmonizing data from multiple sources is achieving a machine-readable integration of variables including their parameters. Due to specific study designs and individual modelling of recorded measurements, both the composition and encoding of variables (e.g., data types, value ranges or units) differs largely between datasets. A harmonization of given data down to formatting can be performed via the CCDM, thus supporting further cross-study analyses (see section 6: “Clinical Data Viewer”). Moreover, the CCDM approach enables structured analysis of captured meta data, i.e., a comparison of individual studies’ data landscapes.

4.4.1. Description of the common clinical data model (overview)

By its sub-structure, the CCDM reflects the tripartition into a) a core set of internal variables, b) a collection of mappings dedicated to particular external data resources as well as c) the data itself. These three levels refer to each other and are shortly described in the following sections. See Appendix for details, including a UML diagram.

The integration of new sources of clinical data into the CCDM requires two subsequent multi-step processes, preferably executed by a data steward. An overview is shown in Fig. 4.

[5] https://cordis.europa.eu/project/id/826421/results

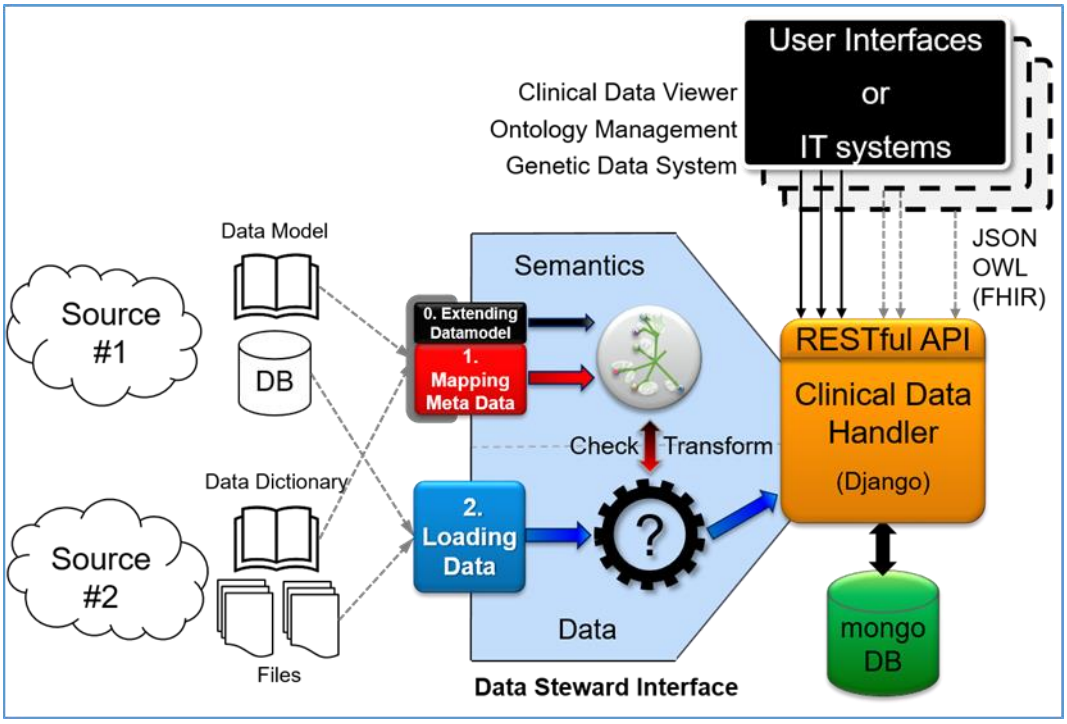


Figure 4: Overview on the data integration process. External sources get integrated by both meta data and data. Integrating elements of external data models or dictionaries requires a formatted mapping onto CCDM-internal variables (red box). If necessary, the data model's core (gray) might be extended with new variable declarations (black box). When integrating actual data (blue box) from the respective sources, each data point will be checked for compliance with the modelled rule sets (core + mappings) before integration into a clinical data handling backend (orange), communicating with a mongoDB instance (green). Where necessary, data will be transformed according to the source-specific mapping set. Finally, integrated data appears clean and queryable via a RESTful API, e.g. supplying the Clinical Data Viewer with data.

Integration of data from various sources starts with mapping the given meta data onto the common clinical data model. Resulting assignments and rules get implemented automatically in the central data handling unit (clinical backend). Similarly, data will be loaded into this central, database-driven environment; here, tabular data will be parsed, filtered and quality controlled according to the setup source-specific declarations encoded in data model's core and mapping sets. Both integration processes are supported by dedicated web interfaces (see 4.4.2 Description of the Data-Steward), each implemented as a form-driven file uploads (bulk import) or stepwise dialog-driven wizards.

4.4.2. Description of the Data-Steward

This service provides a user interface for several features described before such as editing and updating the CCDM in relation to new studies that will be added and the final data upload for further (meta)-analysis using software tools such as the Clinical DataViewer (see section 6).

Welcome to the IDSN Data-Steward.

IDSN

Integrative Datasematics in Neurodegenerative Diseases

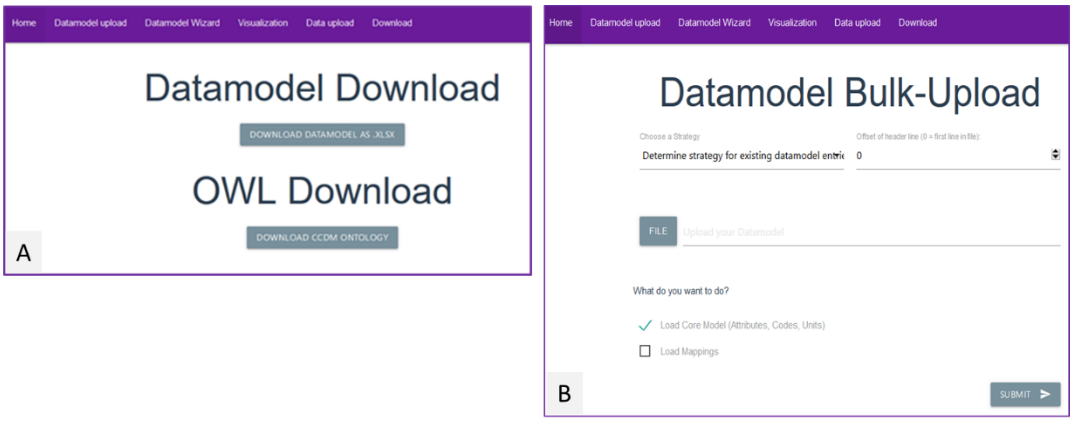
Data Steward Version 1.0

Useful Links

OLS
Clinical Viewer
DZNE
IDSN Homepage

Figure 7: Data-Steward entry page. Screenshot taken from http://idsn.dzne.de/data-steward.

The Data-Steward services supports mapping and alignment of the attributes, definition of new attributes (“Datamodel Wizard”), visualization of the study related attributes and their relations (“Visualization”) and the upload of data in the CCDM (“Data-upload”). In fact, the data model could be downloaded in .xlsx format (“Download”), edited e.g., by adding or mapping new study related attributes and uploaded back to update the system (“Datamodel upload”) (Figure 8).



Datamodel Download

OWL Download

Datamodel Bulk-Upload

Choose a Strategy: Determine strategy for existing datamodel entrie

Offset of header line (0 = first line in file): 0

Upload your Datamodel

What do you want to do?

✓ Load Core Model (Attributes, Codes, Units)

□ Load Mappings

Figure 8: Data-Steward pages for up- and download of the model. Download datamodel as .xlsx or .owl file **(A)**. The .xlsx file could be used as template for curation of new data sets and subsequent upload into the model using the upload module “Datamodule upload” **(B).**

A key feature of the Data-Steward is the **graph visualization** of the data model (Figure 9). Edges in that graph represent the connections between different sources and its attributes as well as the mappings between those attributes.

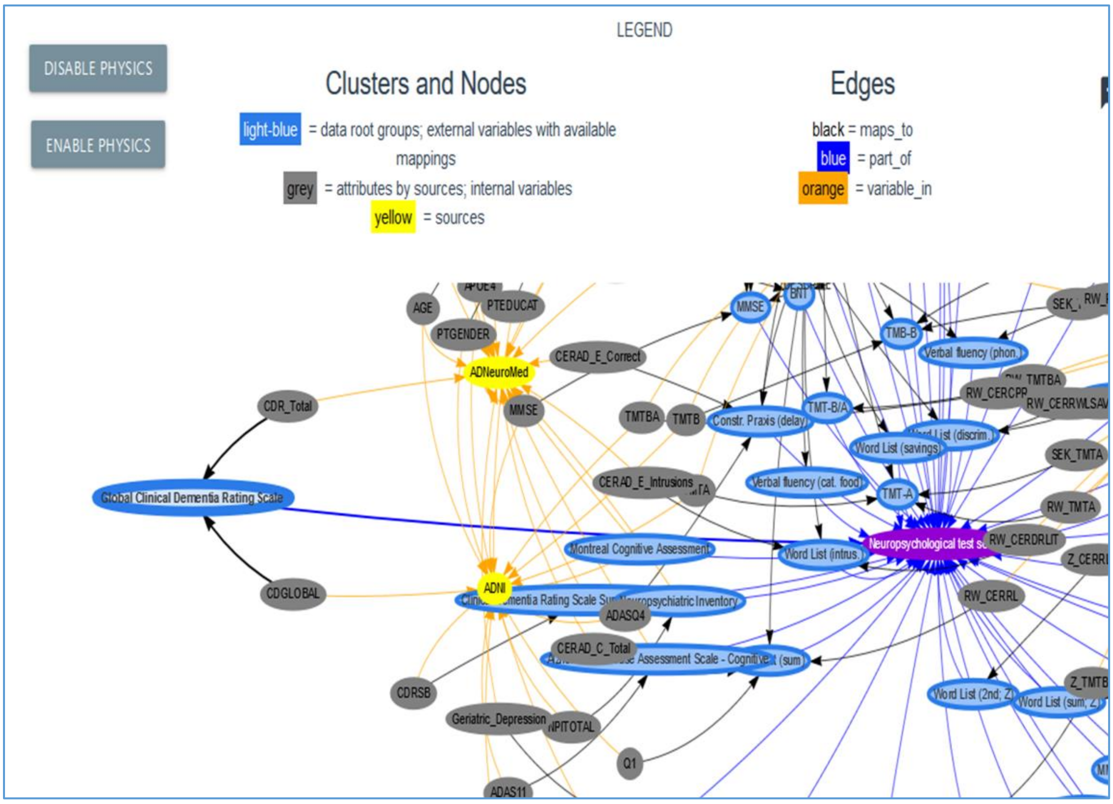


Figure 9: Graph Visualization of the Data Model. The "CDR_Total" variable from the ADNeuroMed study and the CDGLOBAL variable from ADNI study are mapped to CCDM variable "Global Clinical Dementia Rating Scale", which is part of the "Neurophysiological test scores" in CCDM.

An additional functionality in the Data-Steward ("SHOW MORE IN OLS"; right-click context menu on nodes) enables sematic normalization of concepts to EBI – OLS[6] (Fig. 10). Within a TVB-Cloud version of the CCDM this function could be linked to Fraunhofer SCAI OLS containing TVB-Cloud specific ontologies for concept harmonization.

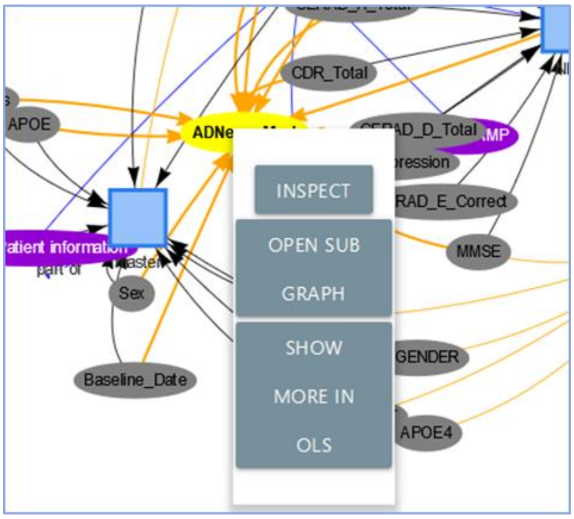


Figure 10: Concept harmonization via OLS

[6] https://www.ebi.ac.uk/ols/search?q=sources_ADNI

5. Results

5.1. AD Data Catalogue (AD Data Landscape)

Data collected in cohort studies lay the groundwork for a plethora of Alzheimer's disease (AD) research endeavors. Therefore, the AD cohort data landscape was systematically sketched out.

Former undertakings attempted to evaluate the AD data landscape solely on the basis of metadata and literature, without investigating the underlying patient-level data. However, reviewing study protocols can only explain the original design of a given study and thereby neglects unforeseen changes in procedures or participant recruitment throughout study runtime. The most comprehensive and granular approach is a patient-level and data-driven evaluation of the AD data landscape, which is a tedious and time-consuming endeavor and should better be performed by the data owner.

