Methods of Information in Medicine

The Leipzig Health Atlas - An open platform to present, archive and share biomedical data, analyses and models online


Affiliations below.

DOI: 10.1055/a-1914-1985

Please cite this article as: Kirsten T, Meineke F, Löffler-Wirth H et al. The Leipzig Health Atlas - An open platform to present, archive and share biomedical data, analyses and models online. Methods of Information in Medicine 2022. doi: 10.1055/a-1914-1985

Conflict of Interest: The authors declare that they have no conflict of interest.

This study was supported by German Ministry of Education and Research, 031L0026, Innovative Medicines Initiative 2 Joint Undertaking, 853988

Abstract:
Background: Clinical trials, epidemiological studies, clinical registries and other prospective research projects, together with patient care services, are main sources of data in the medical research domain. They serve often as basis for secondary research in evidence-based medicine, prediction models for disease and its progression. This data is often neither sufficiently described nor accessible. Related models are often not accessible as functional program tool for interested users from the healthcare and biomedical domains.

Objective: The interdisciplinary project Leipzig Health Atlas (LHA) has been developed to close this gap. LHA is an online platform that serves as a sustainable archive providing medical data, metadata, models, and novel phenotypes from clinical trials, epidemiological studies and other medical research projects.

Methods: Data, models, and phenotypes are described by semantically rich metadata. The platform prefers to share data and models presented in original publications but is also open for non-published data. LHA provides and associates unique permanent identifiers for each data set and model. Hence, the platform can be used to share prepared, quality assured data sets and models while they are referenced in publications. All managed data, models, and phenotypes in LHA follow the FAIR principles, with public availability or restricted access for specific user groups.

Results: The LHA platform is in productive mode (https://www.health-atlas.de/). It is already used by a variety of clinical trial and research groups and is becoming increasingly popular also in the biomedical community. LHA is an integral part of the forthcoming initiative building a national research data infrastructure for health in Germany.

Corresponding Author:
Toralf Kirsten, Leipzig University Medical Center, Medical Data Science, Leipzig, Germany, tkirsten@uni-leipzig.de

Affiliations:
Toralf Kirsten, Leipzig University Medical Center, Medical Data Science, Leipzig, Germany
Toralf Kirsten, Leipzig University, Institute for Medical Informatics, Statistics and Epidemiology, Leipzig, Germany
Toralf Kirsten, Leipzig University, LIFE Research Centre for Civilization Diseases, Leipzig, Germany
Markus Löfler, Leipzig University, Interdisciplinary Centre for Bioinformatics, Leipzig, Germany

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
The Leipzig Health Atlas - An Open Platform to Present, Archive and Share Bio-medical Data, Analyses and Models Online

Toralf Kirsten, Frank A. Meineke, Henry Loeffler-Wirth, Christoph Beger, Alexandr Uciteli, Sebastian Stäubert, Matthias Löbe, René Hänsel, Franziska G. Rauscher, Judith Schuster, Thomas Peschel, Heinrich Herre, Jonas Wagner, Silke Zachariae, Christoph Engel, Markus Scholz, Erhard Rahm, Hans Binder, Markus Loeffler on behalf of the LHA team

*Authors contributed equally
§ shared senior authorship
# further LHA team members in alphabetical order: Anika Groß, Ying-Chi Lin, Katja Rillich, Samira Zeynalova, Marita Ziepert

Abstract / Summary

Background Clinical trials, epidemiological studies, clinical registries and other prospective research projects, together with patient care services, are main sources of data in the medical research domain. They serve often as basis for secondary research in evidence-based
medicine, prediction models for disease and its progression. This data is often neither sufficiently described nor accessible. Related models are often not accessible as functional program tool for interested users from the healthcare and biomedical domains.

Objective The interdisciplinary project Leipzig Health Atlas (LHA) has been developed to close this gap. LHA is an online platform that serves as a sustainable archive providing medical data, metadata, models, and novel phenotypes from clinical trials, epidemiological studies and other medical research projects.

Methods Data, models, and phenotypes are described by semantically rich metadata. The platform prefers to share data and models presented in original publications but is also open for non-published data. LHA provides and associates unique permanent identifiers for each data set and model. Hence, the platform can be used to share prepared, quality assured data sets and models while they are referenced in publications. All managed data, models, and phenotypes in LHA follow the FAIR principles, with public availability or restricted access for specific user groups.

Results The LHA platform is in productive mode (https://www.health-atlas.de/). It is already used by a variety of clinical trial and research groups and is becoming increasingly popular also in the biomedical community. LHA is an integral part of the forthcoming initiative building a national research data infrastructure for health in Germany.

