Cohort Harmonization and Integrative Analysis from a Biomedical Engineering Perspective

Konstadina D. Kourou, Vasileios C. Pezoulas, Eleni I. Georga, *Student Member, IEEE,* Themis P. Exarchos, *Member, IEEE,* Panayiotis Tsanakas, Manolis Tsiknakis, Theodora Varvarigou, Salvatore De Vita, Athanasios Tzioufas, and Dimitrios I. Fotiadis, *Senior Member, IEEE*

***Abstract*— In this review the critical parts and milestones for data harmonization, from the biomedical engineering perspective, are outlined. The need for data sharing between heterogeneous sources pave the way for cohort harmonization; thus, fostering data integration and interdisciplinary research. Unmet needs in chronic as well as in other diseases, can be addressed based on the integration of patient health records and the sharing of information of the clinical picture and outcome. The stratification of patients, the determination of various clinical and outcome features and the identification of novel biomarkers for the different phenotypes of the disease characterize the impact of cohort harmonization in patient-centered clinical research and in precision medicine. Subsequently, the establishment of matching techniques and ontologies for the creation of data schemas are also presented. The exploitation of web technologies and data-collection tools support the opportunities to achieve new levels of integration and interoperability. Ethical and legal issues which arise when sharing and harmonizing individual-level data are discussed in order to evaluate the harmonization potential. Use cases that shape and test the harmonization approach are explicitly analyzed along with their significant results on their research objectives. Finally, future trends and directions are discussed and critically reviewed towards a roadmap in cohort harmonization for clinical medicine.**

I. Introduction

Cohort studies offer invaluable sources of biological, health, environmental, behavioural and psychosocial data and have given rise to multiscale predictive data analysis. The information gathered from large population-based cohorts allow to leverage public health and clinical medicine. Moreover, access to studies that incorporate different types of research data would permit the investigation of direct and/or indirect disease aetiological determinants. To this end, the analysis of synthesized datasets across population-based studies is set to become increasingly important [1]. The heterogeneity of existing cohorts, stemming from variability in experimental design, measurement and standardization methods, supports the realization of aggregated schemes towards the development of precision medicine solutions. Additionally, making the data findable, accessible, interoperable and reusable enables data sharing. Data sharing includes: (i) the procedures for data access, (ii) the mechanisms for dissemination, (iii) the tools and software for data re-use, (iv) the definition whether data access should be widely open or restricted, and (v) the prohibition period that may exist [2].

The need for data harmonization enables the cross-national and international comparative research, while it enables the investigation of similarities and differences across longitudinal datasets when compatible data are available [3]. Data harmonization refers to the creation of a standardized, comprehensive database/schema, combining data generated from different sources (e.g. epidemiological studies, clinical studies), and facilitating data integration and interdisciplinary research. The increase of sample size and the improvement of generalizability and validity of research results constitute the most significant benefits of the harmonization process [4, 5]. Global initiatives have incorporated data harmonization in their design in order to investigate and/or identify risk factors of complex chronic diseases. Examples of such initiatives aim to bring out the value of health data through large scale analytics and ensure the development of harmonized measures and standardized computer infrastructures [6]. The ability to effectively harmonize data from different studies and patient cohorts facilitates the rapid extraction of new scientific knowledge on disease onset, disease progression and classification of different disease phenotypes. This type of approach can certainly address the unmet needs of chronic diseases including: (i) validation or identification of novel biomarkers, (ii) patient stratification for different treatments, as well as (iii) selection of new targets for therapy. Furthermore, this approach will enable the decisive evaluation of disease features, co-morbidities, remissions, exacerbations etc. Additionally, it will allow the scientific

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731944 and from the Swiss State Secretariat for Education, Research and Innovation SERI under grant agreement 16.0210.

K.D. Kourou is with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, GR45110 and with the Dept. of Biological Applications and Technology, University of Ioannina, Ioannina GR45110, Greece.

V.C Pezoulas, E.I. Georga, T.P. Exarchos and D.I. Fotiadis are with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, Ioannina GR45110, Greece (corresponding author email: fotiadis@cc.uoi.gr). D.I. Fotiadis and T.P. Exarchos Department of Biomedical Research, Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology-Hellas (FORTH), Ioannina GR45110, Greece.