In this part of the work, nine of the major clinical cohort study datasets available in the AD field were traced down, accessed, investigated, and compared. We comprehensively describe the acquired data and show which data modalities were identified in the datasets as well as their overlap within the studies investigated. More detailed analysis e.g., on the longitudinal follow-up on biomarker-level in order to explore to what extent current AD data cover the progression of the disease are described in TVB-Cloud public deliverable "Evaluating the Alzheimer's Disease Data Landscape", Birkenbihl et al., 2020[7].

In addition, we made all analysis and results available through an interactive web-portal (https://adata.scai.fraunhofer.de), such that researchers can explore the AD data landscape, which correspond to a data catalogue on available AD related studies, with respect to the investigated variables.

The prerequisite for working across multiple datasets is interoperability with regard to their features. This aspect includes the availability of features, the same naming conventions and comparable representations. While we explored the representations of important features in the "Cohorts" section, here we focus on naming conventions and to conceptually map similar features to their corresponding counterparts in other datasets. This work is not a complete mapping of all features available in these cohorts. It serves as an overview on the current state of dataset interoperability and showcases that there is no common data model for dementia datasets in use. This hinders data-driven approaches across cohorts substantially in the summary of the mapping

In the following, some detailed results are provided. The selection and amount of data modalities measured in a cohort study often depend on the study's aims and available funding. Thus, often only partially overlapping sets of data modalities are assessed in distinct cohort studies. To analyze, which data is available in our investigated cohorts and to explore the overlap between them, we assessed the grade of availability per data modality. Therefore, each dataset was manually curated according to defined criteria (Supplementary Table 1).

Figure 11 shows an overview of the data modalities and their availability score in all acquired cohort datasets. Commonly assessed throughout all studies were demographic variables (e.g., participant age, sex, and education), as well as some clinical assessments such as MMSE (Mini-Mental State Examination).

[7] https://cordis.europa.eu/project/id/826421/results

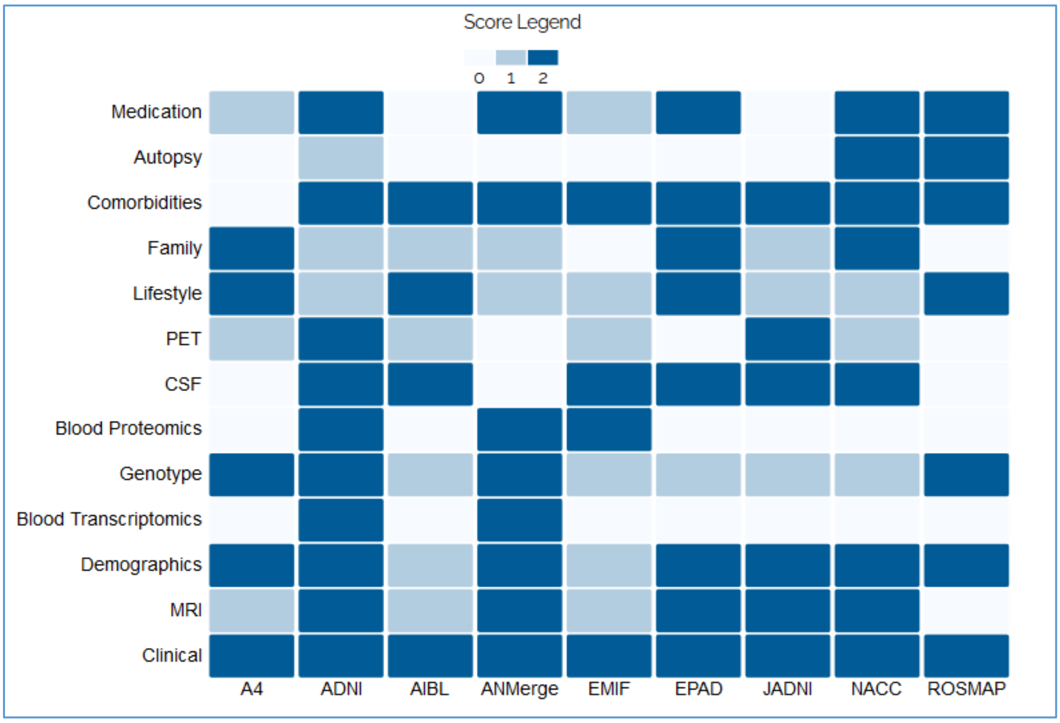
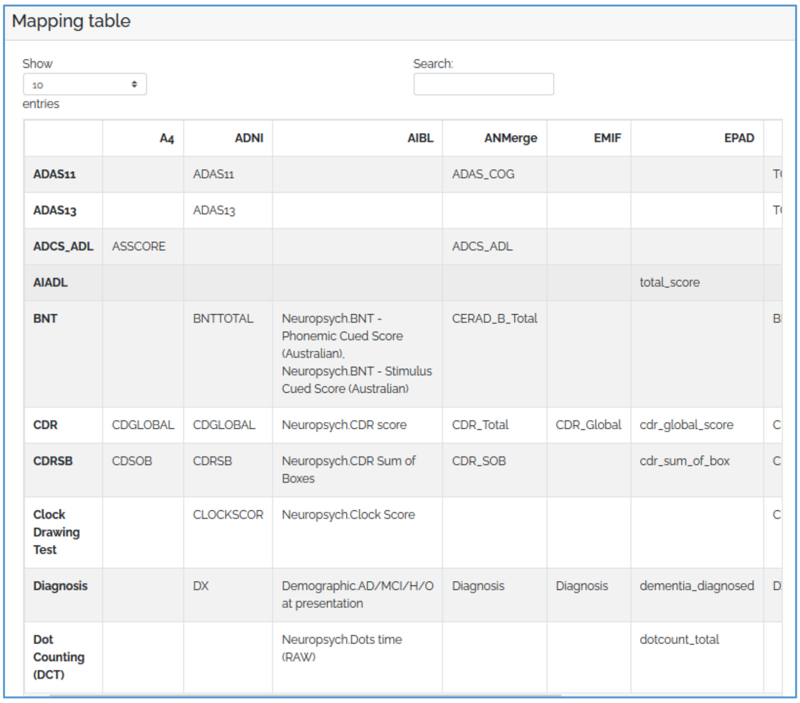


Figure 11: The heatmap shows the availability scores we assigned to denote how completely a certain data modality is available and shared in the investigated cohort datasets. The corresponding criteria for score assignments are shown in supplementary table. PET: Positron emission tomography, CSF: Cerebrospinal fluid, MRI: Magnetic resonance imaging, Family, Blood Transcrip.: Transcriptomic data gathered from blood samples (Figure taken from https://adata.scai.fraunhofer.de/).

Table 2 shows a section of the variable mappings from all studies investigated. Comprehensive information on all mapped variables is given in the AData Viewer (https://adata.scai.fraunhofer.de/feature_comparison)



Mapping table

Show 10 entries

Search:

	A4	ADNI	AIBL	ANMerge	EMIF	EPAD	
ADAS11		ADAS11		ADAS_COG			T
ADAS13		ADAS13					T
ADCS_ADL	ASSCORE			ADCS_ADL			
AIADL						total_score	
BNT		BNTTOTAL	Neuropsych.BNT - Phonemic Cued Score (Australian), Neuropsych.BNT - Stimulus Cued Score (Australian)	CERAD_B_Total			B
CDR	CDGLOBAL	CDGLOBAL	Neuropsych.CDR score	CDR_Total	CDR_Global	cdr_global_score	C
CDRSB	CDSOB	CDRSB	Neuropsych.CDR Sum of Boxes	CDR_SOB		cdr_sum_of_box	C
Clock Drawing Test		CLOCKSCOR	Neuropsych.Clock Score				C
Diagnosis		DX	Demographic.AD/MCI/H/O at presentation	Diagnosis	Diagnosis	dementia_diagnosed	D
Dot Counting (DCT)			Neuropsych.Dots time (RAW)			dotcount_total	

Table 2: Screenshot of the Mapping Table for clinical parameter. Comprehensive information for all variables is available under https://adata.scai.fraunhofer.de/feature_comparison.

The same information on mapped variables from studies investigated in relation to different modalities such as clinical, demographics, lifestyle or comorbidities is visualized in Figure 12 which provides an overview on availability and pairwise relations of clinical parameter within the investigated studies. Figure 12 is a screenshot taken from https://adata.scai.fraunhofer.de/feature_comparison. The AData Viewer website provides an inactive view on all harmonized modalities.

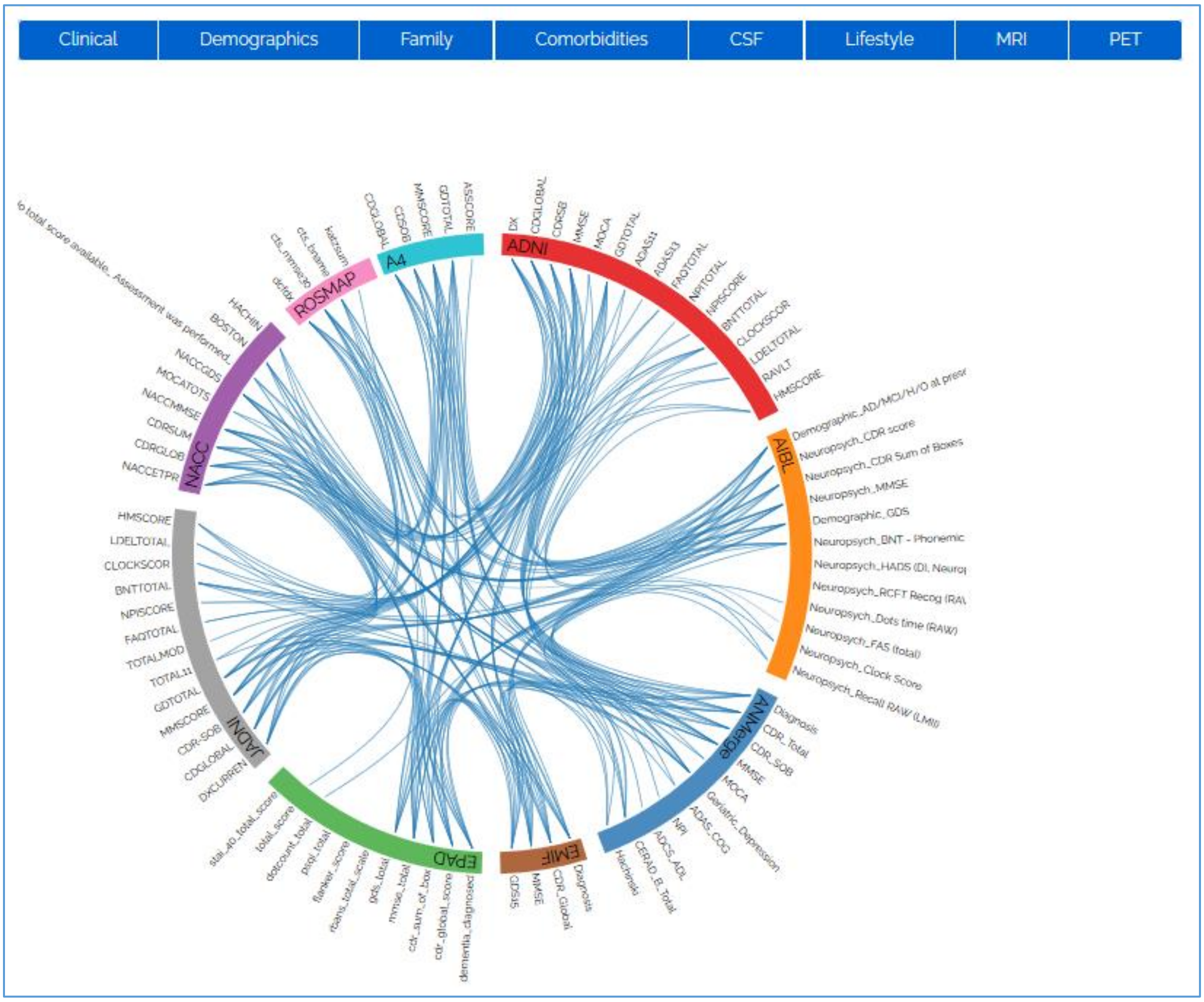
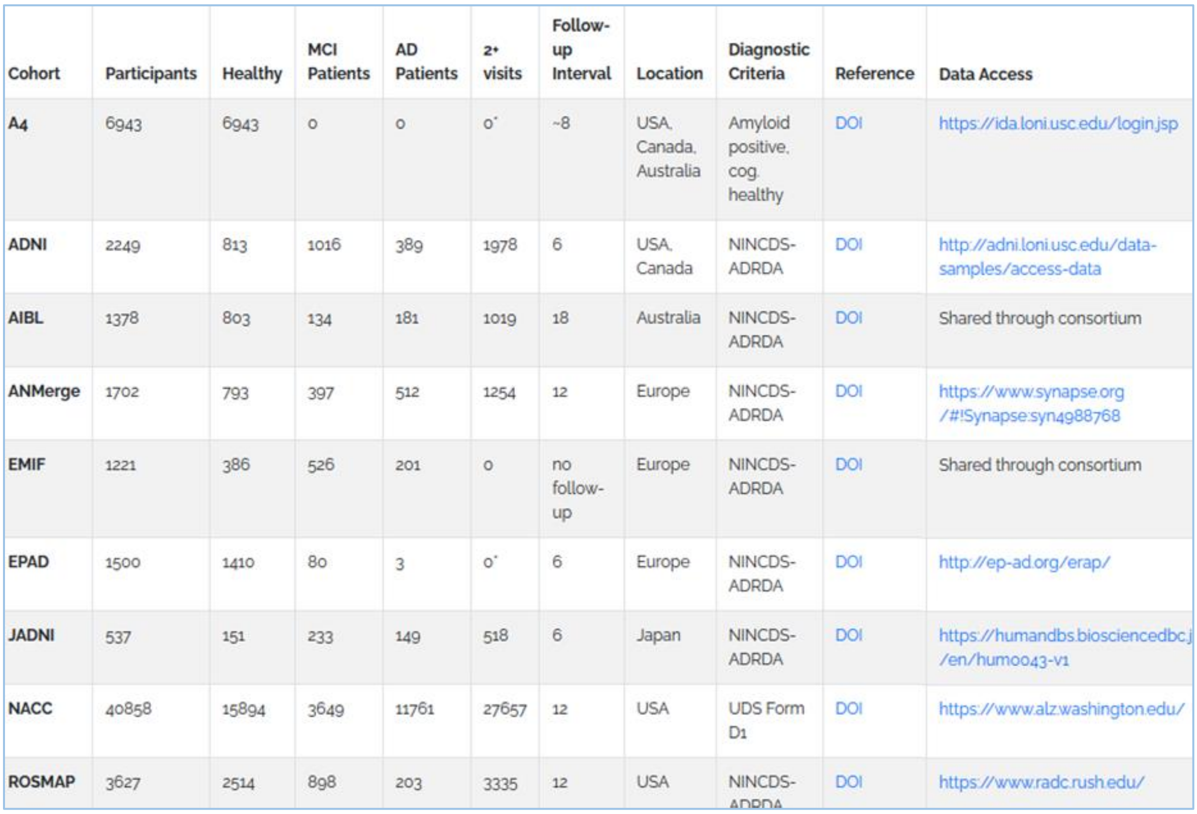