Keywords
FAIR
data semantics
research data management
structured medical data / metadata
clinical trial data
biomathematical modelling
ontologies
1 Introduction

Scientific results are typically reported in (peer-reviewed) publications. In the medical research domain, clinical trials, epidemiological and molecular biological studies are the main sources for creating data for evidence-based medicine finally aiming at improving healthcare. While publications describing the essence of the scientific finding are accessible, the publication data and derived models are often not available for interested readers which hampers findability, accessibility, and further interpretability of the data, by limiting their re-usability.

In recent years, the publication requirements in many medical journals, among them the Nature publishing group, Cell Science, PLOS journals have changed. Typically, these journals require the availability of fundamental data, not just in order to check the validity of generated scientific results but also to guarantee their sustainability for future research. Multiple platforms became available allowing managing scientific data. Dryad \(^1\) and Gene Expression Omnibus (GEO) \(^2,3\) are prominent examples of such platforms used widely in the last decade.

Most of these platforms offer a storage service, which is, in some cases, not free of charge. Data can be uploaded in nearly every format. Comma separated value structures (CSV) allow nearly every tabular format with and without column headers. However, some platforms are lacking additional metadata, which often makes data usage difficult. Moreover, data is often not interoperable because common data formats are not used. Shared medical data should include metadata describing the clinical, technical and semantical context. Because of the semantic heterogeneity of the used terminologies data from different publications (uploads) usually cannot be combined without additional efforts. Moreover, clinical projects often use different consent forms in order to capture the permission from participants giving rise to
variable consent limitations. In the same way, privacy regulations need to be considered to be in line with laws, like the General Data Protection Regulation (GDPR) and country-specific laws. All these technical and organizational requirements have been bundled with the FAIR principles for making research data Findable, Accessible, Interoperable, and Reusable, that came up in parallel with the project start of LHA in 2016. Numerous articles have discussed FAIR practices and principles (and references cited therein).

To address these challenges, we have set up the Leipzig Health Atlas (LHA) as a web-based platform to make data, analyses and models accessible for interested research communities in a FAIR-conform way. The goal is to provide data of different nature together with their metadata on different levels of abstraction, cf. a classification of data - from an ontological viewpoint. Moreover, and in addition to many other data platforms, the LHA also aims to implement different kinds of biomathematical models and data analysis tools as interactive web-applications.

By now (February 2022), the LHA contains 327 data sets and 34 models / tools associated with 891 publications and 34 scientific projects, mainly from the domains of medical research. The paper is organized as follows: We introduce the main concepts and methods (Section 2), provide an overview of available models and applications (Section 3), and finally discuss special aspects in comparison with similar platforms (Section 4).

2 Methods

2.1 Sharing Approach

The functional structure of the LHA is based on concepts of the reference model for an Open Archival Information System (OAIS). This model provides a framework, including terminology and concepts, for describing and comparing architectures and operations of archives and, thus, for sharing their content. Figure 1 shows an overview of functional entities implemented in the Leipzig Health Atlas. Clinical and epidemiological study groups and consortia provide data and describing metadata on different levels to the LHA in a
publication-oriented manner. Each producer of content (left side in Figure 1) can create new projects and publication entries which can be associated with data sets. The upload process is available for registered LHA users. Each entry is associated with a unique identifier and a stable URI allowing identifying and accessing projects and publications as well as data sets within and from outside the LHA. The LHA is also suited for pure data publications. Hence, the generated identifier can be used in the corresponding publication and also for the journal review.

The LHA platform provides basic rules for medical data management each user should follow. According to our terms and conditions, every user must provide data without patient identifications. It stays in the user responsibility to replace identification data by study-specific pseudonyms for each individual beforehand. This applies to structured data, images, and other types of data as well. Secondly, the LHA does not manage patient consents, neither in printed form nor electronically; this must be managed on the user side. Thirdly, the LHA prefers data files following current interoperability standards, such as CDISC Operational Data Model (CDISC ODM) \(^9\). We use these formats throughout the platform for sharing with interested researchers but also to generate additional data formats to feed specific software applications like SPSS and R.

In addition to the scientific data, the LHA offers tools for a spectrum of models that have been created and evaluated in medical research. Such models include risk and prognosis analysis for different diseases such as hereditary cancer and lymphoma and also normative epidemiological data for different phenotypes. According to the Open Archival Information System (OAIS) standard, data sets and applications (as implementation for different kinds of models) can be accessed and used by users of the LHA platform (right side of Fig. 1). The LHA uses different access levels as described below. Although OAIS is a high-level reference model including neither implementation details nor tool recommendations, it is helpful in structuring the functional entities such as Ingest or Access components \(^8\) within the LHA. We
experienced this benefit also in other projects such as at the LIFE Research Centre for Civilization Diseases (University Leipzig) when data from multiple large epidemiological studies needed to be integrated for analysis-focused sharing on request \textsuperscript{10}. The same methodology is used in the German Medical Informatics Initiative aiming at integrating and preparing patient care data in so-called data integration centers at University Hospitals in Germany \textsuperscript{11}.

