Accepted Manuscript online: 2022-08-01 Article published online: 2022-12-23

The Leipzig Health Atlas—An Open Platform to Present, Archive, and Share Biomedical Data, Analyses, and Models Online

Toralf Kirsten[1,2,3,*] Frank A. Meineke[2,*] Henry Loeffler-Wirth[4] Christoph Beger[2] Alexandr Uciteli[2] Sebastian Stäubert[2] Matthias Löbe[2,3] René Hänsel[2] Franziska G. Rauscher[2,3] Judith Schuster[2] Thomas Peschel[2] Heinrich Herre[2] Jonas Wagner[2,3] Silke Zachariae[2] Christoph Engel[2,3] Markus Scholz[2] Erhard Rahm[5] Hans Binder[4,§] Markus Loeffler[2,3,4,§] on behalf of the LHA team[#]

[1] Department of Medical Data Science, Leipzig University Medical Center, Leipzig, Germany
[2] Institute for Medical Informatics, Statistics, and Epidemiology, Leipzig University, Leipzig, Germany
[3] Interdisciplinary Centre for Bioinformatics, Leipzig University, Leipzig, Germany
[4] LIFE Research Centre for Civilization Diseases, Leipzig University, Leipzig, Germany
[5] Department of Computer Sciences, Leipzig University, Leipzig, Germany
[6] Anhalt University of Applied Sciences, Köthen, Germany

Address for correspondence Toralf Kirsten, Department of Medical Data Science, Leipzig University, Härtelstraße 16-18, 04107 Leipzig, Germany (e-mail: toralf.kirsten@medizin.uni-leipzig.de).

Methods Inf Med 2022;61:e103–e115.

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 a basis for secondary research in evidence-based medicine, prediction models for disease, and its progression. This data are often neither sufficiently described nor accessible. Related models are often not accessible as a functional program tool for interested users from the health care and biomedical domains.

Objective The interdisciplinary project Leipzig Health Atlas (LHA) was 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 nonpublished data. LHA provides and associates unique permanent

Keywords

- FAIR
- data semantics
- research data management
- metadata
- clinical trials
- biomathematical models
- ontology
- risk prediction models
- omics data

* *These authors contributed equally.*
§ *These authors shared senior authorship.*
\# *Further LHA team members in alphabetical order: Anika Groß,[5,6] Ying-Chi Lin,[5] Katja Rillich,[2] Samira Zeynalova,[2] Marita Ziepert[2].*

received
March 10, 2022
accepted after revision
June 11, 2022
accepted manuscript online
August 1, 2022

DOI https://doi.org/10.1055/a-1914-1985.
ISSN 0026-1270.

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

identifiers for each dataset and model. Hence, the platform can be used to share prepared, quality-assured datasets 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.

Introduction

Scientific results are typically reported in (peer-reviewed) publications. In the medical research domain, clinical trials and epidemiological and molecular biological studies are the main sources for creating data for evidence-based medicine finally aiming at improving health care. 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 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[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 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 are 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 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[4] and country-specific laws. All these technical and organizational requirements have been bundled with the FAIR principles[5] for making research data Findable, Accessible, Interoperable, and Reusable, that came up in parallel with the start of project Leipzig Health Atlas (LHA) in 2016. Numerous articles have discussed FAIR practices and principles (see Wilkinson et al[6] and references cited therein).

To address these challenges, we have set up the 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.[7] 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 datasets and 34 models/tools associated with 891 publications and 34 scientific projects, mainly from the domains of medical research. The article is organized as follows: we introduce the main concepts and methods, provide an overview of available models and applications, and finally discuss special aspects in comparison with similar platforms.

Methods

Sharing Approach

The functional structure of the LHA is based on concepts of the reference model for an Open Archival Information System (OAIS).[8] This model provides a framework, including terminology and concepts, for describing and comparing architectures and operations of archives and, thus, for sharing their content. ►Fig. 1 shows an overview of functional entities implemented in the LHA. Clinical and epidemiological study groups and consortia provide data and describe metadata on different levels to the LHA in a publication-oriented manner. Each producer of content (left side in ►Fig. 1) can create new projects and publication entries which can be associated with datasets. 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 datasets 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 journal review.

The LHA platform provides basic rules for medical data management each user should follow. First, 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. Second, the LHA does not manage patient consents, neither in printed form nor

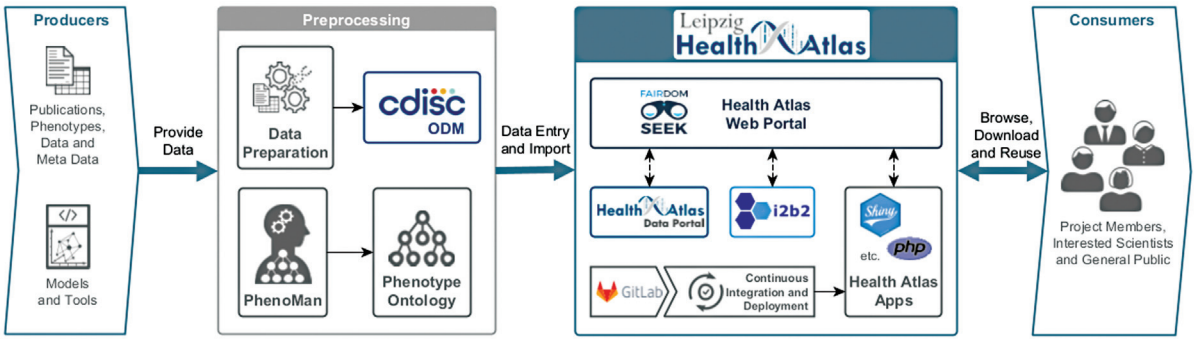

Fig. 1 Overview of OAIS functional entities in the context of the LHA platform. Displayed trademarks are CDISC, FAIRDOM SEEK, GitLab, i2b2, PHP, and Shiny. Shiny is a trademark of RStudio, PBC. We are not associated with or sponsored by any of the trademark owners.