Figure 12: Overview on availability and relation of clinical parameter within the investigated studies. Screenshot taken from https://adata.scai.fraunhofer.de/feature_comparison. The website provides an interactive view on all modalities investigated.

Although the purpose of the work is to provide a comprehensive overview on the availability of data per modality, the presented results of this analysis are strongly dependent on our defined curation criteria, and different criteria could lead to different results. Additionally, all investigated datasets could potentially hold more information than presented in the current version of the AData Viewer. The AData Viewer is an ongoing project that will be further updated and substantiate by new features and the addition of new data and datasets.

Detailed analyses are described in the AData Viewer and in TVB-Cloud public deliverable "Evaluating the Alzheimer's Disease Data Landscape", Birkenbihl et al., 2020[8]. Here we show an example on deviation of cognitively impaired patients from healthy controls in the different cohorts.

To evaluate how severely patients from each cohort have been affected by AD, we compared the distributions of both cognitive outcomes and key biomarkers for the cognitively affected patient subgroups (i.e., participants with Mild Cognitive Impairment (MCI) or AD diagnosis). Table 3 shows the distributions for each complete cohort including healthy controls, MCI and AD patients. Detailed data per diagnosis subgroup are given in Table 4. Comprehensive information on all modalities investigated can be found at https://adata.scai.fraunhofer.de/cohorts.

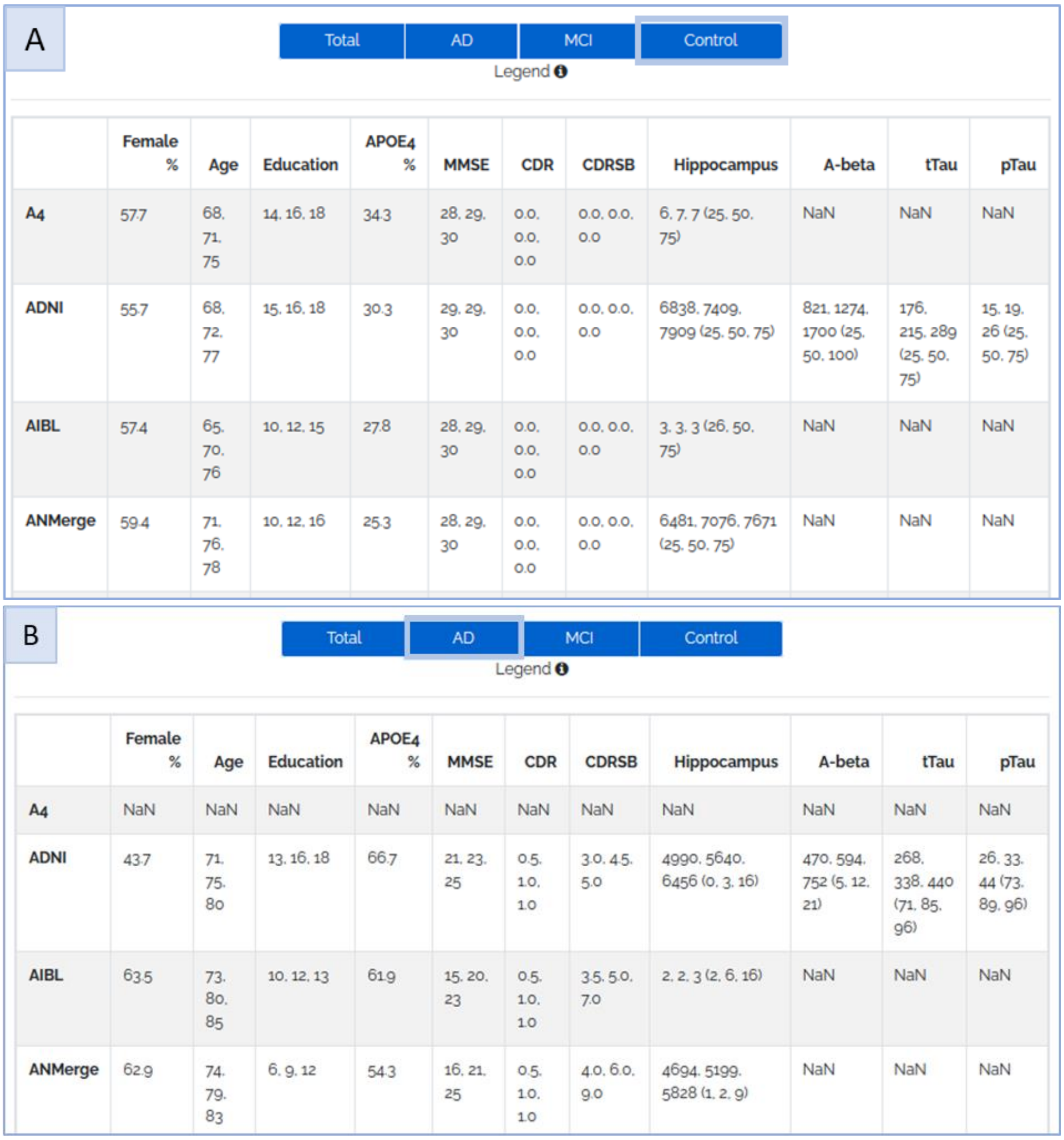


Cohort	Participants	Healthy	MCI Patients	AD Patients	2+ visits	Follow-up Interval	Location	Diagnostic Criteria	Reference	Data Access
A4	6943	6943	0	0	0*	~8	USA, Canada, Australia	Amyloid positive, cog. healthy	DOI	https://ida.loni.usc.edu/login.jsp
ADNI	2249	813	1016	389	1978	6	USA, Canada	NINCDS-ADRDA	DOI	http://adni.loni.usc.edu/data-samples/access-data
AIBL	1378	803	134	181	1019	18	Australia	NINCDS-ADRDA	DOI	Shared through consortium
ANMerge	1702	793	397	512	1254	12	Europe	NINCDS-ADRDA	DOI	https://www.synapse.org/#!Synapse:syn4988768
EMIF	1221	386	526	201	0	no follow-up	Europe	NINCDS-ADRDA	DOI	Shared through consortium
EPAD	1500	1410	80	3	0*	6	Europe	NINCDS-ADRDA	DOI	http://ep-ad.org/erap/
JADNI	537	151	233	149	518	6	Japan	NINCDS-ADRDA	DOI	https://humandbs.biosciencedbc.j/en/hum0043-v1
NACC	40858	15894	3649	11761	27657	12	USA	UDS Form D1	DOI	https://www.alz.washington.edu/
ROSMAP	3627	2514	898	203	3335	12	USA	NINCDS-ADRDA	DOI	https://www.radc.rush.edu/

Table: 3: Overview on the cohort studies included in AData Viewer. Number of recruited participants and key information on the respective study design is given. Additionally, links are provided which facilitate access to the underlying datasets. Screenshot taken from https://adata.scai.fraunhofer.de/.

According to the MMSE scores, AD patients from AIBL (Quantiles: 15, 20, 23), AddNeuroMed (Quantiles: 16, 21, 25) showed the worst cognitive performance. ADNI (Quantiles: 21, 23, 25) contained patients with fewer cognitive symptoms. The CDR sum of boxes scores (CDRSB) slightly shifts the perspective. Here, AddNeuroMed is the most affected cohort with its 25%, 50% and 75% quantiles of the CDRSB scores being 4, 6 and 9 respectively. AIBL patients scored 3.5, 5, 7, which slightly contradicts the image painted by the MMSE scores. Again, ADNI shows the least cognitive symptoms with its CDR-SB quantiles being 3, 4.5, 5, (Table 4).

[8] https://cordis.europa.eu/project/id/826421/results



A

	Female %	Age	Education	APOE4 %	MMSE	CDR	CDRSB	Hippocampus	A-beta	tTau	pTau
A4	57.7	68, 71, 75	14, 16, 18	34.3	28, 29, 30	0.0, 0.0, 0.0	0.0, 0.0, 0.0	6, 7, 7 (25, 50, 75)	NaN	NaN	NaN
ADNI	55.7	68, 72, 77	15, 16, 18	30.3	29, 29, 30	0.0, 0.0, 0.0	0.0, 0.0, 0.0	6838, 7409, 7909 (25, 50, 75)	821, 1274, 1700 (25, 50, 100)	176, 215, 289 (25, 50, 75)	15, 19, 26 (25, 50, 75)
AIBL	57.4	65, 70, 76	10, 12, 15	27.8	28, 29, 30	0.0, 0.0, 0.0	0.0, 0.0, 0.0	3, 3, 3 (26, 50, 75)	NaN	NaN	NaN
ANMerge	59.4	71, 76, 78	10, 12, 16	25.3	28, 29, 30	0.0, 0.0, 0.0	0.0, 0.0, 0.0	6481, 7076, 7671 (25, 50, 75)	NaN	NaN	NaN

B

	Female %	Age	Education	APOE4 %	MMSE	CDR	CDRSB	Hippocampus	A-beta	tTau	pTau
A4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ADNI	43.7	71, 75, 80	13, 16, 18	66.7	21, 23, 25	0.5, 1.0, 1.0	3.0, 4.5, 5.0	4990, 5640, 6456 (0, 3, 16)	470, 594, 752 (5, 12, 21)	268, 338, 440 (71, 85, 96)	26, 33, 44 (73, 89, 96)
AIBL	63.5	73, 80, 85	10, 12, 13	61.9	15, 20, 23	0.5, 1.0, 1.0	3.5, 5.0, 7.0	2, 2, 3 (2, 6, 16)	NaN	NaN	NaN
ANMerge	62.9	74, 79, 83	6, 9, 12	54.3	16, 21, 25	0.5, 1.0, 1.0	4.0, 6.0, 9.0	4694, 5199, 5828 (1, 2, 9)	NaN	NaN	NaN

Table 4: Demographic characteristics of the cohorts. Distributions of important Alzheimer's disease biomarkers at study baseline. Distributions of demographic and biomarker values are described using the 25, 50 and 75% quantiles of the raw measured values. To make absolute values of biomarker measures more comparable, we additionally report the deviation between the healthy control group of one respective cohort and a selected participant subgroup of the same study (i.e. all, healthy, MCI and AD). This deviation is given as the quantiles received when evaluating the raw value quantiles under the empirical cumulative distribution function (ECDF) of the control group of the respective cohort. Table 4A and 4B show results from control group and AD group, respectively. Screenshot taken from https://adata.scai.fraunhofer.de/.