2.2 Data Access and Reuse

Data is represented in two possible ways by the LHA platform: Firstly, users can upload data as data files, preferably in CDISC ODM format. We currently offer the service generating this standardized format from tabular data files along with separately collected, standardized metadata descriptions. By uploading a data file, the user creates an entry in the LHA platform that can be associated with other entries. In this way, data files are linked with studies in which they have been generated, with the respective publication(s) as well as with the uploading user who is registered within the platform. These different references interrelate between different types of entries. They form a network allowing the user to navigate throughout the LHA platform and, thus, to access entry by entry, i.e., from a single research project to relevant publications as its outcome to their authors. In this way, the user can access all entries where each entry is publicly available. The visibility of the whole content of an entry depends on its access level and users access rights. For instance, a user who is interested in the LIFE-Adult Study\textsuperscript{12} can search the corresponding entry and see all relevant metadata including the study leaders (links to other entries), descriptions providing an overview of the goals and associated publications (links to publication entries). Each publication can refer to publication data (entries). While the user can navigate to a single publication data entry in the platform, the corresponding data file is only visible if its access level has been defined as “public” during the upload process. The access level can be changed at any time by a user with access rights. There are further access levels provided by the LHA platform. Data files
can be accessible for a user group which is defined within the system. In this way, the data file will become visible and, thus, accessible when a user authenticates (logins) at the platform and the user belongs to the predefined user group. Moreover, an authenticated user can request access to a specific data file. In this case, the platform sends automatically an e-mail to the uploading content provider who then can provide access rights or reject the request. The requirements and rules for reusing medical data are on bilateral agreements. The LHA platform supports neither negotiations nor accounting with regard to providing access to data, models, and their implementation in software applications.

Secondly, the LHA also includes query tools, such as i2b2 \(^{13}\) and a self-developed data portal \(^{14}\). This necessitates to transfer data from uploaded data files to these query platforms. The goal of these platforms is allowing users to execute case count queries, i.e., for analyzing the cardinality of the (study) data by formulating the conditions for specific data items. Each condition consists of an attribute (feature name) which is compared with a constant (single value) or any kind of range, e.g., “age > 20”. Multiple conditions can be combined by logic operators, such as AND, OR, and NOT. These query platforms allow to get insights about the (study) data without accessing the data directly. We have defined two restrictions within the tools to avoid compromising published data. Firstly, the number of combined conditions is restricted to seven. Secondly, the smallest number of patients allowed is 10, otherwise the system displays “< 10” as message. The access to these query platforms is similar as to the access of data files. Instead of downloading data files, the user can get insights about the structure and value distribution of the data set. Finally, the LHA also provides different access levels to software applications implemented as models and methods. Applications are either publicly accessible, or accessible for users after authentication at LHA.

2.3 LHA Platform Components

The LHA platform consists of multiple, mostly pre-existing components. It uses the research data sharing platform FAIRDOM SEEK \(^{15}\) as a central component. It is enhanced by some
features to provide additional functionalities. For instance, contributors can annotate their uploaded content with concepts from the Human Disease Ontology, which support searching and filtering the content on the home page. Namely, entries can be linked to other (potential external) web pages and software tools. In this way, publication entries can be referred to original publications on journal or conference web sites. In the same way, entries are linked to installed software systems. We use i2b2 and a self-developed data portal system to allow running case count queries according to user specified conditions without having direct access to data. Associated frequency characteristics are generated online, i.e., whenever the user adds a new condition including or excluding patients, the data portal system instantly updates the number of patients/probands, e.g., taking the 10,000 probands of the LIFE-Adult-Study into account. Due to its distributed computing infrastructure, the data portal system scales well in both directions, towards increasing number of patients and towards the number of items which have been observed and recorded. Similarly, we link software applications as implementation for models accepted in different communities, e.g., for risk prediction of cancer subtypes, to entries in the central SEEK system. While the link to each (possibly external) web page and subsystem is stable, the content of referenced web pages and the version of subsystems (and, thus, their functionality) can change over time. Therefore, we carefully link different subsystems and especially external tools or web pages with the central SEEK hub. We suppose that scientific journal web pages will be stable for a long time, in particular, when we (re)use the DOI (Digital Object Identifier) and the online reference system doi.org. We use a continuous integration process to develop and deploy all software applications developed in the frame of the LHA-project at the University of Leipzig. Their source code is managed by a GitLab system from which the compiled code is deployed online to a cloud-based management system. Most of the current applications are developed as R-Shiny apps.