P. Tsanakas is with the National Technical University of Athens, Greece, Athens GR115 23, Greece.

M. Tsiknakis is with the Biomedical Informatics and eHealth lab, Technological Educational Institute of Crete and with the Computational Medicine laboratory, Institute of Computer Science, FORTH, Heraklion GR70013, Greece.

T. Varvarigou is with the Division of Communication, Electronic and Information Engineering, School of Electrical and Computer Engineering, National Technical University of Athens, Athens, GR15780, Greece.

S. De Vita is with the Clinic of Rheumatology, Department of Medical and Biological Sciences, Udine University, Udine, IT33100, Italy.

A. Tzioufas is with the Dept. of Pathophysiology, School of Medicine, University of Athens GR15772, Greece.

societies to collaborate efficiently with patient associations as well as the regulatory mechanisms in each country.

Research guidelines for rigorous data harmonization encompass: (i) the definition of the research questions, objectives and protocol, (ii) the collection of appropriate information and selection of studies, (iii) the definition of variables and evaluation of harmonization potential, (iv) the processing of data, and (v) the estimation of the quality of the harmonized dataset [7]. Evaluation of harmonization potential refers to the clinical evaluation of whether a clinical dataset under examination is capable of being harmonized. This procedure is related to two factors, namely: (i) the data sharing regulations, i.e., whether the sharing of the clinical data is possible or not to enable harmonization, and (ii) the quality of the dataset, i.e., if the dataset meets the minimum requirements. In fact, the quality of the dataset determines the harmonization potential in a technical manner. For example, a well-qualified dataset will lead to a larger matching probability with the reference model than a less qualified dataset. A reference model consists of the minimum number of features (e.g. clinical parameters) which enable the explicit description of a disease's domain knowledge. Data integration strategies, taking advantage of contemporary Semantic Web technologies, aim at maximizing the exploitation potential of synthesized datasets across multidisciplinary studies. Indicative techniques rely on automated or supervised creation of data schemas using vocabularies and ontology mapping [8, 9].

This review analyses previous research and presents new trends on longitudinal cohort harmonization, while it discusses the significant challenges occurring when managing and harmonizing large amounts of data from different sources. Ethical, legal and related restrictions associated with sharing and pooling individual-level information are also discussed. Current techniques, ranging from data standardization models to advanced ontology-based and semantic interlinking technologies, enabling the creation of links among diverse datasets, are presented. Several case studies, related to the data pooling and harmonization of disparate cohorts, are presented and discussed. These approaches reveal the benefits of integrating diverse sets of multilevel biomedical data so as to compare and enhance the statistical power of validated models, as well as improve the management of the healthcare system. Furthermore, future directions in the field are discussed. Considering that new technologies have not been holistically validated and evaluated, the present review contributes positively to the adoption of new practices in data harmonization and scalable cohort analytics for knowledge discovery.

Data harmonization is more prominent in chronic diseases, in order to address unmet needs, such as: (i) patient stratification and (ii) biomarker identification and/or validation. This comprises a fundamental basis of our work, while additional references to other diseases with their related studies are considered (e.g. obesity and ageing). The following sections of this review constitute a comprehensive description of (i) data harmonization from the clinical perspective and its importance on addressing the unmet needs in chronic diseases, (ii) the technical perspectives when data integration should be conducted, (iii) the ethical and legal issues, (iv) the existing case studies related to data harmonization and the (iv) discussion for supporting co-analysis of harmonized datasets from a biomedical engineering perspective.

II. HARMONIZATION IN CHRONIC DISEASES

The application of existing or newly identified therapeutic targets in everyday's clinical practice is hampered by the incomplete understanding of chronic disease pathogenesis, the lack of validation of potential indices in large patient cohorts and the disease heterogeneity. In addition, the complex pathogenetic mechanisms involving, to a different extent, various systems (immune, endocrine, neuroendocrine, etc.) further hinder the successful application of a sole omics' technology (as attested by genome-wide association studies) for the discovery of novel pathogenetic/therapeutic targets. These observations highlight the need for the classification of patients with chronic disease in subgroups with homogeneous clinical profiles, disease duration and similar outcome, as well as for trans-disciplinary research involving clinical and basic researchers for the delineation of the underlying pathogenetic pathways and the identification/validation of novel therapeutic targets.