electronically; this must be managed on the user side. Third, 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 OAIS standard, datasets 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.[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.[11]

Data Access and Reuse

Data are represented in two possible ways by the LHA platform: first, users can upload data as data files, preferably in the 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[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.

Second, 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 getting insights about the (study) data without accessing the data directly. We have defined two restrictions within the tools to avoid compromising published data. First, the number of

combined conditions is restricted to seven. Second, 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 dataset. 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.

LHA Platform Components

The LHA platform consists of multiple, mostly pre-existing components (see ►Fig. 1). 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 (DO), which support searching and filtering the content on the home page (see ►Fig. 2). 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 websites. In the same way, entries are linked to installed software systems. We use i2b2 and a self-developed data portal system[14] 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, toward increasing number of patients and toward 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.

Method Annotation with Human Disease Concepts

To improve the findability and accessibility of the content in the LHA, users are able to annotate their uploaded publications, data, models, and analyses with disease concepts extracted from the 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. ►Fig. 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 are 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 the SEEK core.

Results and Applications

Since the start of the project in 2016, we have 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.

Platform

The LHA has been designed and implemented as a web-based platform (see ►Fig. 3) 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 was 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.

Networking/FAIRing Data and Researchers

With the launch of the LHA, we were well aware of the problem of "empty shelves"[21] 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: (1) from the beginning, a series of research consortia were invited,

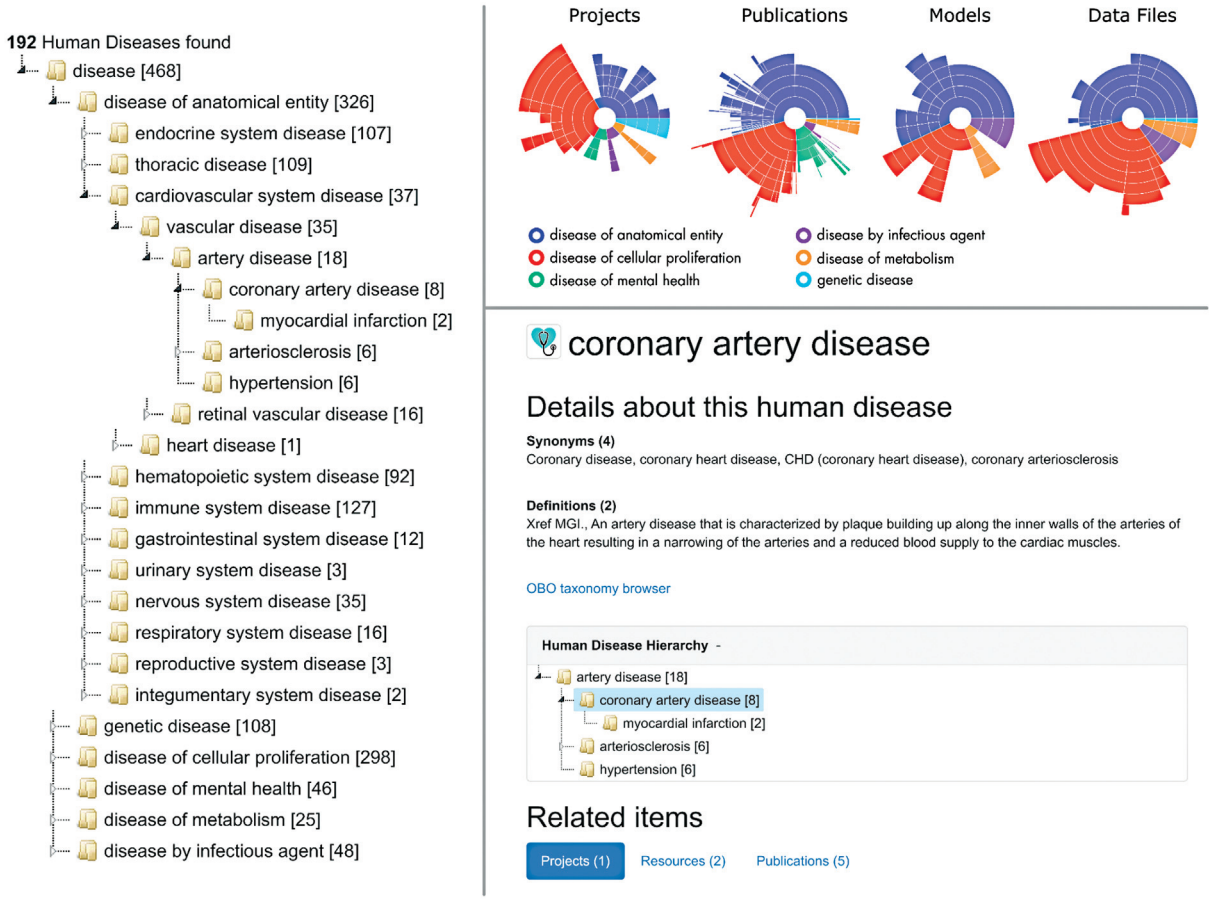

Fig. 2 Integration of the disease ontology in the LHA web portal. The sunburst charts illustrate the distribution of LHA content in the disease ontology. Each layer of the chart illustrates a level in the taxonomy tree and the center of the chart is the root (i.e., "disease"). LHA, Leipzig Health Atlas.