2.4 Method Annotation with Human Disease Concepts
In order to improve the findability and accessibility of the content in the Leipzig Health Atlas, users are able to annotate their uploaded publications, data, models and analyses with disease concepts extracted from the Human Disease Ontology (DO), which provides a common standard of human disease terms and their etiological descriptions \(^{16-18}\). Such a standardized corpus of disease concepts is critical for data sharing, effective interpretation of contextual data and rigorous computational analysis \(^{18}\) to annotate the content in the LHA. Figure 2 illustrates the integration of DO into LHA. Disease concepts are manually added by using the NCBI BioPortal \(^{19}\) search widget. It enables users to search for the desired disease concept where the actual metadata of the concept are extracted afterwards. In addition to definitions and available synonyms of the disease concept, the overview page also provides a link to Ontobee \(^{20}\), which lists further metadata, annotations and relations to other concepts. Finally, hierarchical relations to other concepts of the ontology and annotated content is listed below. The described DO functionalities are part of a new extension for SEEK, which has been developed in the LHA project and which was recently added to SEEK core.

2.5 Ethical consideration

All personal data provided by the LHA are from studies and projects, for which ethical approval has been obtained, together with informed written consent of all individuals.

3 Results and Applications

Since project start in 2016, we designed and implemented the platform architecture consisting of several software systems (see Fig. 1) and invited collaborating project consortia and study groups from the medical research domain to submit publications, publication related data, and models. In this way, LHA enabled transformation of available publication data into standardized formats and supported the development of innovative software applications and models of interest. In the following we sketch selected highlights provided by the platform.

3.1 Platform
The Leipzig Health Atlas has been designed and implemented as a web-based platform running in the production mode from early days on. The platform has a modular structure and consists of different components which are linked together (see Fig. 1). Initially LHA has been created only for sharing research data. Today its functionality is extended for sharing models and their applications, which enables the user for quickly applying and utilizing a model without any further implementation efforts. The platform is currently mainly used by medical and clinical scientists as well as by epidemiologists, medical computer scientists, biometricians and bioinformaticians. According to standard SEEK statistics the mean download frequency of content of the LHA is 28.60 downloads (sd = 59.52, max = 1,019), as of autumn 2021. The mean number of content accesses (i.e. requests to pages containing metadata of content) is 77.89 (sd = 114.53, max = 1,229) where requests send from web crawlers were not considered.

3.2 Networking/FAIRing data and researchers

With the launch of the LHA, we were well aware of the problem of "empty shelves" as one of the problems of data sharing, meaning that technically carefully constructed archives sometimes remained largely empty. In this project, we addressed the problem in several ways: (i) from the beginning, a series of research consortia were invited, involved and collaborations were established; (ii) we included the option to apply legal access barriers into FAIR policy, if asked for by our contributors; (iii) LHA uses permanent URLs for sustainable data sharing (last column in Table 1), (iv) LHA enables integrating own group web presences and, (v), it provides complex interactive models, which overall in this combination makes LHA unique, to the best of our knowledge.

Subsequently we provide examples of the models we are supporting in the LHA.

3.3 Tools for risk prediction in hereditary cancer

Hereditary breast and ovarian cancer (HBOC) and hereditary non-polyposis colorectal cancer (HNPCC / Lynch syndrome, LS) are cancer predisposition syndromes caused by germline
mutations in DNA repair genes and leading to considerably increased risks of cancer. These syndromes are clinically characterized by familial clustering of cancer and/or early age of disease onset. Valid predictions of mutation probabilities and cancer risks are indispensable to assist genetic counsellors and clinicians in their decisions to offer genetic testing or targeted preventive measures such as intensified cancer surveillance. A number of risk prediction models for HBOC and LS have been developed in the past years, but only few models have so far been implemented in convenient computer programs. Our goal was to implement and to provide the most important of these previously published models as an easy-to-use web-based application.

The "Manchester Scoring System" allows the calculation of BRCA1/2 mutation probabilities based on aggregated family history 22. The "GC-HBOC Mutation Frequency Explorer" enables the flexible assessment of mutation risks in BRCA1/2 and other genes for different sets of familial cancer histories based on a large dataset from the German Consortium for Hereditary Breast and Ovarian Cancer (GC-HBOC) 23. The "Extended Claus Model" (as implemented in the commercial pedigree drawing software Cyrillic 2.1.3, which is no longer supported and no longer works on newer operating systems) predicts both mutation and breast cancer risks based on structured pedigree data. "MMRpredict", "PREMM 1,2,6", and "PREMM 5" predict mutation risks in mismatch repair genes for patients from families suspected of having LS 24-26. Another tool calculates the age when a woman should be re-invited for the reassessment of her individual breast cancer risk to decide whether intensified surveillance is needed.