Besides the clinical and medical research observations that support the need for integrating different cohorts in chronic diseases, there are additional considerations and rationales that make the integration of cohorts an urgent need. In terms of theoretical rationale, the integration of different cohorts can solve the issue of generalizability. The importance of generalization ability of study outcomes has been discussed in the literature. For cohort studies, limited representativeness may mean that results cannot be assumed to be true for un-surveyed or underrepresented subsets of the population of interest. The second theoretical rationale has to do with the interest in known or potential sources of heterogeneity, which are mainly syntactic, semantic and/or conceptual and may also exist due to the different geographical locations of the patients. In addition, other sources of such heterogeneity refer to factors that can influence differences within or between study outcomes (e.g. genetics). Cross-national studies allowing assessment of the influence of social policy and environment on health outcomes are perhaps the best known examples of such investigations. The third theoretical consideration has to do with the specific unmet needs of chronic disease. The inferences, causalities, associations or new knowledge that researchers wish to draw from combined data determine whether pooling individual data is a worthwhile undertaking. Finally, a fourth theoretical consideration deals with the directions of future research. The areas for future research can be identified based on gaps in current knowledge. However, since most of the studies and literature in chronic diseases refer to single regional or national cohort analysis, pooled data analyses of integrative cohorts may assist in addressing new research questions. The second type of rationale has to do with the statistics and the benefits of increased sample size. Low prevalence outcomes are infrequently reported in local cohorts. Hence, larger samples are required to obtain adequate data for the analysis

Table I. Unmet needs in chronic diseases that can be addressed through data harmonization.

Unmet Needs		Addressing unmet needs through data harmonization	Related work(s)
1. Validation or identification of novel biomarkers	Candidate disease biomarkers are further studied by systems biology approaches aiming to validate or identify clinical, laboratory and molecular biomarkers for early disease diagnosis, follow-up and response to treatment.	Harmonization of cohorts will further enable the search and identification of novel biomarkers. The effective data harmonization facilitates the rapid extraction of new scientific knowledge on disease progression.	[78]
2. **Stratification of patients to distinct subgroups**	Patient can be stratified based on the clinical picture (i.e. mild or severe disease), the histological observations and the molecular characteristics.	Harmonization of cohorts will increase the sample size of the research results. Thus, patient subgroups can be easily extracted based on their clinical, histological and molecular characteristics.	[1], [4] and [80]
3. Selection of new targets for therapy	According to the stratification of patients to homogeneous patient subgroups, targeted therapy can be defined.	Harmonization of cohorts will enable the identification of novel therapies according to the stratification of all patients. Homogeneous patient subgroups will further allow clinicians to find new targets for therapy.	[4], [73] and [78]

of predictors and consequences of these outcomes. The last category of rationale has to do with practical considerations. When using existing data, results may be obtained in a timely manner compared to initiating a new study. Consequently, answers to research questions can be accelerated, particularly where outcomes are of importance to the public, or when the unmet needs have already been identified. Such results are relatively cost efficient, and can often be obtained without unnecessary duplication of tasks or additional burden on the target population. In the case of survey based cohort studies, reducing the necessity of identifying additional participants, and sending, receiving and processing survey data for analysis, may mean that the greatest costs associated with survey methods are eliminated.

Facilitating the integration of patient records and the sharing of information of the clinical picture and outcome will permit the stratification of patients, the determination of various clinical and outcome features, such as co-morbidities and mortality ratios, which up to now have been studied in various small patient groups, the development of diagnostic, follow-up and therapeutic algorithms, as well as biomarkers for the different phenotypes of the disease. Thus, unmet needs in chronic diseases can be addressed, namely: (i) the validation or identification of clinical, laboratory and molecular biomarkers for early disease diagnosis, follow-up and response to treatment, (ii) the stratification of patients to distinct subgroups according to clinical picture and outcome for different treatments and (iii) the selection of new targets for therapy. Table I summarizes the unmet needs that characterize chronic diseases and how they can be addressed through data harmonization and data sharing, along with related works which have been published in the scientific and clinical research literature. Additionally, Fig.1 illustrates a conceptual step-wise categorization in chronic diseases along with the unmet needs that can be addressed through harmonization. Towards this categorization particular steps are considered with reference to: (i) the clinical stratification, (ii) the histologic stratification, (iii) the systems biology approaches for patient molecular stratification and (iv) the homogeneous patient groups for targeted therapy. Hence, unmet needs in chronic as well as other diseases can be further considered and studied thoroughly regarding the utilization of the harmonization perspective [10]. In order to achieve harmonization and address the unmet needs specific technical steps which contribute to the definition of the research objectives and the analysis of harmonized datasets should be followed. Moreover, the analysis of the output results can be further achieved through appropriate IT infrastructures and algorithms that allow the federated analysis of pooled datasets.