involved, and collaborations were established; (2) we included the option to apply legal access barriers into FAIR policy, if asked for by our contributors; (3) LHA uses permanent URLs for sustainable data sharing (last column in ►**Table 1**), (4) LHA enables integrating own group web presences and, (5), 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.

Tools for Risk Prediction in Hereditary Cancer

Hereditary breast and ovarian cancer (HBOC) and hereditary nonpolyposis 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. Several 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 HBOC (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;

Table 1 List of Consortia for which the LHA contains publications, data, and models

Consortium name	Research topic of consortium	Type of data shared C: clinical, E: expression Me: methylation, G: genetic, PH: phenotype, P: proteoma, M: model	Health-Atlas ID https://www.health-atlas.de/lha/<ID>
GLA	German Lymphoma Alliance (molecular mechanisms in malignant lymphoma, molecular and genetic profiling of DLBCL, randomized clinical trials on diffuse large B cell lymphoma and follicular lymphoma)	G, E, Me, P, Phe-signatures, M (risk model)	7RX4165T77–8
GGN	German Glioma Network	C, E	7Q0CF98QUE-7
GC-HNPCC	Hereditary Colorectal Cancer (Lynch)	C, G, risk models	7Q0CEYUJ25-4
GC-HBOC	Hereditary breast and ovarian cancer	C, G, risk models	7Q0CE8DEEW-2
LIFE	Population-based epidemiology ADULT patient-based cohort HEART	C, E, G, Phe-signatures	7Q0CG2J40X-2
SMITH	Medical Informatics Initiative, electronic medical records	C, Phe-signatures	81CN2DP7WT-9
e:Med	Different consortia and topics: colorectal tumors (Lynch syndrome), German Lymphoma Alliance, Community-acquired pneumonia	E, Me, G, P, M (dynamic models)	880R6A9NVP-0
DFG	Single-cell RNAseq of early development and treatment resistance of melanomas	E (single cell), Ge	7QFYTCQMN4-6
imSAVAR	Immune safety avatar: nonclinical mimicking of the immune system effects of immunomodulatory therapies	E (single cell), Phe	8AC22EUC5P-7
oBIG	Omics bioinformatics for health	G	8AC22MUVW9-4

Abbreviation: LHA, Leipzig Health Atlas.
Note: The depicted Health Atlas ID leads to the landing page of the whole consortia and the overview of all related contend in LHA.

authorization is needed. They have been reviewed by clinical users with regard to their practical applicability.

Models of Blood Formation during Chemotherapy

Blood formation is a highly regulated physiological process. Cytotoxic chemotherapy and growth factor applications during cancer therapy result in complex interactions of various detrimental and stimulating effects on the hematopoietic system. In consequence, hematotoxic side effects are common but also highly heterogeneous and difficult to predict at an individual level. We developed several physiological models of blood formation under chemotherapy in the past with the aim to allow predictions of hematotoxic risks at an individual level, based on dynamical patient data. These dynamic models showed superior prediction performance compared with 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 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 ►**Fig. 4**).

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 with. Usually device-based reference data are 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 RNFLT. This accounts for individual differences in refraction, age, sex, and eye side[27–29] via software