A comparison of raw biomarker measurements between cohorts proved to be impossible, since encountered values are on different scales and may be subject to batch effects. Thus, we analyzed how much measurements diverged from their respective control population in each cohort.

The prerequisite for comparative approaches involving biomarker measurements across datasets is an alignment of their underlying data models (i.e., making data interoperable). In this analysis, each study had defined its own data model and variable names differed between them. This forced us to individually map variables to their corresponding counterparts in other studies to enable comparisons in the first place (e.g., combine "lh_hippo_volume" and "rh_hippo_volume" and map to "Hippocampus").

Another difficulty is that numerous datasets reported values of equivalent variables in different ways. For example, CSF biomarker measurements are reported to be either normal (0) or abnormal (1) in NACC (National Alzheimer's Coordinating Center Study), while other studies provide numerical values, which themselves were capped at different thresholds in some of those studies (e.g., ">1700"). All these factors led to a severe lack of interoperability between datasets which significantly limits comparative approaches and restricts them to more standardized variables like clinical assessment scores.

Our evaluation exposed critical limitations in the AD data landscape that impede comparative approaches across multiple datasets. Comparing our results to those gained by metadata-based approaches highlights that thorough investigation of real patient-level data is inevitable to assess a data landscape.

In summary, assessing the availability of data modalities in clinical cohort datasets involves intensive and meticulous manual curation of the acquired datasets and thereby, the definition of applicable curation criteria specifying under which circumstances each data modality is considered as "available". Furthermore, it is often necessary to define a gradual categorization to represent the degree of availability.

Our analysis exposed major challenges that severely impede comparative approaches on AD cohort data. The investigated cohort datasets neither followed a common naming system for variables, nor represented values of the same measurement in equal manner. On top of that, some studies only shared processed values instead of the underlying raw data. This further impedes interoperability since differences in applied processing pipelines inevitably introduce systematic biases into the data.

In contrast, a patient-level and data-driven evaluation 1) is not subject to these assumptions, 2) allows for a quantitative investigation of important cohort statistics and 3) illustrates the amount and quality of the data accessible to the field.

A promising approach to increase dataset interoperability is the here provided and described comprehensive **common data model** that facilitates the alignment and mapping of variables for acquired datasets.

5.2. Common Clinical Data Model

Generating a data catalog in relation to the available AD related studies ("AD data landscape"), including identification and mapping by variable (attribute) name and comparing data from different studies was set up by the AData Viewer and described in sections 4.3. and 5.1.

As pointed out there, a common data model that would facilitate the alignment and mapping of variables for acquired datasets is a promising approach to increase dataset interoperability. Comparability and interoperability is a major issue, also within FAIR data handling, when comparing clinical datasets. Therefore, a comprehensive harmonization down to single patient data is necessary.

Here we describe the **Common Clinical Data Model** (CCMD) that was developed for mapping and harmonization of patient level data from diverse clinical trials to enable cross studies analysis, including automatic analysis performed by computer.

Initially, the system was set up within the IDSN project (https://www.idsn.info/de/idsn.html). Here, the primary aim was the integration of the multi-centric DZNE study 'DELCODE', the DESCRIBE patient registry and analogous clinical routine data from local hospital IT systems, thereby generating a basic set of general clinical trial related attributes such as patient ID or Sex; ethnicity and a core set of attributes (variables) related to dementia research ranging from “father has dementia” to “Global Dementia Rating Scale”, according to exemplary needs of the clinical research community.

As a proof of concept for the usability of the system within TVB-Cloud, in order to enable the integration of diverse types of NND-related studies in particular the datasets in TVB-Cloud, we initially mapped two international dementia studies; namely the Alzheimer Disease Neuroimaging Initiative (ADNI)[9] and the AddNeuroMed collaboration. In difference to the AData Viewer, which also contains those two studies the study-related attributes were aligned to the CCDM on the level of patient data and should thereby become comparable and computable.

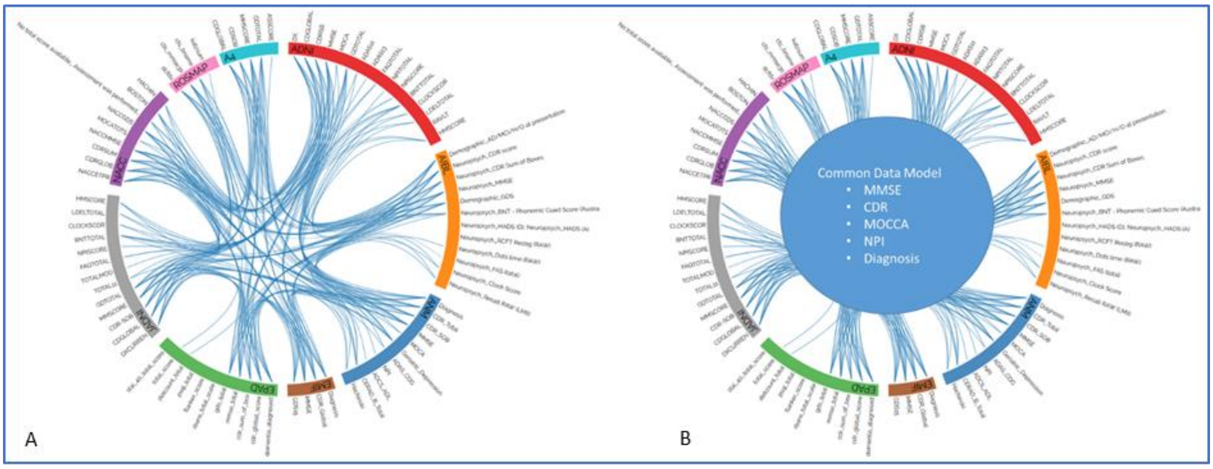


Figure 13: Extension of the AData Viewer (A) to a common data model (B).

Many clinical studies have been performed over the last years with the focus on NDD. Interoperability of variables is indispensable to enable comparative analysis on measured parameters. That includes common parameters such as age and sex as well as disease specific measurement entities such as MMSE. Most often only pairwise comparison between different studies (e.g., comparing new studies to ADNI data, Westmann et al., 2011[10], TVB-Cloud public deliverable “Evaluating the Alzheimer’s Disease Data Landscape”, Birkenbihl et al., 2020[11]) are made (Fig. 13A), but reliable meta-analysis based on patient data are not possible because of missing interoperability of the data variables (e.g., Balsis et al., 2015[12]). Applying our CCDM, i.e., mapping variables from different studies to a common data model, should enable arbitrary comparisons between the integrated studies (Fig. 13B).

The CCDM contains features that allows a semi-automated mapping of the attributes from different studies (see supplementary: “Description of the Common Clinical Data Model (CCDM)” for details). Here, mapping of study-related attributes was done manually, using the dedicated template from the

[9] Mueller S.G. et al.Ways toward an early diagnosis in Alzheimer’s disease: the Alzheimer’s Disease Neuroimaging Initiative (ADNI). (2005) Alzheimers Dement (N Y). 1(1):55-66.

[10] Westman, E. et al. AddNeuroMed and ADNI: similar patterns of Alzheimer's atrophy and automated MRI classification accuracy in Europe and North America. (2011) NeuroImage 58, 3: 818

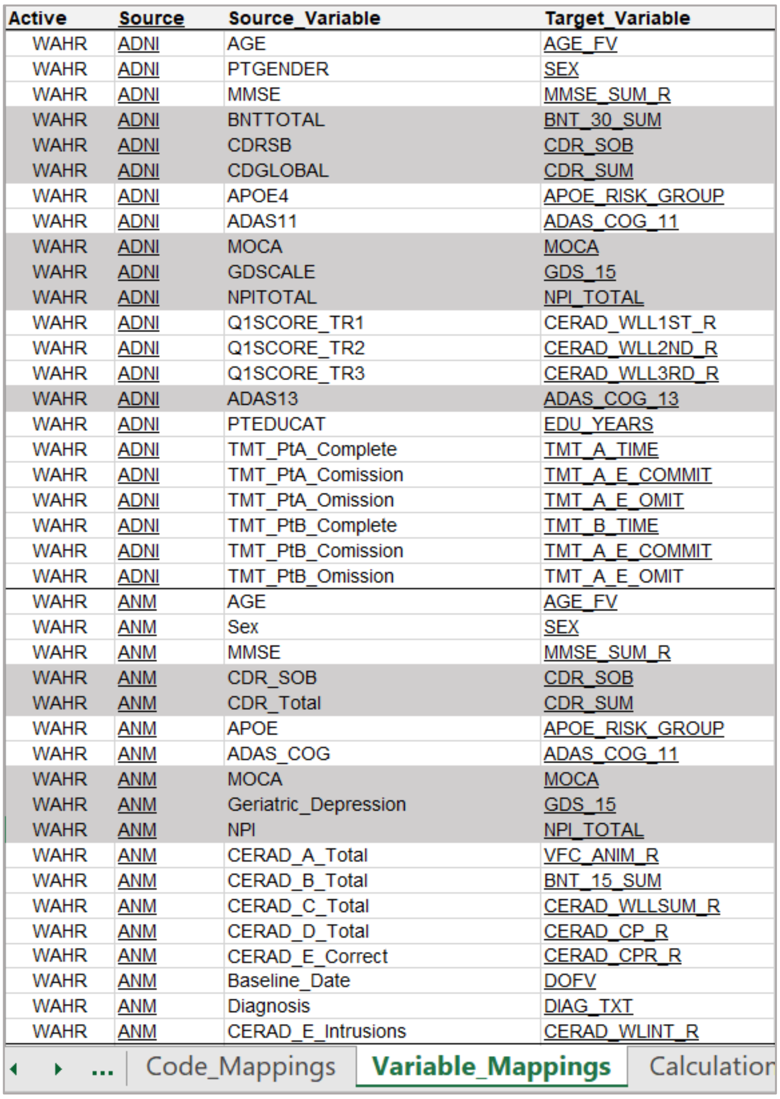
[11] https://cordis.europa.eu/project/id/826421/results

[12] Balsis S. et al. How Do Scores on the ADAS-Cog, MMSE, and CDR-SOB Correspond? (2015) Clin Neuropsychol. 29(7):1002

Data-Steward that enable editing and subsequent uploading of the variable mapping to the data model ("bulk upload").

Four different possible settings need to be considered for variables mapping. **First**, attributes of the curated studies are simply *equal* to the attributes of the CCDM and have to be aligned by name, only. This is often true for basic parameters such as age, sex or patient ID, but also for disease-specific, standardized measurements such as the MMSE (Mini-Mental State Examination) score. **Second**, although the attributes could be semantically aligned (equal meaning), the variables encodings of the measurements differ from the CCDM analogues by e.g., value ranges or default units. Thus, defining an adequate transformation statement is necessary in order to establish a computable link between internal and external variables. Transformations comprise mathematical conversions (numerical variables), named mappings between known states (e.g., conversion from named months to their numerical equivalents: 'Jan' → '1'). Similarly, various encoding styles of ApoE get harmonized this way and assigned to basic interpretations like risk groups (high, medium, low), according to measured genotypes (two, one or no ApoE ε4 alleles). **Third**, it might be necessary to consult multiple variables from the data source and calculate the target value – a simple example is the patient's age at a certain measurement. For instance, the age of onset of a disease could be inferred by offsetting the date of the first symptoms with the date of birth. Obviously, such calculations have to be carried out as post-processing, as all necessary inputs have to be imported first. **Fourth**, attributes from the new study that is supposed to be mapped to the CCDM cannot be aligned to an existing attribute at all. The reason could be that the related endpoint was not covered yet by the given CCDM (e.g., NPI, Neuropsychiatric Inventory) or a full harmonization is principally not possible. The latter case is often true for items which show in detail deviating meanings (e.g., BNT, Boston Naming Test, see below), that could therefore not be aligned. In both cases, new attributes need to be defined, thus increasing the CCDM's overall semantic coverage.

In the first approach, we mapped 15 variables from ADNI and 13 variables from ANM to the existing common data model and added seven new variables for NDD-specific clinical measurements (Table 5).