III. TECHNICAL PERSPECTIVE

This section introduces the research tools which enables researchers to address the unmet needs in chronic diseases through their utilization based on the harmonization initiative. Achieving data harmonization as a rigorous scientific process is a demanding task. Data harmonization, from the technical perspective, is an iterative process consisting of related and inter-dependent steps. To better understand the key steps towards data harmonization approach, effective and rigorous technical methods and tools have been developed [1].

Recently, the iterative key steps towards harmonization of research data have been delineated by Maelstrom Research and its partners [7]. These principles aim to promote a systematic procedure for data harmonization and a methodological guidance for researchers which integrate heterogeneous data from different cohorts for the synthesis of harmonized datasets [10]. An overview of the key steps towards data harmonization process is depicted in Fig. 2. The four closely related steps that should be considered during the harmonization workflow include: (A) Definition of the objectives and research questions, (B) Data discovery and

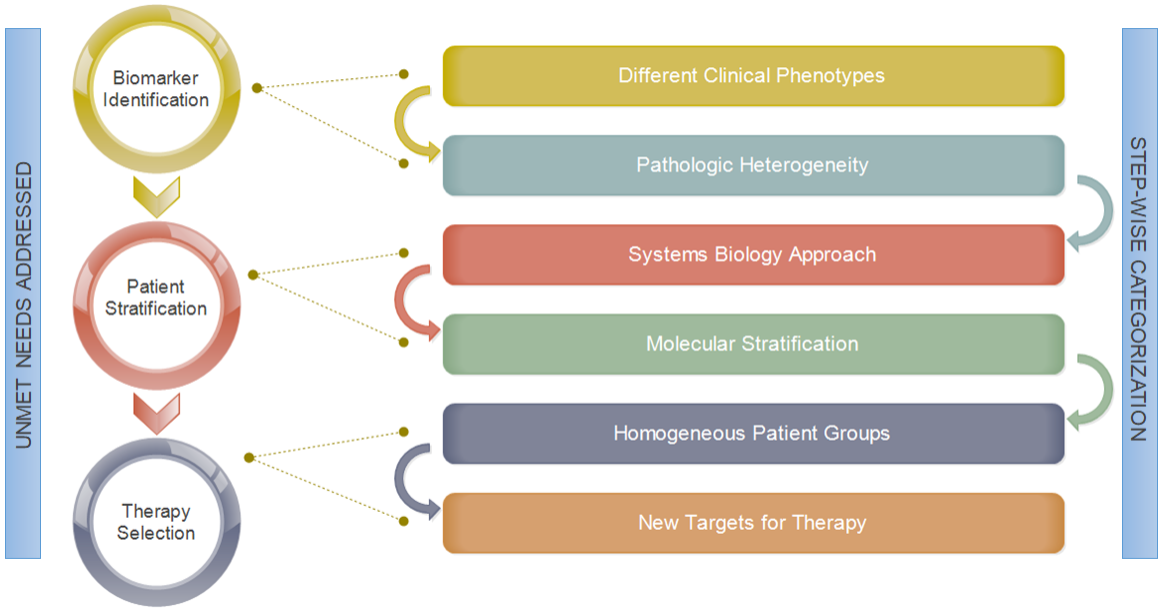

Figure 1. Step-wise categorization (right) in chronic diseases along with the unmet needs that can be addressed (left).

study selection (C) Definition of the target variables and Data Processing, and (D) Co-analysis of the harmonized dataset(s).