All web applications have been implemented in the programming language R (R Core Team, www.r-project.org) using the R-package shiny. Due to the legal requirements of the European Medical Device Regulation, all applications are currently available for scientific purposes only; authorization is needed. They have been reviewed by clinical users with regard to their practical applicability.
3.4 Models of blood formation during chemotherapy

Blood formation is a highly regulated physiological process. Cytotoxic chemotherapy and growth factor applications during cancer therapy results in complex interactions of various detrimental and stimulating effects on the haematopoietic system. In consequence, haematotoxic side effects are common but also highly heterogeneous and difficult to predict at an individual level. We developed a number of physiological models of blood formation under chemotherapy in the past with the aim to allow predictions of haematotoxic risks at an individual level, based on dynamical patient data. These dynamic models showed superior prediction performance compared to statistical models or simple physiological models. To facilitate usage of our models in clinical practice, we implemented them as Shiny applications. After uploading individual patient data and therapy specification, the model learns individual parameter estimates from the data and a large set of external data resources. We provided an easy-to-use tool allowing therapy adaptations in silico in order to extrapolate possible toxic responses of that patient under therapy alterations, which can be evaluated by the physician to select effective and tolerable treatment schedules (see Figure 4).

3.5 Normative epidemiological data for retinal nerve fiber layer thickness

Early detection of eye diseases is crucial to retain functional vision as long as possible. Eye health can be monitored by investigating retinal structure. Retinal nerve fiber layer thickness (RNFLT) is measured with optical coherence tomography (OCT). Those data can be used to detect changes potentially leading to optic neuropathies, such as Glaucoma. Analyses like such require normative data from extensive epidemiological cohorts to which individual data of a patient can be compared to. Usually device-based reference data is only based on a small sample size. Here we utilize a large sample from the LIFE-Adult study 12 obtained by high-resolution OCT- scans of the retina 27 and present new normative data for retinal nerve fiber layer thickness. This accounts for individual differences in refraction, age, sex and eye side28,29,27 via software application, called RNFLT(D)-Visualizer30. Via this app, eye specialists
are able to upload a pdf document of individual patient measurements. These values are extracted and compared with our new normative data. As a feature of this application, the user can comparatively evaluate the current state per eye (left, right), moreover, one can directly compare the current state of both eyes (differential evaluation). For these purposes, the application provides different visualizations to easily detect differences between user measurement and the normative data (device-specific as well as our new norms) as well as between different measurements itself. Figure 5 shows an example visualization of the application for one example patient compared to our new norms obtained from 4,483 healthy persons.

3.6 Phenotypic Profiles and PhenoMan

While the SEEK platform only allows to represent research data as data files, the LHA Data Portal \textsuperscript{14,10,31} supports efficient but flexible storing, querying and providing the research- and metadata making use of the CDISC ODM (Operational Data Model \textsuperscript{9}) format. It also uses the Study Data Query Language (SDQL), a novel domain specific language we created to specify filter criteria for querying the data \textsuperscript{32,33}. The SDQL reuses the conceptual abstract ODM entities to utilize known and well-defined vocabularies to enable an ODM-compliant data retrieval. Queries supporting the identification and classification of individuals or characteristics that meet specific phenotypic criteria (e.g., ‘select men aged 40-60 with myocardial infarction’) have been called phenotypic queries \textsuperscript{31}. They are used, e.g., for feasibility studies or to define study cohorts. To model and represent phenotypes as well as for generating phenotypic queries in SDQL, we developed the Phenotype Manager \textsuperscript{34,33} (PhenoMan), an ontology-based phenotyping framework \textsuperscript{34,33}. The PhenoMan imports the phenotype specifications (including generated queries) into the LHA using the extended REST interface of SEEK. After importing, the phenotype specification can be searched and referenced within the LHA platform. Figure 6 shows the LHA representation of the cut-off points for waist circumference as an indicator for risk of metabolic complications.
recommended by the World Health Organization. The queries are integrated as links to the LHA Data Portal and can be executed by clicking the links (magnifier icon). When the user is logged in, the Data Portal returns the corresponding results, i.e., the number of available persons meeting all criteria of the query.

3.7 oposSOM-Browser: An interactive access to health-related omics data

The LHA provides access to a series of studies addressing gene (dys-)regulation based on whole-genome transcriptomic data in the context of different diseases such as cancers, sepsis and pneumonia, Coeliac disease as well as of the blood transcriptomes of an epidemiological cohort of more than 3,300 healthy individuals. In addition, whole genome DNA-methylation and genetic single nucleotide variant data are provided for gliomas and human populations, respectively. These data are made available as so-called SOM-data for in-depth views using an interactive browser developed for the LHA. SOM-data were generated in the ‘selforganizing maps (SOM) high-dimensional data portraying’ workflow implemented in the oposSOM R-package. This application integrates functionalities such as expression landscape visualization, biomarker selection, function mining, sample stratification, diversity analyses, and phenotype mapping. The Leipzig Health Atlas makes this browsing tool available by providing a series of interactive functionalities such as the discovery of gene-expression landscapes and of their functional context, of associations with clinical phenotypes and also to search for prognostic predictions. Seven bulk transcriptome data sets, one single cell transcriptome, one methylome and two genome data sets are currently available. Figure 7 illustrates browser functions for the dataset of about 3,300 blood transcriptomes of healthy citizens of Leipzig collected in the LIFE-ADULT-Study. The screenshots refer browsing biological function, associations with phenotypic features, and cellular signalling cascades. This application is publicly accessible under https://apps.health-atlas.de/opossom-browser.