Active	Source	Source_Variable	Target_Variable
WAHR	ADNI	AGE	AGE_FV
WAHR	ADNI	PTGENDER	SEX
WAHR	ADNI	MMSE	MMSE_SUM_R
WAHR	ADNI	BNTTOTAL	BNT_30_SUM
WAHR	ADNI	CDRSB	CDR_SOB
WAHR	ADNI	CDGLOBAL	CDR_SUM
WAHR	ADNI	APOE4	APOE_RISK_GROUP
WAHR	ADNI	ADAS11	ADAS_COG_11
WAHR	ADNI	MOCA	MOCA
WAHR	ADNI	GDSCALE	GDS_15
WAHR	ADNI	NPITOTAL	NPI_TOTAL
WAHR	ADNI	Q1SCORE_TR1	CERAD_WLL1ST_R
WAHR	ADNI	Q1SCORE_TR2	CERAD_WLL2ND_R
WAHR	ADNI	Q1SCORE_TR3	CERAD_WLL3RD_R
WAHR	ADNI	ADAS13	ADAS_COG_13
WAHR	ADNI	PTEDUCAT	EDU_YEARS
WAHR	ADNI	TMT_PtA_Complete	TMT_A_TIME
WAHR	ADNI	TMT_PtA_Comission	TMT_A_E_COMMIT
WAHR	ADNI	TMT_PtA_Omission	TMT_A_E_OMIT
WAHR	ADNI	TMT_PtB_Complete	TMT_B_TIME
WAHR	ADNI	TMT_PtB_Comission	TMT_A_E_COMMIT
WAHR	ADNI	TMT_PtB_Omission	TMT_A_E_OMIT
WAHR	ANM	AGE	AGE_FV
WAHR	ANM	Sex	SEX
WAHR	ANM	MMSE	MMSE_SUM_R
WAHR	ANM	CDR_SOB	CDR_SOB
WAHR	ANM	CDR_Total	CDR_SUM
WAHR	ANM	APOE	APOE_RISK_GROUP
WAHR	ANM	ADAS_COG	ADAS_COG_11
WAHR	ANM	MOCA	MOCA
WAHR	ANM	Geriatric_Depression	GDS_15
WAHR	ANM	NPI	NPI_TOTAL
WAHR	ANM	CERAD_A_Total	VFC_ANIM_R
WAHR	ANM	CERAD_B_Total	BNT_15_SUM
WAHR	ANM	CERAD_C_Total	CERAD_WLLSUM_R
WAHR	ANM	CERAD_D_Total	CERAD_CP_R
WAHR	ANM	CERAD_E_Correct	CERAD_CPR_R
WAHR	ANM	Baseline_Date	DOFV
WAHR	ANM	Diagnosis	DIAG_TXT
WAHR	ANM	CERAD_E_Intrusions	CERAD_WLINT_R

Table 5: Attribute mapping of ANM and ADNI variables. Source attribute is from the curated dataset. Target attribute is defined in the data model. Newly added variables were labelled in grey.

The importance of this mapping procedure could be exemplified by mapping of the Boston Naming Test (BNT), which is one of the most frequently used measures of confrontation naming for cognitive assessment. Originally established by Kaplan et al. in 1983[13], the full BNT consists of 60 drawings of various items that have to be named by the test person. In between, multiple short forms of the BNT (mostly containing 30 or 15 items) have been developed (Mack et al., 1992[14], Calero et al., 2002[15]) to reduce cost, time, but not least stress of elderly test persons. Consequently, the test setups vary, especially considering declining performance of patients in the course of testing time. Generally, the possibility to recalculate the measurements from one assay to the other is at least questionable (Katsumata et al., 2015[16]; Hobson et al., 2011[17]). As part of the standardized CERAD test battery (Morris

[13] Kaplan E, Goodglass H, Weintraub S. The Boston Naming Test. Philadelphia, PA: Lea & Febiger; 1983

[14] Mack W.J. et al., Boston Naming Test: shortened versions for use in Alzheimer's disease. (1992) J Gerontol. 47(3):154

[15] Calero M.D. et al., Usefulness of a 15-item version of the Boston Naming Test in neuropsychological assessment of low-educational elders with dementia. (2002) J Gerontol B Psychol Sci Soc Sci. 57(2):187.

[16] Katsumata Y. et al., Assessing the discriminant ability, reliability, and comparability of multiple short forms of the Boston Naming Test in an Alzheimer's disease center cohort. (2015) Dement Geriatr Cogn Disord. 39(3-4):215.

[17] Hobson V.L. et al., An examination of the Boston Naming Test: calculation of "estimated" 60-item score from 30- and 15-item scores in a cognitively impaired population. (2011) Int J Geriatr Psychiatry. 26(4):351.

et al., 1989[18]) which was a founding element of the initial CCDM, a 15-item BNT was integrated in the basic data model ('BNT_15_SUM') (Table 8). Although ANM and ADNI are performing a BNT, only the variable from ANM ('CERAD_B_Total') could be directly mapped to the CCDM – obviously, the respective values originate from CERAD battery tests, thus equal conditions. In contrast, for the ADNI variable mapping cannot be defined as being equal, as ADNI used a 30-item test ('BNTTOTAL'), resulting in a different variable space. We therefore defined an additional variable ('BNT_30_SUM) for the CCDM. Consequently, the decision whether and exactly how to merge these data is intentionally left to future analysts for dedicated judgement. However, values are thus semantically as well as technically integrated. For the same reason a second ADAS COG assay ('ADAS_COG_13'; value range: [0:85]) has been added to the CCDM, as the already integrated score ('ADAS_COG_11') only considered 11 items (value range: [0:70]). Both the original ADAS11 score as well as the extended ADAS13 score are common in the field. In addition, a thorough evaluation of the clock drawing assay which is used in different forms is also necessary.

The revised and extended data model was added to the CCDM service using the "Datamodel Bulk-Upload" tool in the Data-Steward toolbox. The information on the new incorporated studies, the variables and the relation to existing variables, upper level concepts and other studies' study variable could also be visualized in the Data-Steward visualization tool (Fig. 14 and Fig. 15).

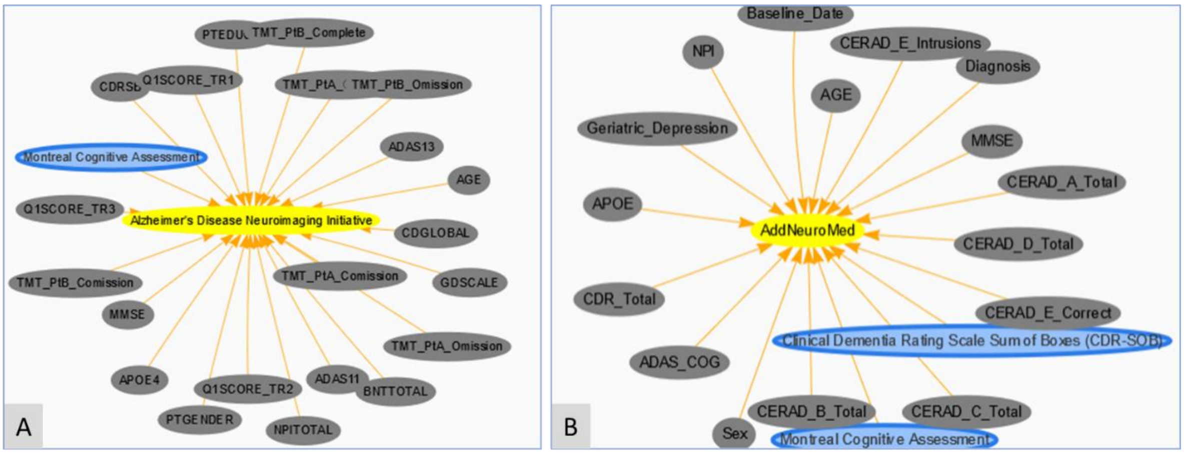


Figure 14: Screenshots from Data-Steward Tool: Features of ADNI (A) and AddNeuroMed study (B)

[18] Morris J.C. et al., The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. (1989) Neurology. 39(9):1159

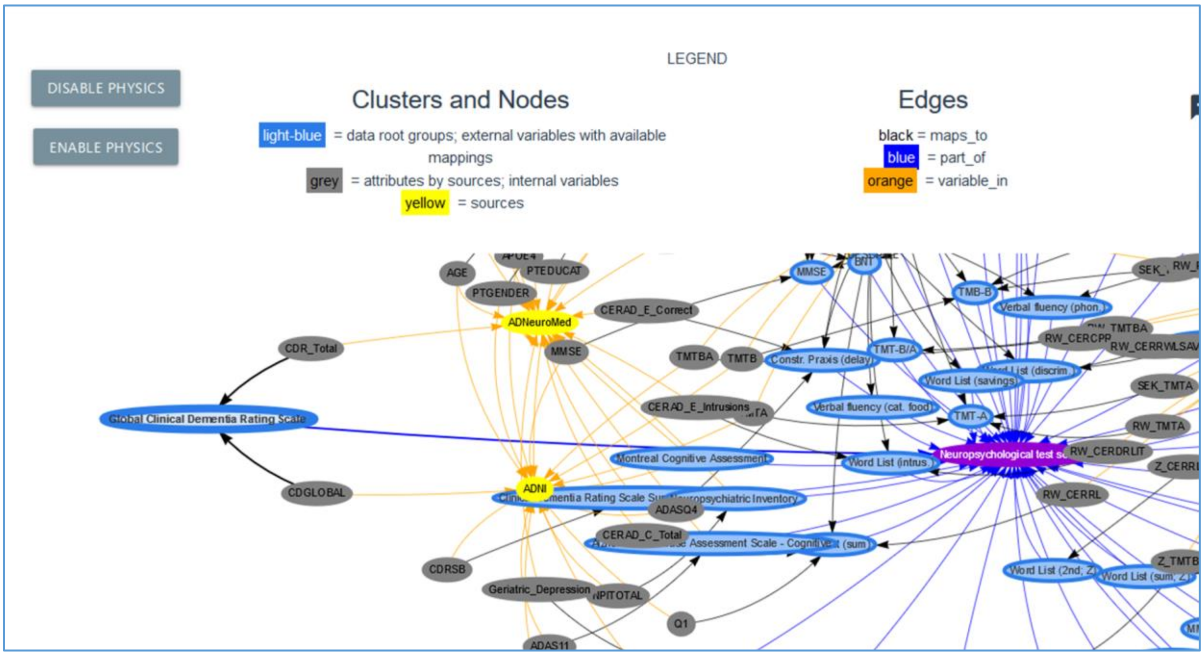


Figure 15: Screenshot from Data-Steward tool: Graph Visualization of the Data Model. The "CDR_Total" variable from the AddNeuroMed study and the CDGLOBAL variable from ADNI study are mapped to CCDM variable "Global Clinical Dementia Rating Scale ('CDR_SUM'), which is part of the "Neurophysiological test scores" in CCDM.

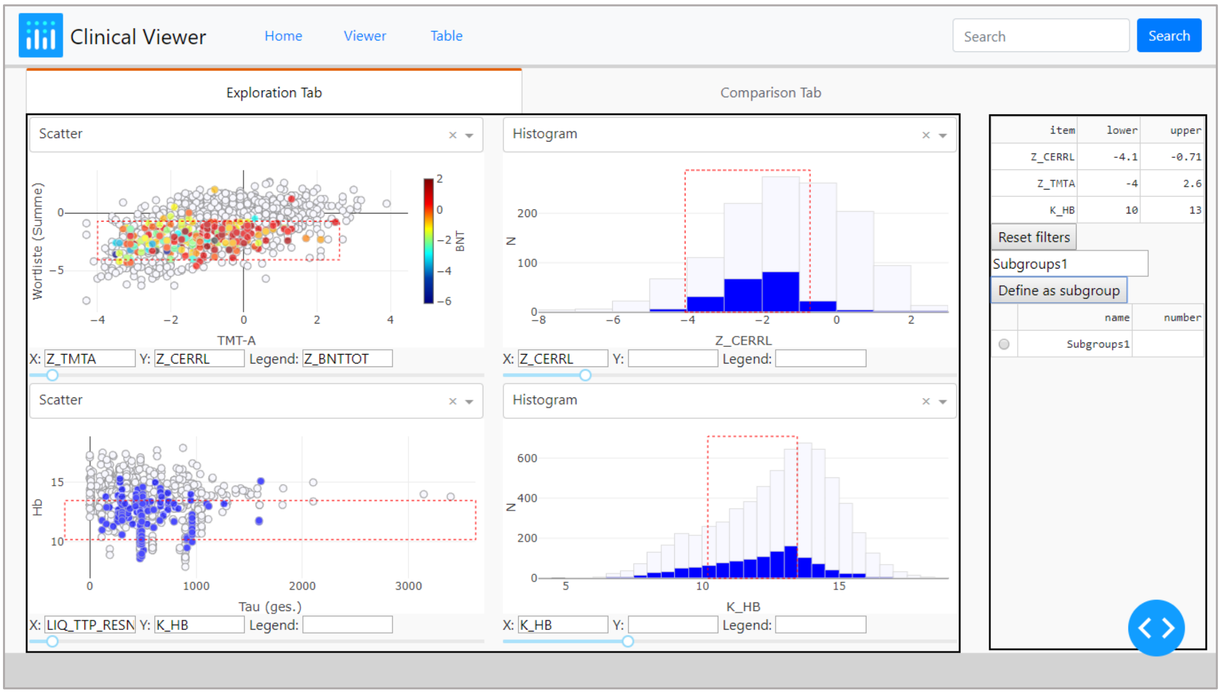
Normalization in the CCDM goes beyond the mapping that was done by the AData Viewer, now comparison of different studies on the patient level is possible. With that we are able to build new "cohorts" over several studies or validate results based on new studies.