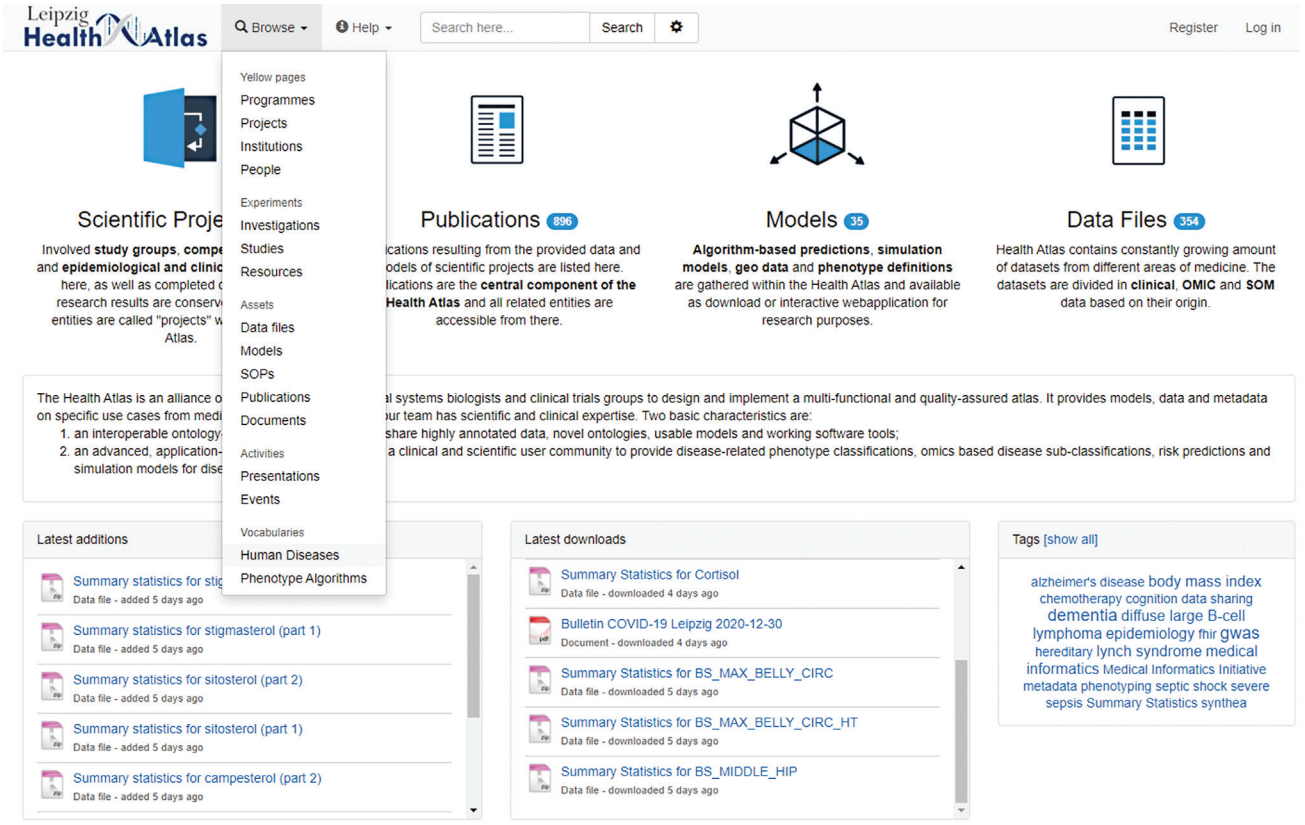

Fig. 3 Web user interface of the LHA platform. LHA, Leipzig Health Atlas.

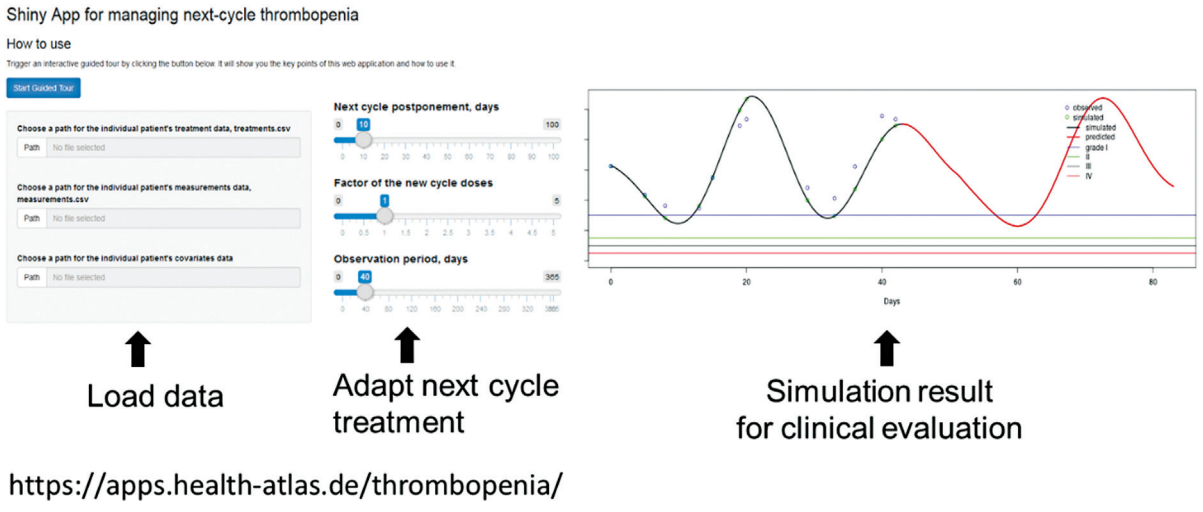

Fig. 4 Individual patient data can be uploaded to derive individual parameter estimates. Future course of the therapy can be modified by sliders resulting in different predictions regarding hematotoxic response of the selected patient. This application is publicly accessible at: https://apps.health-atlas.de/thrombopenia.

application, called RNFLT(D)-Visualizer.[30] 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. ►Fig. 5 shows an example visualization of the application for one example patient compared with our new norms obtained from 4,483 healthy persons.

Fig. 5 Z-score 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 the 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/.

Phenotypic Profiles and PhenoMan

While the SEEK platform only allows to represent research data as data files, the LHA Data Portal[10,14,31] supports efficient but flexible storing, querying, and providing the research and metadata making use of the CDISC ODM[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.[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.[33] 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 (PhenoMan), an ontology-based phenotyping framework.[33,34] 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. ►Fig. 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.[35] 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.[33]

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, celiac 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.[36] SOM-data were generated in the "self-organizing maps (SOM) high-dimensional data portraying" workflow[37] implemented in the oposSOM R-package.[38] This application integrates functionalities such as expression landscape visualization, biomarker selection, function mining, sample stratification, diversity analyses, and phenotype mapping. The LHA 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.[36] Seven bulk transcriptome datasets, one single cell transcriptome, one methylome, and two genome datasets are currently available. ►Fig. 7 illustrates browser functions for the dataset of approximately 3,300 blood transcriptomes of healthy citizens of Leipzig collected in the LIFE-ADULT-Study.[39] The screenshots refer to browsing biological function, associations with phenotypic features, and cellular signaling cascades.