3.8 COVID-19
The Leipzig Health Atlas hosts an interactive monitoring tool which enables inspection of the dynamics of COVID-19 infection numbers worldwide, namely in 187 countries, using a trajectory-approach 40. Comparison of trajectories between countries and regions supports developing hypotheses and models to better understand the epidemiology of COVID-19. This application is publicly accessible under https://apps.health-atlas.de/covid-19-grapher.

In the frame of the oposSOM browser (previous subsection) the LHA provides SOM-portrayal data of the genomes of SARS-COV-2 variants observed until spring 2021 which enables studying the diversity of their mutational patterns, a task of emerging impact due to evolving immune and vaccination evasive virus variants 41. An update with novel variants, particularly Omecron, is presently in preparation.

We also provide pandemic data for the single states of Germany on a daily basis. These data comprise for example age specific incidences, deaths and occupied intensive care beds (https://apps.health-atlas.de/covid19-de-by-age/). Moreover, IMISE continuously publishes epidemiologic bulletins including model simulations and recommendations. All these bulletins are present in the Leipzig Health Atlas (https://www.health-atlas.de/, on the top of the start page).

4 Discussion and Outlook

The Leipzig Health Atlas is functional as a web-based, registered re3data 42 platform. It manages research studies, publications and the corresponding data and metadata as well as models. The LHA is designed first of all for scientists working in the medical domain including clinicians, epidemiologists, molecular and human geneticists, pathologists, biostatisticians and modellers. LHA is seeded around collaborative research projects and large national study consortia with involvement of Leipzig research groups, which is now becoming used by wider scientific communities. The LHA enables publishing extensive data as supplement to the information given in original research papers where simple access is realized via the LHA-ID. Thus, LHA supports open access to data and resources as required.
by many publishers. The LHA offers this option especially for scientists from the medical and epidemiological domains who want to make their data available for secondary use and/or more detailed views in the respective communities. All data and implemented models not necessarily need to be directly uploaded to LHA platform but can be managed somewhere else. LHA links such externally managed data and applications in a third-party research management system; their availability is tested automatically on regular basis; their content and functional state (reachable but does not work as it should be) is however never checked. By capturing relevant metadata, LHA provides a FAIR view on local and remote data and implemented models.

The LHA also offers an interesting solution under a technological perspective. The HL7 FHIR standard as well as popular Common Data Models (OMOP CDM, i2b2, PCORnet CDM) are implemented to focus on interoperability for data sharing. LHA uses international medical terminologies such as ICD-10, LOINC, or SNOMED CT. The use of these common information models (and data formats) and terminologies is checked by local administration whenever new data files are uploaded to the system (not for remotely managed data); we do only a formal check but leave the responsibility for the content to the scientist who is initially informed about and needs to accept the terms of use.

LHA functionalities overlap with a series of software and research data platforms in both, academic and commercial domains. For example, the FAIRDOMHub is a web-based repository for publishing FAIR data, operating procedures and models for the Systems Biology community. Similar to LHA, FAIRDOMHub utilizes the SEEK system as portal software infrastructure but manages research data for another community and, thus, with a slightly different focus. The Comprehensive Knowledge Archive Network (CKAN) is another powerful data management software that makes data accessible by providing tools to streamline data publishing, sharing and searching used by a number of governmental bodies e.g. for the European Data Portal. Dataverse is another open source web application to
share, preserve, cite, explore and analyze research data. Similarly, openBIS\textsuperscript{53,54} is a free and open source data management tool developed by ETH Zurich that supports the entire data lifecycle from project initiation, data production and analysis to make research data available to the public. Also OneData\textsuperscript{55} is a global data management system providing easy access to distributed storage resources and supporting a wide range of use cases from personal data management to data-intensive scientific computations.

In contrast to other platforms, the LHA aims at standardizing management of data and metadata. It reuses the CDISC ODM standard for data structure. A major difference to many other platforms in the medical domain is, that the LHA contains several applications as implementation of published and community accepted models, e.g., for risk prediction and comparative evaluation of different scores. These applications are for research use only since they are not certified for clinical use according to medical device regulations. The LHA platform implements such a series of applications to support different study groups and research projects and opens within the research community.