Please note: based on data protection related restrictions we were not able to upload and analyze patient level data from ADNI or ANM to the CCDM, yet. Therefore, the results from related data analysis using the clinical data viewer is demonstrated with the data from IDSN-related studies from the ataxia field.

6. Demonstrator: Clinical Data Viewer

The **Clinical Data Viewer** (http://idsn.dzne.de/react-viewer/) is under development in the IDSN project with project partners from the University Hospital Bonn.

The aim is to display semantically integrated data from the clinical data handling backend and allow for user interaction. Thus, it communicates tightly with the respective API, and works on a generalizing concept enabling flexible front end semantics; i.e., the viewer is independent of a certain variable set, but communicates with the backend on available variables (by data type; e.g., maps integer and float to numerics, codes to categorical etc.), sending them to a set of graph elements. Thus, variables are selectable only where applicable, and plots' behavior depends on the CCDM definitions. More important, the viewer transparently generates (clinical) concepts like patients and visits, considers the time dimension (e.g., links data points in X-Y scatterplots according to configurable tolerances in time; "time delta"), enables longitudinal analyses, the immediate definition and comparison of named sub-cohorts from a dataset, as well as the temporary declaration of new variables, e.g. for calculating values (Fig. 16). While data is acquired as already harmonized sets from the backend, meta analyses across studies is an intrinsic feature and just limited to availability and permission (user management).



X: Z_TMTA Y: Z_CERRL Legend: Z_BNTTOT

X: Z_CERRL Y: Legend:

X: LIQ_TTP_RESN Y: K_HB Legend:

X: K_HB Y: Legend:

item	lower	upper
Z_CERRL	-4.1	-0.71
Z_TMTA	-4	2.6
K_HB	10	13

Subgroups1

	name	number
○	Subgroups1	

Figure 16: Screenshot from the Clinical Data Viewer tool (draft version): One possible endpoint for handling, visualizing and analyzing harmonized data from the CCDM-driven clinical data backend. Data is gathered on click, variables assigned to all applicable plots and dimensions. Axis assignments support auto-completion. All displayed data refers to the same data pool ("multi-view"), thus actions like filtering on one graph (or table) influences all others. Filtering conditions and subgroup definitions of patients are located on the right.

7. Conclusion

Within the Virtual Brain Cloud (TVB-Cloud) project, we are building a reference infrastructure for sharing and processing Health and Biomedical research data, specifically data in the domain of neurodegenerative disease (NDD).

One goal was to generate a data catalogue which contain harmonized annotations of all data sets used in this project and can be searched in an integrative viewer.

As datasets from project partners were up to now not available to be used for the establishment of the system, we acquired major AD cohort studies to set up a data catalogue that represent a dataset landscape of AD related studies.

While there has been work on standardizing data collection as well as on guidelines defining an AD specific data model, we still experience a deficit in interoperability across AD datasets. The investigated cohort datasets neither followed a common naming system for variables, nor represented values of the same measurement in equal manner. On top of that, some studies only shared processed values instead of the underlying raw data. This further impedes interoperability since differences in applied processing pipelines inevitably introduce systematic biases into the data.

Irrespective of that, the AData Viewer developed in TVB-Cloud displays a comprehensive AD data landscape and a valuable tool to analyze, map and visualize all kind of modularity measured in AD-related clinical studies. Thereby it also enables comparisons across different studies. The AData Viewer is under further development including the integration of new studies and additional attributes. With that also TVB-Cloud related studies can be integrated, analyzed, and visualized in the AData Viewer based on the described procedure. It should be notified that this variable based integration could be done on the basis of data modalities and does not necessarily need patient derived raw data. It is recommendable that study data owner would be the best qualified experts to do that work.

As pointed out the analysis exposed major challenges that severely impede comparative approaches on the AD cohort data. The finding of common modalities across cohorts does not imply that the measured variables are interoperable or even comparable on a semantic level. By mapping a variety of variables across the datasets, we also established an overview of their interoperability.

One promising approach to increase dataset interoperability is a comprehensive, NDD-specific common data model. Such a data model could support the alignment and mapping of variables by providing easy-to-follow guidelines and a dedicated interface for retrospective data harmonization.

To that end and in order to focus (bundle) research activities and tool development in the wider area of the EU research landscape, we applied the common data model that was under development at SCAI Fraunhofer within the IDSN project for its applicability for AD related datasets using the ADNI and AddNeuroMed datasets that were also analyzed in the AData Viewer.

Focused on the assay included in neurophysiological test batteries (e.g., CERAD) or common variables (e.g., sex, education), we integrated more than 30 different variables into the model. Thereby, we realized that the detailed analysis of study-specific variables on the level of single patient raw data is inevitable, but also cumbersome. When analyzing the subset of variables, it became also clear that this work needs to be done by experienced persons, best from the data owners. Therefore, the CCDM offers a user interface ('Data-Steward') that enables a semi-automatic harmonization of the variables, not at least by using domain-specific ontologies that could be hosted in the Fraunhofer SCAI OLS instance. In addition, the system also enables raw data upload and cross-study analyses using the 'Clinical Data Viewer' or programming environments like Jupyter.

Based on data protection legislation (i.e., GDPR), sharing of patient data might often not be possible. Most components of the system presented here could be executed or implemented in the protected environment of the data owner allowing at least harmonization and comparison against freely accessible data set.

Last, not least the system can provide support within the study planning supporting the identification of alignment parameter.

Overall, the systems presented here were developed/adapted for The TVB-Cloud project and include the following components

- a **data catalogue on AD related studies** with harmonized variables (AData Viewer),
- a **common clinical data model** as a semantic framework that could be used for data harmonization and mapping (Data Steward), and
- finally, for the analysis and comparison of studies the **clinical data viewer**.

8. Supplementary Information

8.1. DKAN Clinical Data Repository

DKAN is a community-driven, free and open-source open data platform (inspired by CKAN (Comprehensive Knowledge Archive Network)), that gives organizations and individuals access to structured information. In our case, the Fraunhofer SCAI DKAN system is an in-house platform enabling the storage of clinical study data in an access-restricted environment in compliance with the FAIR data handling principles 'interoperability' and 're-usability'. Note: This is not the clinical data repository for the TVB-Cloud partners due to license restrictions. Sharing of tools and risk and progression models will be realized via a cloud solution.

The procedural method to import data to the platform includes mainly the following three steps:

Creation of datasets and resources – DKAN's data publishing model is based on the concept of datasets and resources. A dataset is a container for one or more resources; a resource is the actual "data" being published, such as a CSV table, a JSON data file, or a TIFF image. The dataset and resource content types in DKAN are provided by the DKAN Dataset module.

Addition of further resources to the dataset – After creating a dataset, we're prompted to add one or more data resources to it. There are three types of resources that can be added to a Dataset, depending on the type and location of the resource, e.g. upload files, API or website URL (e.g. a database link), and linked remote file.

Adding metadata to datasets – In this last step additional metadata are added to the dataset. All these fields are optional, but provide valuable information about your dataset to both human visitors to the website and machines discovering your dataset through one of DKAN's public APIs. Examples for such metadata information are Author, Spatial/Geographical Coverage Area, Spatial/Geographical Coverage Location, Frequency, Temporal Coverage, Granularity, Data Dictionary, Additional Info, and Resources.

There are further optional modules to enhance DKAN platforms like Datastore, Harvester, Workflow and others. To conclude, DKAN allows to publish data for site visitors handled as content. Treating data like content makes it easier to access and take up information. Equally important is making open data accessible for computers (i.e. machine-readable) to interact programmatically. Both approaches for the data access follow the FAIR paradigm. It needs to be printed out that most of these modules are not applicable as clinical data often underlie data security rules that prohibits the free distribution.

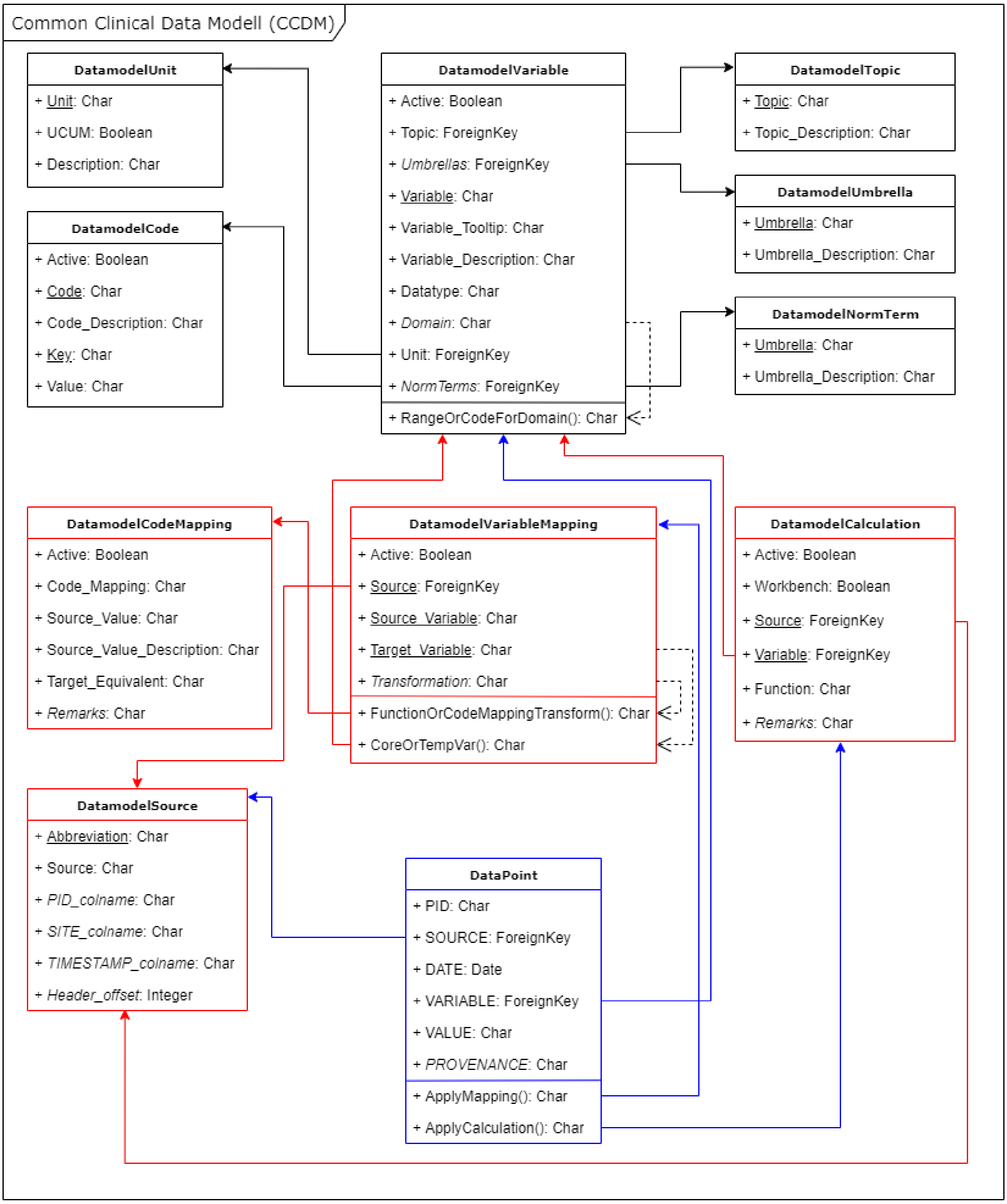
Demographics	Blood Transcriptomics	Genotype	Blood Proteomics	CSF	PET
Not available	Not performed	Not performed	Not performed	Not performed	Not performed
Sex, Age	Only processed data	APOE	Only processed data	At least 1 of: A-beta, pTau, tTau	At least 1 of: FDG PET, Amyloid PET, AV PET
Sex, Age, Education, Race/Ethnicity	Raw data	Broad genotyping	Raw data	A-beta, pTau, tTau	FDG PET, Amyloid PET, AV PET

Lifestyle	Family History	Comorbidities	Autopsy	Medication
Not available	Not available	Not available	Not performed	Not available
At least 1 of: Substance consumption; physical activity; nutrition	Dementia state of biological parents	Information on at least 1 comorbidity	Basic information (e.g. brain weight)	AD medication
Information on at least 2 of the above	Additional Information	Information on 5 ≤ other diseases	At least 1 post-mortem omics dataset	Detailed medication

Supplementary Table 1: Criteria used to define the availability scores for each investigated data modality.

8.2. Description of the Common Clinical Data Model (CCDM)

The CCDM reflects by its sub-structure the separation into a core set of internal variables, a collection of mappings dedicated to particular external data resources as well as the data itself (suppl. Figure 1) These three levels refer to each other and are described in the following sections.