This application is publicly accessible at: https://apps.health-atlas.de/opossom-browser.

COVID-19

The LHA hosts an interactive monitoring tool (see ►Fig. 8) 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

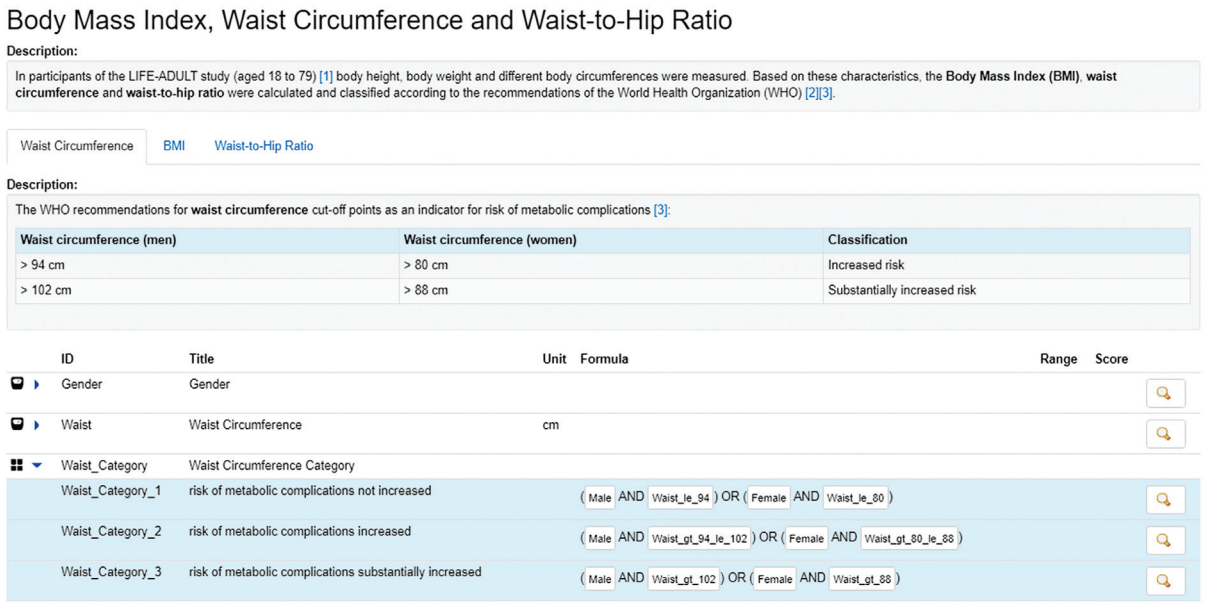

Fig. 6 Cut-off points for waist circumference as an indicator for risk of metabolic complications.