The LHA platform implemented own models to support own study groups and research projects for several reasons. Firstly, because of specific requirements of our user communities to adapt the platform, e.g., to allow differentiation between entity types including project and project consortia, for browsing through the entity research network. Secondly, because of privacy restrictions in managing study data. We collaborate with the Leipzig Clinical Trial Coordination Center ZKS-Leipzig to reuse expertise regarding data privacy and study organization.

With the European Open Science Cloud (EOSC), the Health Data Space, and the German National Research Data Infrastructure (NFDI) and specifically, the large consortium NFDI4Health (https://www.nfdi4health.de/en) within NFDI focusing on health research data management infrastructure, there are ongoing initiatives in Europe and Germany (NFDI) which set up a distributed and cloud-based infrastructure for medical research as European
counterpart to NCBI and NCI portals and toolsets. In particular, NFDI4Health, led by the German National Library of Medicine (ZBMED), will establish local (institutional) research hubs of university study centers, epidemiological studies, and clinical trials network which are then connected to a centralized search hub acting as global mediator and, therefore, allows to joins local research activities within Germany and with the connection to EOSC on European level as well. In NFDI4Health, LHA serves as template (in terms of infrastructure architecture, use and access rules, and content generation) for such local / institutional research hubs and, hence, acts as incubator for new infrastructure and medical research projects; latter sustainable leave project results at local research hubs or reuse available data from the overall infrastructure.

5 Conclusion

The LHA is an online platform to present and (securely) share data for researchers and medical experts. The users (data producer) can upload data by themselves and can link them to publication(s), project(s) / studies, and diseases. Moreover, LHA also includes novel applications as implementation of widely community accepted models, e.g., for risk prediction and other scenarios. All platform entries for projects, studies, publications, data and models / applications are described by rich metadata and are associated with Human Disease Ontology allowing for classification. The advantage of LHA is the joint availability of data, analysis results and models which makes the atlas a comprehensive sharing platform for research in the biomedical domain. Have a look and be very welcome at https://www.health-atlas.de/.

6 Acknowledgements

The Leipzig Health Atlas project was funded by the German Ministry of Education and Research with the reference number 031L0026 within the program i:DSem – Integrative Data Semantics in Systems Medicine. The project further has received funding from the Innovative
Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 853988 including the EU-(Horizont2020) funded imSAVAR project (to HLW, HB and ML).

References


Figure 1. Overview of OAIS Functional Entities in the context of the LHA platform. Displayed trademarks are CDISC, FAIRDOM SEEK, GitLab, i2b2, PHP and Shiny. Shiny is a trademark of RStudio, PBC. We are not associated with or sponsored by any of the trademark owners.

Figure 2: Integration of the Disease Ontology in the LHA web portal. The sunburst charts illustrate the distribution of LHA content in the Disease Ontology. Each layer of the chart illustrates a level in the taxonomy tree and the center of the chart is the root (i.e., “disease”).

Figure 3: Web user interface of the LHA platform

Figure 4: Individual patient data can be uploaded to derive individual parameter estimates. Future course of the therapy can be modified by sliders resulting in different predictions regarding haematotoxic response of the selected patient. This application is publicly accessible https://apps.health-atlas.de/thrombopenia.

Figure 5: Z-scores-plot of an inter-ocular analysis of the retinal nerve fiber layer thickness differences of both eyes of a patient. As displayed, the obtained difference is clearly established by such inter-eye comparison pointing out this unilateral defect. The data depict unilateral trauma to the left eye resulting in retinoschisis and degeneration of the retinal nerve fiber layer in that eye. Inter-ocular analysis is able to detect subtle differences, this is especially powerful in detection of early degeneration or disease, as all other systemic factors and potential confounders are constant for both eyes of one patient. This application is publicly accessible via https://apps.health-atlas.de/rnflt-visualizer/.
Figure 6: Cut-off points for waist circumference as an indicator for risk of metabolic complications

Figure 7: Screenshots of selected functionalities of the oposSOM-Browser for discovering the blood transcriptome data set as an example: a) The ‘Function browser’ window provides expression patterns of about 2,500 Gene Ontology sets. b) The ‘Phenotype browser’ window shows the sample network, which can be coloured according to different phenotypic features, such as gender, age, transcriptome-types or obesity status. c) The ‘Pathway browser’ provides gene activation patterns in more than 50 KEGG pathways such as the adipocyte signalling pathway, which reveals differing activation levels of the three horizontal signal cascades.

Figure 8: Covid-19Viewer- Screenshots of selected functionalities of the Covid19 trajectory browser: a) Case number trajectories (number of newly confirmed cases as a function of the total number of cases) for countries activated in the left-hand table. b) Time series of several characteristics such as currently and cumulative numbers of cases, deaths, vaccinations, and tests. Scale of y axis can be changed to logarithmic scale and to ‘per 1000 population’.