Supplementary Figure 1: UML diagram of the CCDM. The data model core (black) consists of variables, defined by several attributes. Thereof, topics and umbrella terms as semantic entities are collected separately and referenced. Units (for numeric variables) and codes (for categorical variables) are similarly listed. For semantic integration of data, the main collection of variable mappings (red) is analogously supported by referenceable code mappings and always refer to dedicated sources. Data points (blue) being subject to semantic integration refer to both core model (variables) and mappings, the latter including post-processing of loaded data according to defined calculations to be carried out

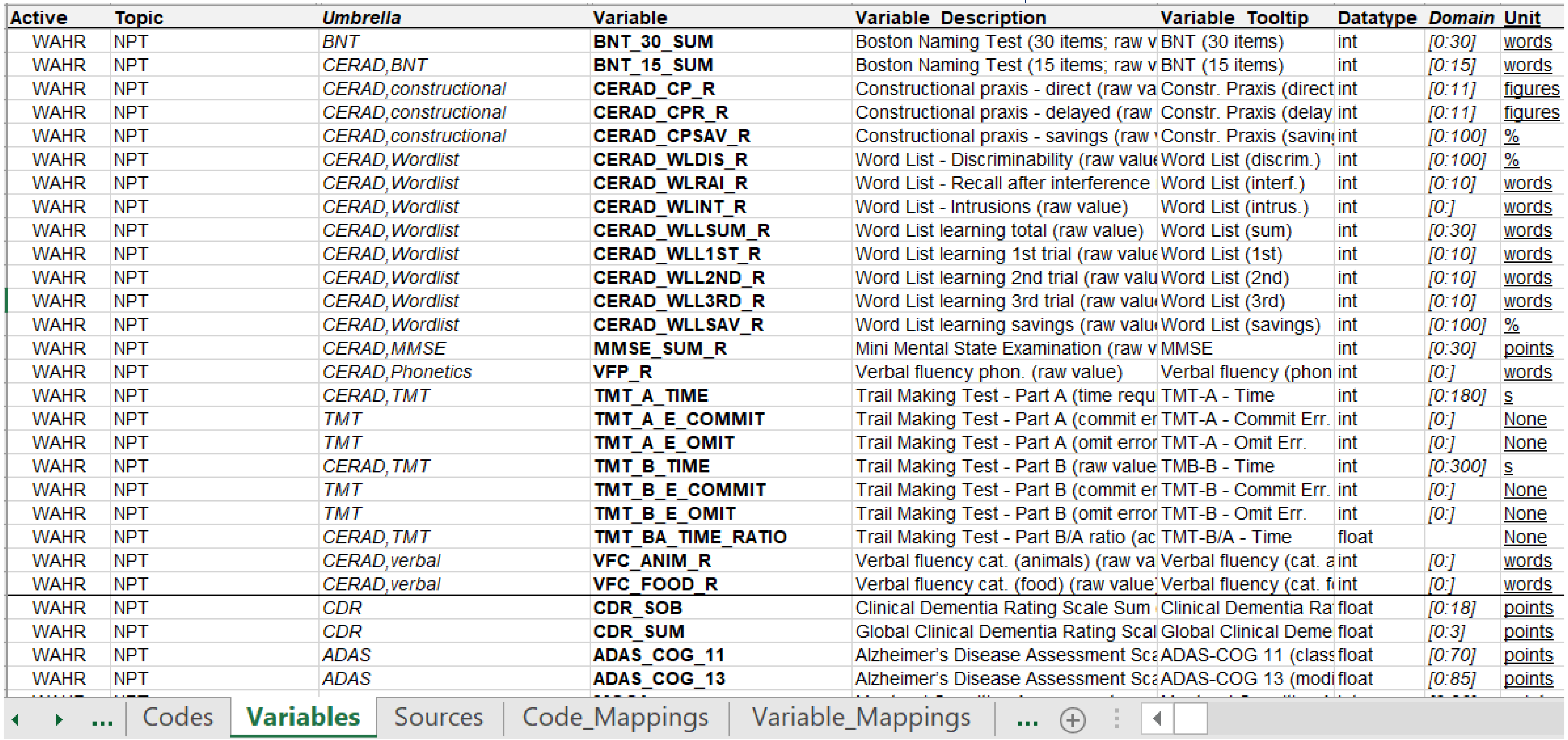
CCDM Core

The data model comprises a set of unequivocally/unambiguously named variables (attributes), carrying several properties (suppl. Tab. 2 and suppl. Fig. 2). Each variable comes with proper textual descriptions and a definition of the data type. Supported types are either textual (strings), numerical (integer, float), categorical (codes; incl. Boolean) or dates (ISO8601-formatted). All data types can be declared as homogenous arrays, so lists of e.g., integer values, dates or code.

For numerical variables, default units can be defined. Although these are preferably consistent with the Unified Code for Units of Measure (UCUM), the 'unit' feature accepts custom entries, e.g. 'points', 'figures' or 'words'. Thereby, modeling of common clinical measures like specifically counted items in a neuropsychological test setup is possible. The 'domain' feature optionally defines legal value ranges, either in a [min]:[max] schema or as comma-separated lists. For variables of the 'code' data type, 'domain' is a reference to a named collection of key-value pairs and obviously mandatory.

Property	Function	Mandatory
Variable	Unique name (ID)	✓
Topic	Unique parent category	
Umbrella	Generalization term wrapping a variable with further, analogous terms (= semantically similar, but not identical).	
Variable_Description	Unique variable's full text description	✓
Variable_Tooltip	Very short description, suitable for tooltips or table headings	✓
Datatype	Possible data type [string, int, float, date, code or arrays of]	✓
Domain	Accepted range of numeric and date values (2-items array in Python notation; empty item is valid) or name of code (for categorical items)	(✓)
Unit	Unique code expressing unit (UCUM preferred)	
Active	Switch for (de-)activation	✓

Supplementary Table 2: Data model fields defining variables



Active	Topic	*Umbrella*	Variable	Variable_Description	Variable_Tooltip	Datatype	*Domain*	Unit
WAHR	NPT	*BNT*	**BNT_30_SUM**	Boston Naming Test (30 items; raw v	BNT (30 items)	int	*[0:30]*	words
WAHR	NPT	*CERAD,BNT*	**BNT_15_SUM**	Boston Naming Test (15 items; raw v	BNT (15 items)	int	*[0:15]*	words
WAHR	NPT	*CERAD,constructional*	**CERAD_CP_R**	Constructional praxis - direct (raw va	Constr. Praxis (direct	int	*[0:11]*	figures
WAHR	NPT	*CERAD,constructional*	**CERAD_CPR_R**	Constructional praxis - delayed (raw	Constr. Praxis (delay	int	*[0:11]*	figures
WAHR	NPT	*CERAD,constructional*	**CERAD_CPSAV_R**	Constructional praxis - savings (raw v	Constr. Praxis (savin	int	*[0:100]*	%
WAHR	NPT	*CERAD,Wordlist*	**CERAD_WLDIS_R**	Word List - Discriminability (raw valu	Word List (discrim.)	int	*[0:100]*	%
WAHR	NPT	*CERAD,Wordlist*	**CERAD_WLRAI_R**	Word List - Recall after interference	Word List (interf.)	int	*[0:10]*	words
WAHR	NPT	*CERAD,Wordlist*	**CERAD_WLINT_R**	Word List - Intrusions (raw value)	Word List (intrus.)	int	*[0:]*	words
WAHR	NPT	*CERAD,Wordlist*	**CERAD_WLLSUM_R**	Word List learning total (raw value)	Word List (sum)	int	*[0:30]*	words
WAHR	NPT	*CERAD,Wordlist*	**CERAD_WLL1ST_R**	Word List learning 1st trial (raw value	Word List (1st)	int	*[0:10]*	words
WAHR	NPT	*CERAD,Wordlist*	**CERAD_WLL2ND_R**	Word List learning 2nd trial (raw valu	Word List (2nd)	int	*[0:10]*	words
WAHR	NPT	*CERAD,Wordlist*	**CERAD_WLL3RD_R**	Word List learning 3rd trial (raw valu	Word List (3rd)	int	*[0:10]*	words
WAHR	NPT	*CERAD,Wordlist*	**CERAD_WLLSAV_R**	Word List learning savings (raw valu	Word List (savings)	int	*[0:100]*	%
WAHR	NPT	*CERAD,MMSE*	**MMSE_SUM_R**	Mini Mental State Examination (raw v	MMSE	int	*[0:30]*	points
WAHR	NPT	*CERAD,Phonetics*	**VFP_R**	Verbal fluency phon. (raw value)	Verbal fluency (phon	int	*[0:]*	words
WAHR	NPT	*CERAD,TMT*	**TMT_A_TIME**	Trail Making Test - Part A (time requ	TMT-A - Time	int	*[0:180]*	s
WAHR	NPT	*TMT*	**TMT_A_E_COMMIT**	Trail Making Test - Part A (commit er	TMT-A - Commit Err.	int	*[0:]*	None
WAHR	NPT	*TMT*	**TMT_A_E_OMIT**	Trail Making Test - Part A (omit error	TMT-A - Omit Err.	int	*[0:]*	None
WAHR	NPT	*CERAD,TMT*	**TMT_B_TIME**	Trail Making Test - Part B (raw value	TMB-B - Time	int	*[0:300]*	s
WAHR	NPT	*TMT*	**TMT_B_E_COMMIT**	Trail Making Test - Part B (commit er	TMT-B - Commit Err.	int	*[0:]*	None
WAHR	NPT	*TMT*	**TMT_B_E_OMIT**	Trail Making Test - Part B (omit error	TMT-B - Omit Err.	int	*[0:]*	None
WAHR	NPT	*CERAD,TMT*	**TMT_BA_TIME_RATIO**	Trail Making Test - Part B/A ratio (ac	TMT-B/A - Time	float		None
WAHR	NPT	*CERAD,verbal*	**VFC_ANIM_R**	Verbal fluency cat. (animals) (raw va	Verbal fluency (cat. a	int	*[0:]*	words
WAHR	NPT	*CERAD,verbal*	**VFC_FOOD_R**	Verbal fluency cat. (food) (raw value	Verbal fluency (cat. f	int	*[0:]*	words
WAHR	NPT	*CDR*	**CDR_SOB**	Clinical Dementia Rating Scale Sum	Clinical Dementia Ra	float	*[0:18]*	points
WAHR	NPT	*CDR*	**CDR_SUM**	Global Clinical Dementia Rating Scal	Global Clinical Deme	float	*[0:3]*	points
WAHR	NPT	*ADAS*	**ADAS_COG_11**	Alzheimer's Disease Assessment Sc	ADAS-COG 11 (class	float	*[0:70]*	points
WAHR	NPT	*ADAS*	**ADAS_COG_13**	Alzheimer's Disease Assessment Sc	ADAS-COG 13 (modi	float	*[0:85]*	points

Supplementary Figure 2: Variable definition in the data model template

Two **generalization concepts** conduct further semantics. Topics are hierarchically ordered, assigning variables to fields of interest. While various lab chemistry measurements are generated from blood samples, others originate from cerebrospinal fluid (CSF) - an important differentiation in neurodegenerative research, modeled by separate topics, which anyhow can be generalized to 'lab measurements'. Meanwhile, defining umbrella terms enables to model similarity between variables, independent of the rigid topics hierarchy. For instance, our model knows a range of variables for the reporting of memory disturbances. According to study designs, these might carry additional information on the reporter; consequently, three variables indicate whether the patient, an attendant (usually an accompanying relative) or an undesignated person made the qualifying statement on experienced memory issues. However, in integrated analysis of data, a researcher is probably primarily interested in any reported occasion, which is modeled by the umbrella term 'memory disturbance'. Similarly, leukocytes are counted in both blood and CSF; while the respective variables are assigned to separate topics, the umbrella term "leukocytes" serves as a direct link. In fact, a variable can be assigned with multiple umbrella terms, enabling various independent, even overlapping generalizations.

For the sake of **normalization**, the CCDM can be mapped to public ontologies. Here, data model variables are linked to defined terms of referenceable semantic resources. Where necessary, exact partial definitions are possible, e.g. assigning encoded variables' keys to particular terms. For instance, while the CCDM's variable 'sex' knows the values '0' (= 'male'), '1' (= 'female') and '2' (= 'other'), these are encoded as separate terms in NCIT: C20197 (male) and C16576 (female). Moreover, C154419 (genderqueer), C154420 (other gender) and C154421 (transgender) fit 'other'. For the mapping of NDD specific variables specific ontologies such as CTO that are hosted in the Fraunhofer OLS will be used (see our published report on "Complete, updated semantic framework for neurodegeneration research documented"[19]).

Mappings

The central mechanism for formally binding external data sources to the CCDM's variable space is the mapping process. Here, declarations are set up to formally describe pair-wise relations between (internal) data model variables and those of the (external) resources. Transformations enable to align external variables with deviating value ranges or default units. These might either be a conversion expressed as a function call (e.g., 'FORMULA((EXTVAR – 100) / 10)') or a code mapping identifier. The latter reference a set of value pairs, linking external values explicitly to those of a categorical variable of the CCDM.