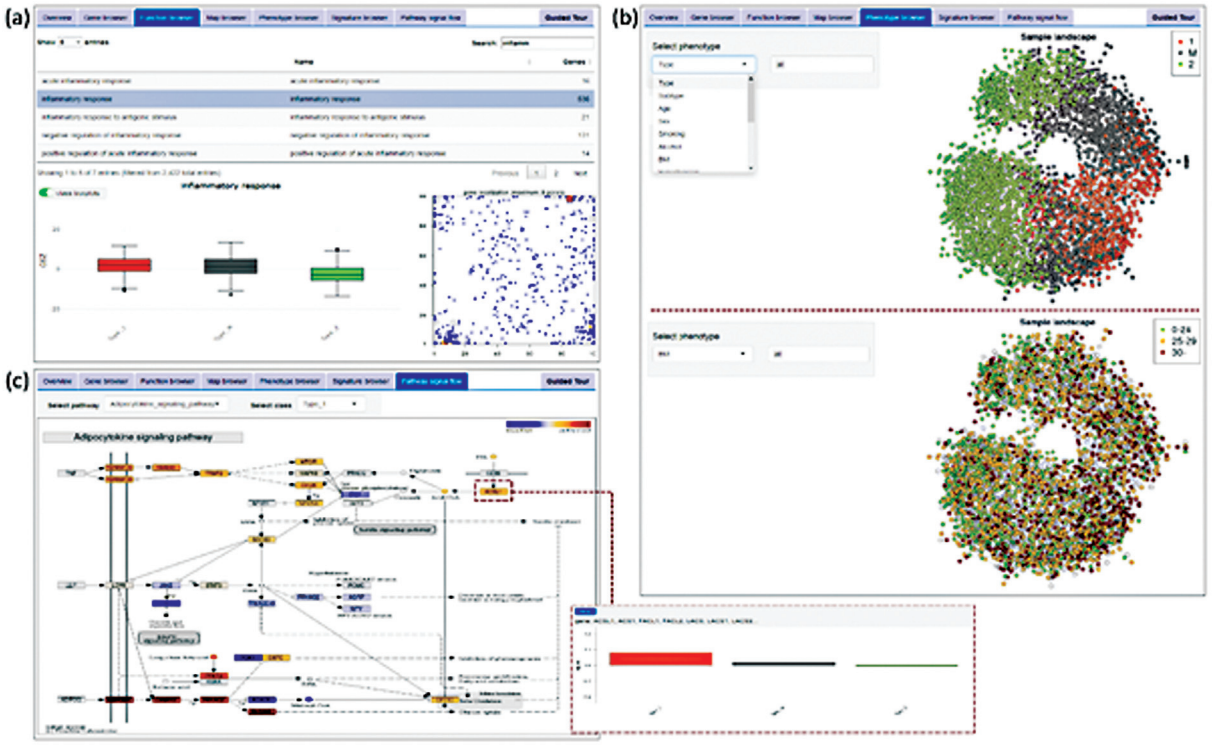

Fig. 7 Screenshots of selected functionalities of the oposSOM browser for discovering the blood transcriptome dataset as an example: **(a)** the "Function browser" window provides expression patterns of approximately 2,500 Gene Ontology sets. **(b)** The "Phenotype browser" window shows the sample network, which can be colored 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 signaling pathway, which reveals differing activation levels of the three horizontal signal cascades.

hypotheses and models to better understand the epidemiology of COVID-19. This application is publicly accessible at: 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 Omicron, is presently in preparation.

We also provide pandemic data for the single states of Germany on a daily basis. These data comprise for example

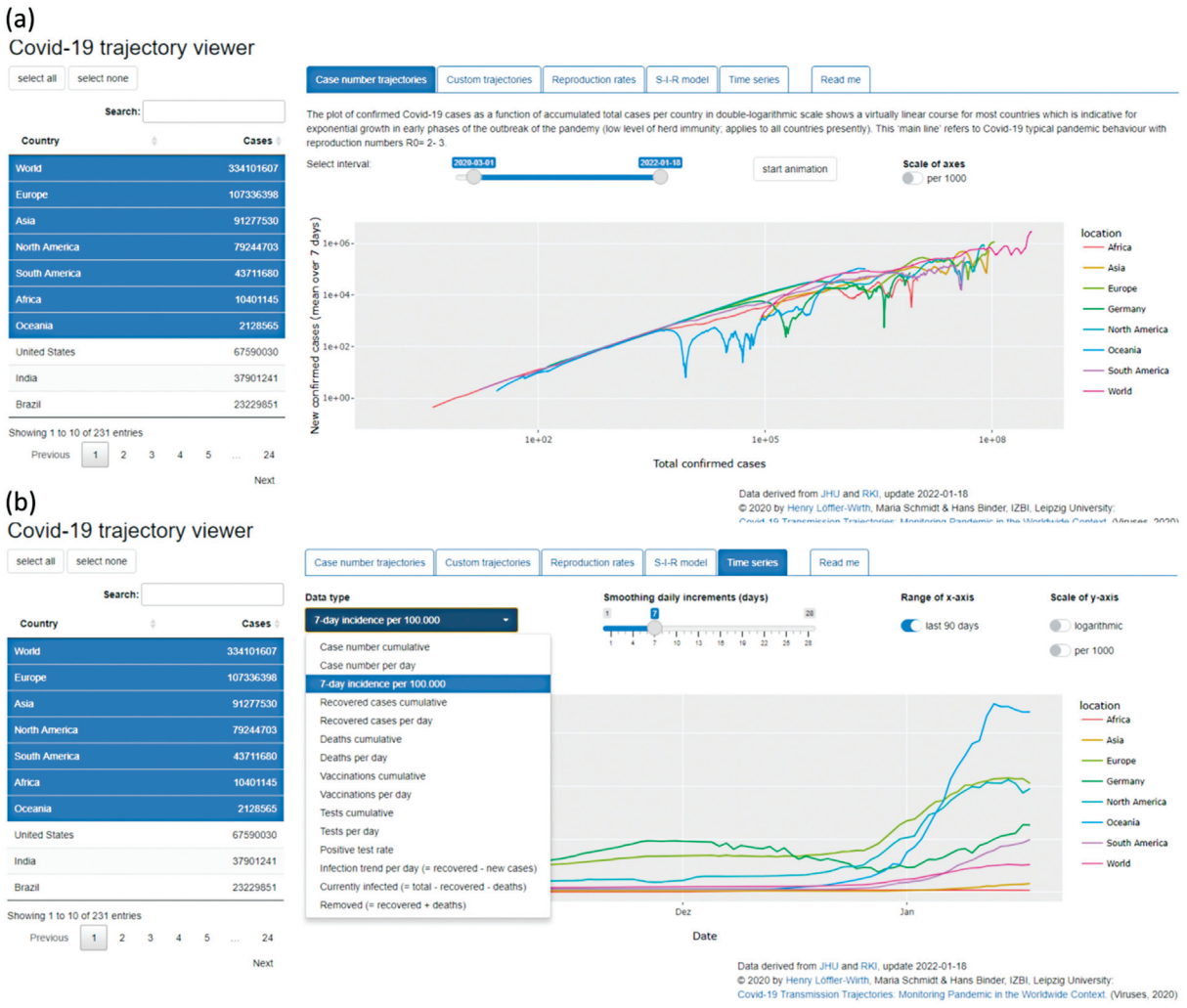

Fig. 8 Covid-19Viewer—screenshots of selected functionalities of the COVID-19 trajectory browser: **(a)** case number trajectories (number of newly confirmed cases as a function of the total number of cases) for countries activated in the left-hand table. **(b)** Time series of several characteristics such as current and cumulative numbers of cases, deaths, vaccinations, and tests. Scale of *y*-axis can be changed to a logarithmic scale and to "per 1000 population."