A comprehensive description of each step is given below along with their basic aims and the key services and software developed to support harmonization and co-analysis of the results. Tasks (A) and (D) compose the main tasks to be addressed at the initial and final stages of data harmonization. The remaining Steps (B) and (C) refer to the underlying principles that generate the synthesized dataset(s) required for further statistical analysis.

The indicative key steps have been addressed by many case studies in the literature highlighting the main contributions of data harmonization in large patient datasets and health records. Moreover, legal and ethical issues should be considered and addressed during the process of data integration for each harmonized cohort.

A. Definition of the objectives and research questions

A research protocol that reflects the potential and the limitations should be developed prior to the data harmonization process. This protocol defines the research questions to be addressed and the objectives that ensure the viability and reproducibility of the process. Issues to be considered in this initial step are study-specific [11] as well as common tasks and are related to: (i) data access procedures and usage, (ii) data infrastructure to be implemented, (iii) study designs and type of information required, (iv) type of data and how they are collected and (v) quality of the data. Moreover, questions related to the proprietary rights of the synthesized datasets and the specific responsibilities of the stakeholders involved in the harmonization process, must be addressed due to the number of the participating individual studies. Based on the research protocol, information regarding the scientific, methodological and administrative tasks is gathered. In addition, ethical and legal concerns are defined towards the harmonization procedure.

Recently, established legal frameworks introduced several requirements on data management and sharing to relevant stakeholders [12]. Those requirements are suggested to be addressed through the development of an adequate data management plan which includes information on: (i) the research data handling, (ii) the type of data to be collected, processed and/or generated, (iii) the methodology and standards to be applied, (iv) the data sharing, curation and preservation [13].

B. Data discovery and study selection

The second step prior to the harmonization workflow is related to: (i) the accumulation of information and (ii) the selection of the participating studies. Gathering information for the high level data concepts and the study designs ensure the appropriate knowledge of each individual study. Procedures for better understanding the legal and ethical issues and the comprehensive characterization of the studies through research protocols, questionnaires and operating procedures, are essential to support study selection and data harmonization.

The selection of the participating studies is based on rigorous criteria to ensure consistency and compatibility among them. Hence, eligible studies involved in a harmonization process must be comparable and share common characteristics to address the research questions and objectives. Several tools and services have been developed for data description, presentation and discovery by the Biobank Standardization and Harmonization for Research Excellence in the European Union (BioShare) project [1] and the Observational Health Data Sciences and Informatics (OHDSI) initiative [6].

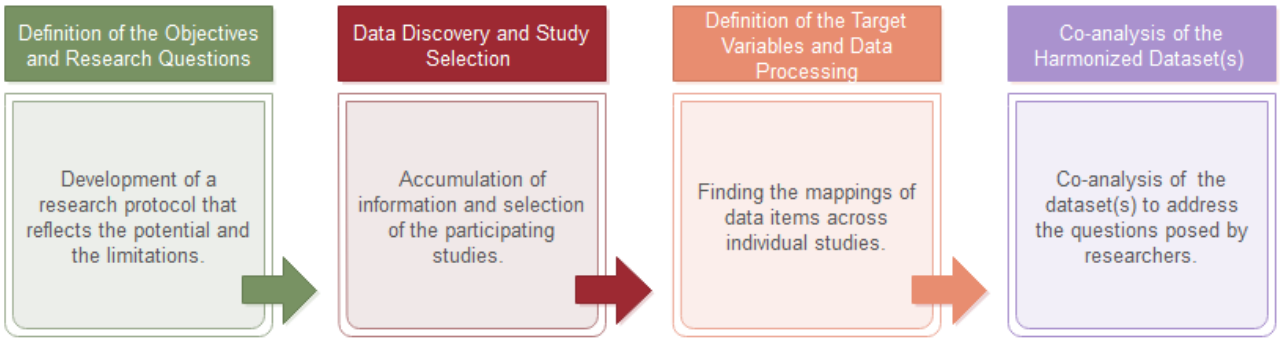

Figure 2. Key steps towards data harmonization.

Café Variome [14] provides a data discovery platform in order to make content available to data scientists and researchers. Users search for data records of desired studies and the list of matching records is exported. Three categories are available for accessing the data, i.e. openAccess, linkedAccess and restrictedAccess. The tool has been built in PHP using the CodeIgniter Web application framework and the Bootstrap front-end framework. The system enables controlled data sharing based on data discovery. Generally, Café Variome is used for genotype-phenotype data searching and sharing and allows for the discovery of patient health records. In addition, it has the ability to modify any data field or attribute of interest and reports all the matched records to the user.