<table>
<thead>
<tr>
<th>Consortium Name</th>
<th>Research topic of consortium</th>
<th>Type of data shared</th>
<th>Health-Atlas ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLA</td>
<td>German Lymphoma Alliance</td>
<td>G, E, Me, P, Ph…</td>
<td>7RX4165T77-8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consortium Name</th>
<th>Research topic of consortium</th>
<th>Type of data shared</th>
<th>Health-Atlas ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLA</td>
<td>German Lymphoma Alliance</td>
<td>G, E, Me, P, Ph…</td>
<td>7RX4165T77-8</td>
</tr>
<tr>
<td>GGN</td>
<td>German Glioma Network</td>
<td>C, E</td>
<td>7Q0CF98QUE-7</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------------------------------</td>
<td>-------------</td>
<td>--------------</td>
</tr>
<tr>
<td>GC-HNPPCC</td>
<td>Hereditary Colorectal Cancer (Lynch)</td>
<td>C, G, risk models</td>
<td>7Q0CEYIJ25-4</td>
</tr>
<tr>
<td>GC-HBOC</td>
<td>Hereditary Breast and Ovarian Cancer</td>
<td>C, G, risk models</td>
<td>7Q0CE8DEEW-2</td>
</tr>
<tr>
<td>LIFE</td>
<td>Population based epidemiology ADULT Patient based cohort HEART</td>
<td>C, E, G, Phe-Signatures</td>
<td>7Q0CG2J40X-2</td>
</tr>
<tr>
<td>SMITH</td>
<td>Medical Informatics Initiative, Electronic medical records</td>
<td>C, Phe-Signatures</td>
<td>81CN2DP7WT-9</td>
</tr>
<tr>
<td>e:Med</td>
<td>different consortia and topics: Colorectal tumours (Lynch Syndrome), German Lymphoma Alliance, Community acquired Pneumonia</td>
<td>E, M, G, P, M (dynamic Models)</td>
<td>880R6A9NVP-0</td>
</tr>
<tr>
<td>DFG</td>
<td>Single cell RNAseq of early development and treatment</td>
<td>E (single cell), Ge</td>
<td>7QFYTCQM4-</td>
</tr>
<tr>
<td>Consortia</td>
<td>Description</td>
<td>ID</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>imSAVAR</td>
<td>Immune safety avatar: nonclinical mimicking of the immune system effects of immunomodulatory therapies</td>
<td>8AC22EUC5P-7</td>
<td></td>
</tr>
<tr>
<td>oBIG</td>
<td>Omics Bioinformatics for health</td>
<td>8AC22MUVW9-4</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: List of Consortia for which the LHA contains publications, data, and models. The depicted Health Atlas ID leads to the landing page of the whole consortia and the overview of all related content in LHA.
coronary artery disease

Details about this human disease

Synonyms (4)
- Coronary disease, coronary heart disease, CHD (coronary heart disease), coronary atherosclerosis

Definitions (2)
- An artery disease that is characterized by plaque building up along the inner walls of the arteries of the heart resulting in a narrowing of the arteries and a reduced blood supply to the cardiac muscles.

Human Disease Hierarchy
- artery disease [18]
  - coronary artery disease [6]
  - myocardial infarction [2]
  - atherosclerosis [6]
  - hypertension [6]

Related items
- Projects (1)  Resources (2)  Publications (5)
Body Mass Index, Waist Circumference and Waist-to-Hip Ratio

Description:
In participants of the LIFE-ADULT study (aged 10 to 79) [1], body height, body weight and different body circumferences were measured. Based on these characteristics, the Body Mass Index (BMI), waist circumference and waist-to-hip ratio were calculated and classified according to the recommendations of the World Health Organization (WHO) [2,3].

<table>
<thead>
<tr>
<th>Waist Circumference</th>
<th>BMI</th>
<th>Waist-to-Hip Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Description:
The WHO recommendations for waist circumference cut-off points as an indicator for risk of metabolic complications [3]:

<table>
<thead>
<tr>
<th>Waist circumference (men)</th>
<th>Waist circumference (women)</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 94 cm</td>
<td>&gt; 80 cm</td>
<td>Increased risk</td>
</tr>
<tr>
<td>&gt; 102 cm</td>
<td>&gt; 88 cm</td>
<td>Substantially increased risk</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Title</th>
<th>Unit</th>
<th>Formula</th>
<th>Range</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Waist</td>
<td>cm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Waist_Circumference</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Waist_Category_1</td>
<td></td>
<td>risk of metabolic complications not increased</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Waist_Category_2</td>
<td></td>
<td>risk of metabolic complications increased</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Waist_Category_3</td>
<td></td>
<td>risk of metabolic complications substantially increased</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>