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 LHA (https://www.health-atlas.de/, on the top of the start page).

Discussion and Outlook

The LHA 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 modelers. LHA is seeded around collaborative research projects and large national study consortia with involvement of Leipzig research groups, which is now being used by wider scientific communities. The LHA enables publishing extensive data as a 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 the 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) are 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[43] as well as popular Common Data Models (OMOP CDM,[44] i2b2,[13] PCORnet CDM[45]) are implemented to focus on interoperability for data

sharing. LHA uses international medical terminologies such as ICD-10[46], LOINC,[47] or SNOMED CT.[48] 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[49] 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[50] is another powerful data management software that makes data accessible by providing tools to streamline data publishing, sharing, and searching used by several governmental bodies, e.g., for the European Data Portal. Dataverse[51,52] is another open-source web application to share, preserve, cite, explore, and analyze research data. Similarly, openBIS[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[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: first, 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; second, 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, NFDI4-Health, led by the German National Library of Medicine, 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 joining local research activities within Germany and with the connection to EOSC on European level as well. In NFDI4-Health, 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 an 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.

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 DO allowing for classification. The advantage of LHA is the joint availability of data, analysis results, and models which make 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/.

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.

Funding/Acknowledgments

The LHA 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 (Horizon 2020)-funded imSAVAR project (to H.L-.W., H.B., and M.L.).

Conflict of Interest

None declared.

References

1 Dryad. Homepage Dryad Digital Repository. Accessed February 10, 2022, at: https://datadryad.org

2 Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30(01):207–210
3 National Center for Biotechnology Information. Homepage Gene Expression Omnibus. Accessed February 10, 2022, at: https://www.ncbi.nlm.nih.gov/geo/
4 European Parliament. European Council. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*. 2016;(L 119):1–88. Accessed December 10, 2021, at: https://eur-lex.europa.eu/eli/reg/2016/679/oj
5 Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018
6 FAIR Principles in Health Research; 2020. Available at: https://www.thieme-connect.de/products/ejournals/issue/10.1055/s-011-50674
7 Herre H. Towards a new foundational ontology of properties, attributives and data. In: Borgo S, ed. Ontology Makes Sense: Essays in Honor of Nicola Guarino. Amsterdam: IOS Press Incorporated; 2019:194–210
8 ISO. Space Data and Information Transfer Systems—Open Archival Information System (OAIS)—Reference Model. 2nd ed. 2012;14721:2012. Published 2012–09. Accessed December 10, 2021, at: https://www.iso.org/standard/57284.html
9 CDISC. Operational Data Model (ODM). Accessed December 10, 2021, at: https://www.cdisc.org/standards/data-exchange/odm
10 Kirsten T, Kiel A, Wagner J, Rühle M, Löffler M. Selecting, Packaging, and Granting Access for Sharing Study Data. In: Eibl M. and Gaedke M., eds. INFORMATIK 2017. Bonn: Gesellschaft für Informatik; 2017:1381–1392
11 Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018;57(Suppl 1):e50–e56
12 Loeffler M, Engel C, Ahnert P, et al. The LIFE-Adult-Study: objectives and design of a population-based cohort study with 10,000 deeply phenotyped adults in Germany. BMC Public Health 2015;15:691
13 Murphy S, Churchill S, Bry L, et al. Instrumenting the health care enterprise for discovery research in the genomic era. Genome Res 2009;19(09):1675–1681
14 LIFE – Leipziger Forschungszentrum für Zivilisationserkrankungen. LIFE Datenportal. Accessed February 10, 2022, at: https://ldp.life.uni-leipzig.de/
15 Wolstencroft K, Owen S, Krebs O, et al. SEEK: a systems biology data and model management platform. BMC Syst Biol 2015;9:33
16 Schriml LM, Arze C, Nadendla S, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res 2012;40(Database issue):D940–D946
17 Schriml LM, Mitraka E. The Disease Ontology: fostering interoperability between biological and clinical human disease-related data. Mamm Genome 2015;26(9–10):584–589
18 Schriml LM, Mitraka E, Munro J, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res 2019;47(D1):D955–D962
19 Whetzel PL, Noy NF, Shah NH, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 2011;39(Web Server issue):W541–5
20 Ong E, Xiang Z, Zhao B, et al. Ontobee: a linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic Acids Res 2017;45(D1):D347–D352
21 Nelson B. Data sharing: empty archives. Nature 2009;461 (7261):160–163
22 Evans DGR, Lalloo F, Cramer A, et al. Addition of pathology and biomarker information significantly improves the performance of the Manchester scoring system for BRCA1 and BRCA2 testing. J Med Genet 2009;46(12):811–817
23 Kast K, Rhiem K, Wappenschmidt B, et al; German Consortium for Hereditary Breast and Ovarian Cancer (GC-HBOC) Prevalence of BRCA1/2 germline mutations in 21 401 families with breast and ovarian cancer. J Med Genet 2016;53(07):465–471
24 Barnetson RA, Tenesa A, Farrington SM, et al. Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. N Engl J Med 2006;354(26):2751–2763
25 Kastrinos F, Steyerberg EW, Mercado R, et al. The PREMM(1,2,6) model predicts risk of MLH1, MSH2, and MSH6 germline mutations based on cancer history. Gastroenterology 2011;140(01): 73–81
26 Kastrinos F, Uno H, Ukaegbu C, et al. Development and validation of the $PREMM_5$ model for comprehensive risk assessment of Lynch syndrome. J Clin Oncol 2017;35(19):2165–2172
27 Baniasadi N, Rauscher FG, Li D, et al. Norms of interocular circumpapillary retinal nerve fiber layer thickness differences at 768 retinal locations. Transl Vis Sci Technol 2020;9(09):23
28 Wang M, Elze T, Li D, et al. Age, ocular magnification, and circumpapillary retinal nerve fiber layer thickness. J Biomed Opt 2017;22(12):1–19
29 Li D, Rauscher FG, Choi EY, et al. Sex-specific differences in circumpapillary retinal nerve fiber layer thickness. Ophthalmology 2020;127(03):357–368
30 Peschel T, Wang M, Kirsten T, Rauscher FG, Elze T. A cloud-based infrastructure for interactive analysis of RNFLT data. [accepted for publication]. In: 17th IEEE eScience 2021 International Conference; 2021. Accessed February 10, 2022, at: https://www.escience2021.org/
31 Kiel A, Wagner J, Rühle M, Twrdik A. Lens - The system behind the LIFE Data Portal. In: Arendt T, Heiker JT, Magin T, Schaefer M, Schulz-Siegmund M, Thiery J, eds. 15th Leipzig Research Festival for Life Sciences; 2019:186. Accessed December 10, 2021, at: http://www.resfest.uniklinikum-leipzig.de/pdf/2019_01_16-Abstract-Book-Research-2019.pdf
32 Wagner J. Softwaregestützte Bereitstellung Von Epidemiologischen Forschungsdaten [Master's thesis]. Leipzig, Germany: Leipzig University of Applied Sciences; 2016
33 Uciteli A, Beger C, Wagner J, et al. Ontological modelling and execution of phenotypic queries in the Leipzig Health Atlas. Stud Health Technol Inform 2021;278:66–74
34 Uciteli A, Beger C, Kirsten T, Meineke FA, Herre H. Ontological representation, classification and data-driven computing of phenotypes. J Biomed Semantics 2020;11(01):15
35 WHO Team Nutrition and Food Safety Waist circumference and waist-hip ratio: report of a WHO expert consultation; 2008. Accessed February 10, 2022, at: https://www.who.int/publications/i/item/9789241501491
36 Loeffler-Wirth H, Reikowski J, Hakobyan S, Wagner J, Binder H. oposSOM-Browser: an interactive tool to explore omics data landscapes in health science. BMC Bioinformatics 2020;21(01):465
37 Wirth H, Löffler M, von Bergen M, Binder H. Expression cartography of human tissues using self organizing maps. BMC Bioinformatics 2011;12:306
38 Löffler-Wirth H, Kalcher M, Binder H. oposSOM: R-package for high-dimensional portraying of genome-wide expression landscapes on bioconductor. Bioinformatics 2015;31(19):3225–3227
39 Schmidt M, Hopp L, Arakelyan A, et al. The human blood transcriptome in a large population cohort and its relation to aging and health. Front Big Data 2020;3:548873
40 Loeffler-Wirth H, Schmidt M, Binder H. Covid-19 transmission trajectories-monitoring the pandemic in the worldwide context. Viruses 2020;12(07):E777
41 Schmidt M, Arshad M, Bernhart SH, et al. The evolving faces of the SARS-CoV-2 genome. Viruses 2021;13(09):1764
42 re3data.org. Leipzig Health Atlas. 2018. Accessed February 10, 2022, at: https://www.re3data.org/repository/r3d100012652