Due to the huge amount of data generated from omics technologies, it is difficult to synthesize and explore such information. OmicsConnect [1] is a software tool that enables viewing and sharing of omics data. It has been designed for presentation, access and mining of complex genomics data.

The Mica software application [1] is a Java-based client-server application developed to create web portals for specific studies. It is used to: (i) create a website for an individual study or a consortium, (ii) create a study catalogue or registry according to the data collected and (iii) enable efficient data access management. It has been designed for database owners and researchers so as to give access to data and spread information about studies as well as search and query data, respectively. Nowadays, Mica is used by several cohort studies and allows to report key characteristics of the participating biobanks and cohorts.

The Molecular Genetics Information System/Observation Entity Model Extensible (MOLGENIS/Observ-EMX) [15-17] has also been designed for data search and analysis. It is a portal for data management and permit the integration, discovery and analysis of scientific data. Moreover, it enables data annotation through public databases, flexible data exploration and data access.

During the OHDSI program, the WhiteRabbit tool [6] has been developed aiming to enable researchers scan datasets of interest and extract a summary information on the contents of the data for further integration and standardization. Subsequently, the dataset of interest can be matched to the desired common data model.

C. Definition of the target variables and Data Processing

Data integration and pooling across biobanks and cohorts is the main process when conducting data harmonization. Concerning this, the identification of the target variables and the evaluation of the harmonization potential is vital. In order to define the list of core variables to be generated, researchers should evaluate first if individual datasets could be used so as to generate the target variables. Moreover, the potential for each specific study to select the variables for harmonization should be defined, based on the fact that not all studies can create all the variables. To this end, quality is ensured and the harmonization procedure can be achieved. The evaluation of the harmonization potential is assessed by comparing the desired data elements within biobanks and finding the matching elements. Furthermore, finding the mappings of data items across individual studies depends on the research questions and objectives defined prior to data harmonization process.

Towards this, Semantic Web technologies [18-20] provide a rigorous solution to automatically integrate disparate information sources and databases. These technologies and principles have been created by the partners of the World Wide Web Consortium (W3C) [21], so as to enable the expression of both data and rules for reasoning about the data. Semantic Web technologies are entitled by the sort of data found in databases and allow reasoning over data through rules from any existing knowledge-representation system to be exported into the Web. Moreover, the vision of linked data on the Web, enabled developers to create data stores, build vocabularies, and write rules for handling data. Modeling heterogeneous information between web resources on the Semantic Web, is empowered by technologies such as: (i) the Resource Description Framework (RDF) [22, 23], (ii) the SPARQL Query Language for RDF [24], (iii) the Web Ontology Language (OWL) [25, 26] and (iv) the Simple Knowledge Organization System (SKOS) [27].

When conducting data harmonization, researchers should be concerned with the fact that different cohorts may use different identifiers for the same data element of same meaning. When comparing or combining information across databases, features that represent common concepts should be discovered. Based on that knowledge, ontologies or vocabularies can be applied for data integration [28]. Ontologies define the terms and the relationships among them

used to describe a specific area of concern. They can help in data integration, when heterogeneity exists on the identifiers used within different cohorts or when new relationships among data elements can be discovered. Besides that, ontologies can be used to organize knowledge regarding the power of linked data on the Web. Different techniques have been proposed so as to specify and determine a common format for the different forms of ontologies (vocabularies) [22, 25, 27, 29]. In addition, Open Biological and Biomedical Ontology (OBO) Foundry [30] which is a collective of ontology developers, incorporate a family of interoperable ontologies that have been developed and are both logically well-formed and scientifically accurate. A reference model is usually co-designed with the assistance of the clinicians so as to meet the minimum requirements of a disease's domain knowledge. This model describes all the terminologies, the types of variables and the related measurement units that a clinical dataset should fulfill. Each heterogeneous clinical dataset is then transformed into this reference model. One possible approach to make this transformation possible is to recruit semantic interlinking mechanisms by (i) constructing an ontology from the heterogeneous clinical dataset and (ii) mapping this ontology to the main ontology, i.e., the one created by the reference model. For example, if the reference model defines the measurement unit of the variable named 'blood pressure' as 'normal' or 'abnormal', and a candidate clinical dataset follows another coding system, e.g., 'normal', 'high', 'very high', then the latter shall be appropriately transformed to fulfill the 'normal', 'abnormal' coding system (e.g., assign the 'high', 'very high' values to 'abnormal'). This example formulates one of the key issues that harmonization wishes to achieve.