Data

The core principle of the CCDM is to assign every particular measurement to a) an individual ('patient') and b) a point in time. Effectively, the data points to be integrated are quintuples of character-separated values according to suppl. Tab. 3, providing one record per row. Thereby, data points follow the principles of the EAV/CR model (Nadkarni et al. 1999[20]): while the person and time information assemble an entity (E), measurements compile from the recorded value (V) and the variable assigned to – the latter referred to as attribute (A) in EAV. Variables as metadata are defined separately and assigned

[19] https://cordis.europa.eu/project/id/826421/results

[20] Nadkarni P.M. Organization of heterogeneous scientific data using the EAV/CR representation. (1999) J Am Med Inform Assoc. 6(6):478

(related) to topics (classes). In contrast to the EAV/CR model, the CCDM's topics are less complex than classes, with less relations in parallel.

Field *(Examples)*	Content
[SITE:]PID *(128b, LocX:12345)*	Person's unique ID in the given context; might be prefixed by a location descriptor SITE, e.g., for multi-center studies
TIMESTAMP *(2012-08-19, 1971, 2016-04, 2018-06-12T10:09:13)*	Time point of the measurement's recording in ISO 8601 format ([YYYY]-[MM]-[DD]T[hh][mm][ss]); the minimum representation is YYYY
VARIABLE *(DIAG_ICD10, MMSE_SUM, HAS_APOE4)*	The name of the variable, either out of the CCDM's internal namespace *or* the defined mapping set for a given data source (external namespace)
VALUE *(0, 23, F03, 12.7g)*	The actually measured value, matching the referred VARIABLE's data type and domain. Numeric values are allowed to be trailed by an alternative unit, if it can be converted accordingly (e.g., µg -> mg)
PROVENANCE *(studyserver:/nfs/data/export/v2.eav, 1234fsdjklfn55, Sally's complete collection)*	Optional text field for providing information on the particular record's origin; might be URLs, notes, encoded paths, hashes or any other information potentially helpful for re-locating data in the source itself if ever necessary.

Supplementary Table 3: Data point elements matching the model's core principle of relating a measurement (VARIABLE + VALUE) to an entity providing both relations to the individual (PID) and the time of recording (TIMESTAMP).

Accordingly, data from external resources finally requires a conversion to the EAV-like format at the entry point to the CCDM in order to get handled. Handling meanwhile comprises a full check whether a particular data point matches the modeled metadata conditions. If so, data points ultimately get stored in an equally formatted data table.

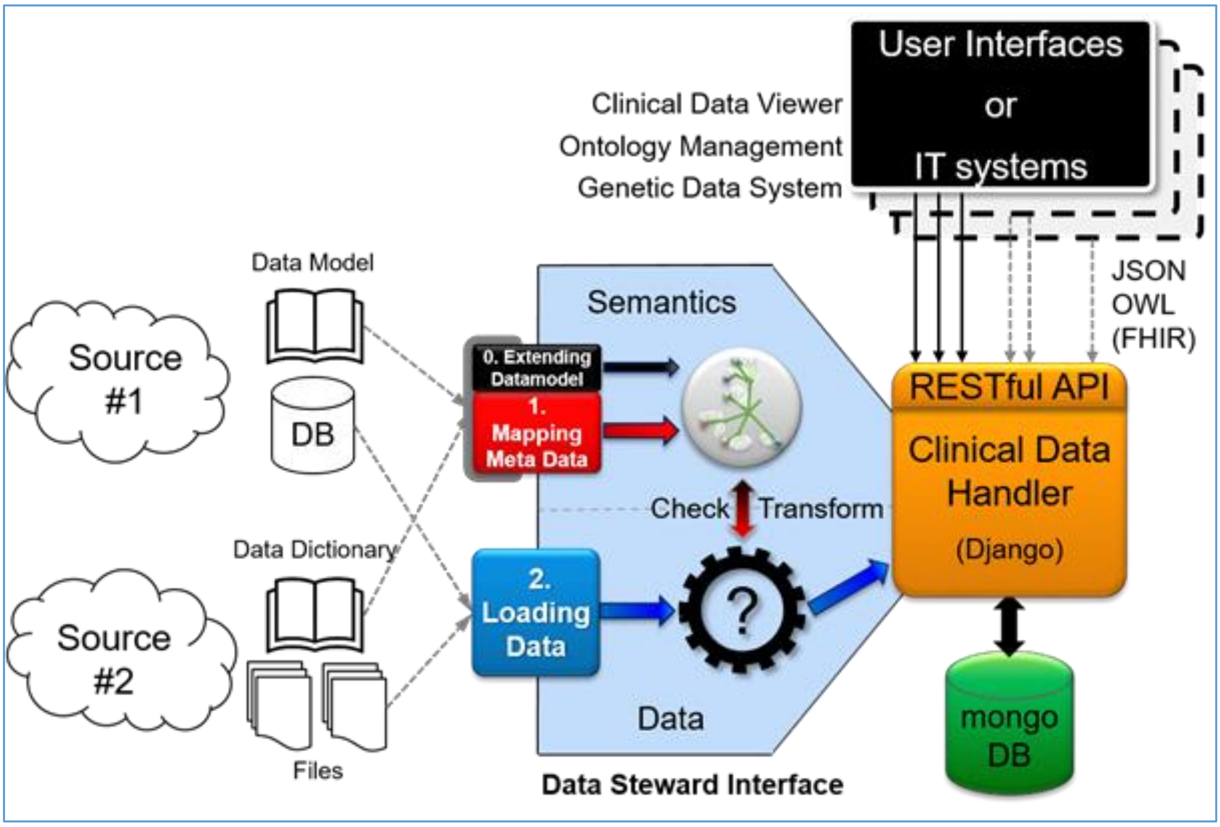
Implementation

In order to make practical use of the common clinical data model, we developed a range of software solutions. These together form a flexible, multi-purpose semantic integration system to pool both data and metadata. Here, the CCDM is implemented as Django models, complemented with dedicated processes for importing, transforming and exporting information from external resources. While the model is fully computable, data's compliance is asserted on the fly at the time of integration into the system. Considering semantic queries on thereby clean and harmonized data, such a system supports interoperability by centrally providing normalizations to public ontologies. Consequently, we offer common formats for exchange of (meta)data with connected systems.

Essentially, our software layer comprises four major elements. The actual backend for storing data, data model and information on mapping sources' metadata is Django-driven, based on MongoDB and access-controlled by KeyCloak. Two importers dedicated to the model/mappings and data, respectively, are processes ensuring the integration of either definitions or data from previously declared sources. For both, we provide web interfaces, primarily supporting data stewards. An API enables clients to query both clean data and the model itself. Here, data property is respected; users will only receive data points upon permission. Finally, the system is capable of providing data in various export formats; these comprise simple tabular, but also JSON-formatted files. While the latter is easily extendable to hospital-oriented HL7-FHIR, exporting to CDISC could connect the environment to study-related IT systems. Both carry data and metadata in an integrative way, allowing for intrinsically semantic data. The CCDM itself including mappings from data resources and to public ontologies can be dumped to Web Ontology Language (OWL).

Integration processes

The integration of new sources of clinical data into our system requires two subsequent multi-step processes, preferably executed by a data steward. An overview is provided suppl. Fig. 3.



Supplementary Figure 3: Overview on the data integration process. Integration of data from various sources starts with mapping the given meta data onto the common clinical data model. Resulting assignments and rules get implemented automatically in the central business intelligence. Similarly, data will be loaded into this central, database-driven environment; here, tabular data will be parsed, filtered and quality controlled according to the set-up source-specific declarations encoded in data model and mappings. Both integration processes are supported by a web interface, implemented as a form-driven file upload each

Modeling & Mapping

In the first process of integrating metadata, the relevant parts of the external data space get semantically connected to the CCDM via pair-wise mappings between external variables and those of the CCDM core. declarations on the expected input data from the given study or dataset ('resource') have to be made. Rarely, the complete data space is subject to integration, as usually an excessive number of variables of partially quite special interest is recorded. In fact, there will be a core set of variables of common interest. Essentially, a respective list of identifiers should be prepared prior to the mapping process. The resource to be referenced for any set up mapping might be declared first. In the actual mapping process, optional transformations can be defined. These might be formulas from a small set of mathematical operations, or particular mappings of expected source values to legal target values. A simple example is the conversion from named months to their numerical equivalents ('Jan' → '1'). These assignments are declared separately as so-called 'code mappings' and are uniquely named. As code mappings are not restricted to data resources, they could be reused.

If an external variable of interest does not match a core set variable even using transformations, the CCDM has to be extended with an adequately defined variable. Here, careful considerations on a generalization of the particular bit of information have to be made. As the data model is intended to be common, it might be favorable to define a widely applicable variable (by name, data type, values, range or arbitrary combinations of those) and introduce a sophisticated mapping for the given source.

Similarly, the assignments of hierarchy (topics) and similarities (umbrella terms) might either refer to elements of the CCDM's given pool or created newly. While both know identifier and description, the former allows the definition of more generalized 'parent topics'.

Finally, in certain complex settings the model variables might depend on multiple inputs. For instance, if not provided explicitly, an age of onset of a disease is necessary to calculate from the time of measurement (e.g. the clinical diagnosis) and the date of birth of an individual. Obviously, this calculation can be performed earliest when all necessary inputs have been read. Separate 'calculation' definitions serve as rule sets being executed in a post-processing procedure. Similar to the mapping definitions with transformations, functions from a provided set of operators allow relating multiple inputs, assigning the result to a variable. Several of those functions know a parameter to express a temporal tolerance ('time delta'); if measurements have not been recorded in a certain period of time, they will not be set off against each other.

All declarations on the CCDM's core and mapping sets can be generated from either an interactive web interface or a mechanism for uploading bulk tables. While the former provides a fine-grained, guided user dialog and a number of support mechanisms (e.g. auto-complete), the CCDM's table representation can be handled offline, generated programmatically and, not least, provides a comprehensive overview on the entire model. However, both options report issues on syntax, semantics as well as, to a certain degree, plausibility.

Data integration

The second, similar process of integrating actual data requires valid, source-specific declarations as indicated above. The reason is that input data is atomically checked for compliance, and all undeclared variables or measurements get ignored. Accordingly, all data finally integrated is guaranteed to comply with the declared specifications of the respective source, formalized in the data model core and mapping tables.

From the data steward perspective, data has to be uploaded with the reference to a declared source. Regarding the input format, either in the quite commonly used 2D CSV table (row = patient at date; column = variable measured) or an EAV-like file has to be provided. While the one-dimensional format natively matches the CCDM’s data point class, CSV files can be converted, if all information is present.

In terms of processes and interfaces we provide a file upload web dialog as well as an API. Depending on the user-selected mode, data is persistently written to the database by single approved data points or completely model-compliant dataset only; also, a sanitizing option is provided, e.g., for correcting location-specific notations of decimal points or dates.

9. Glossary

AD – Alzheimer’s Disease

AData Viewer – Alzheimer’s Disease related Data Viewer

ADNI – Alzheimer’s Disease Neuroimaging Initiative

AIBL – Australian Imaging Biomarkers and Lifestyle Study of Aging

ANM – AddNeuroMed study

ANMerge – extension of ANM study

API – Application Programming Interface

BNT - Boston Naming Test

CCDM - Common Clinical Data Model

CDISC – Clinical Data Interschange Standards Consortium

CDR – Clinical dementia rating

CDRSB - CDR sum of boxes scores

CERAD – Consortium to establish a registry for Alzheimer’s Disease

CKAN - Comprehensive Knowledge Archive Network

CSF – Cerebrospinal fluid

CTO – Clinical Trials Ontology

CSV – Comma-separated values

Django - Python-based free and open-source web framework that follows the model-template-views architectural pattern.

DKAN - DKAN Clinical Data Repository

EAV/CR model – Entity-attribute-value model

ECDF - empirical cumulative distribution function

EPAD - European Prevention of Alzheimer's Dementia

FAIR – Findable, Accessible, Interoperable, Reproducible

DZNE – Deutsches Zentrum für Neurodegenerative Erkrankungen e. V.

GUI – Graphical User Interface

GDPR – General Data Protection Regulation

HBP – Human Brain Project

HL7-FHIR – Level 7 International Fast Healthcare Interoperability Resources

ID – Identifier

IDSN – Integrative Data Semantics for Neurodegenerative research

JSON – JavaScript Object Notation

MCI - Mild Cognitive Impairment

MMSE - Mini-Mental State Examination

MRI - Magnetic resonance imaging

MCI – Mild Cognitive Impairment

NDD – Neurodegenerative disease

NACC - National Alzheimer's Coordinating Center Study

NCIT – National Cancer Institute Thesaurus

OWL – Web Ontology Language

PET - Positron emission tomography

PID – Process identifier

TIFF - Tag Image File Format

OLS – Ontology Lookup Service

TVB-Cloud – The Virtual Brain Cloud

UCUM - Unified Code for Units of Measure

UML – Unified Modeling Language