43 HL7. FHIR Release 4 (Technical Correction #1) (v4.0.1). Accessed February 10, 2022, at: https://www.hl7.org/fhir/R4/
44 Observational Health Data Sciences and Informatics. Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). 2022. Accessed February 10, at: https://www.ohdsi.org/data-standardization/the-common-data-model/
45 National Patient-Centered Clinical Research Network. PCORnet Common Data Model. Accessed February 10, 2022, at: https://pcornet.org/data/
46 World Health Organization. ICD-10: International Statistical Classification of Diseases and Related Health Problems: Tenth Revision. 2nd ed. 2004. Accessed December 10, 2021, at: https://apps.who.int/iris/handle/10665/42980
47 Regenstrief Institute. Logical Observation Identifiers Names and Codes (LOINC). Accessed February 10, 2022, at: https://loinc.org/
48 SNOMED International. Systematized Nomenclature Of Medicine Clinical Terms (SNOMED CT). Accessed February 10, 2022, at: https://www.snomed.org
49 HITS gGmbH. FAIRDOMHub. Accessed February 10, 2022, at: https://fairdomhub.org/
50 Open Knowledge Foundation. Comprehensive Knowledge Archive Network (CKAN). Accessed February 10, 2022, at: https://ckan.org/
51 King G. Homepage Dataverse. Accessed February 10, 2022, at: https://dataverse.org/
52 King G. An Introduction to the Dataverse Network as an infrastructure for data sharing. Sociol Methods Res 2007;36(02): 173–199
53 ETH Zurich Scientific IT Services. Homepage openBIS. Accessed February 10, 2022, at: https://openbis.ch
54 Bauch A, Adamczyk I, Buczek P, et al. openBIS: a flexible framework for managing and analyzing complex data in biology research. BMC Bioinformatics 2011;12:468
55 onedata. Homepage onedata. Accessed February 10, 2022, at: https://onedata.org