During the last decade, the huge amount of information available from different databases and the variations on the concepts and relationships of their data elements, led to the semantic heterogeneity problem, which is the different terminologies used for describing equivalent concepts. Consequently, in order to overcome the problem related to the management of heterogeneous information sources, ontology matching techniques have been proposed [31, 32]. Matching ontologies enables the identification of correspondences among semantically related terms of ontologies, based on their meaning. The applications of ontology matching are characterized of models with heterogeneity, such as the database schemas of specific individual studies. Several matching systems have been proposed in the literature. These matching systems are based on the kind of data that are utilized, i.e. (i) strings (terminological or lexical systems), (ii) structure (structural systems), (iii) instances (extensional systems), and (iv) models (semantics methods) [8, 31, 33-47]. Fig. 3 illustrates the semantic matching process along with an indicative example [31]. Different ontologies can be matched according to the related correspondences of their entities. More specifically, Fig. 3(A) demonstrates the general strategy for semantic matching, whereas Fig. 3(B) depicts an indicative example of semantic interlinking among the correspondences of two different ontologies. Ontology 1 corresponds to a treatment schema and Ontology 2 to disease diagnosis.

A number of integrated tools and methods have been developed during the BioSHaRE project [1, 48] related to research data harmonization. Most of these tools have been deployed for data pooling and for identifying the mappings of variables across different datasets. More specifically, the BiobankConnect software [9] enables the connection of data across biobanks for pooled analysis by employing ontology and lexical matching. Data mapping is achieved and subsequently harmonization of biobanks data dictionaries can be fostered. In a similar manner, the DataSchema and Harmonization Platform for Epidemiological Research (DataSHaPER) approach [4, 5] provides an integrated method for harmonization of large population studies. The development of the Generic DataSchema tool enables the generation of a list of core variables among participating studies through a template that defines common format measures. Opal software [1] allows users to perform data harmonization and data integration across studies. Furthermore, the processing and implementation of algorithms through Opal, support the transformation of study specific variables into common formats. Another framework, namely the System for Ontology-based Re-coding and Technical Annotation (SORTA) [49], has been developed for the annotation of biomedical and phenotype data. SORTA overcomes the problem of semantic heterogeneity and matches original data values to a target scheme. Specifically, SORTA solves the obstacles of heterogeneity of data contents by mapping text descriptions and/or coded data values with standard codes such as ontologies or local terminologies. Suppose that we need to retrieve the most relevant matches of a list of terms such as: (i) protruding eyeball, (ii) hearing impairment, and (iii) hyperextensibility at elbow joint, against the Human Phenotype Ontology. SORTA, will retrieve the most relevant concepts for data values from the established knowledge base and users can pick the correct matches from the list of concepts. In our example, SORTA can select the ontology term for protruding eyeball, i.e proptosis, and a number of synonyms will be retrieved, i.e prominent eyes and prominent globes.

Furthermore, the Molecular Genetics Information System (MOLGENIS/connect) system [16] also addresses the challenge of linking and harmonizing disparate archives of heterogeneous phenotype data. Molgenis/Connect is an additional system to BiobankConnect which matches data elements from biobanks to target variables/schemes. In general, it is a semi-automatic system that effectively defines the transformation algorithms in order to produce integrated datasets. Specifically, the data elements and the DataSchema can be viewed by the users. A semantic similarity facility is provided to find the best matching elements according to the DataSchema. Towards this approach, other tools have been also developed, such as the RD-Connect software [50]. However the main objective of this tool is to simply aggregate data from different sources, creating a platform of linked data which are further processed. The Sample availability (SAIL) method [51] has been created for harmonizing and integrating biomedical and clinical data across cohorts in order to further support research. Within the OHDSI interdisciplinary collaborative [6], resources to convert a wide variety of health

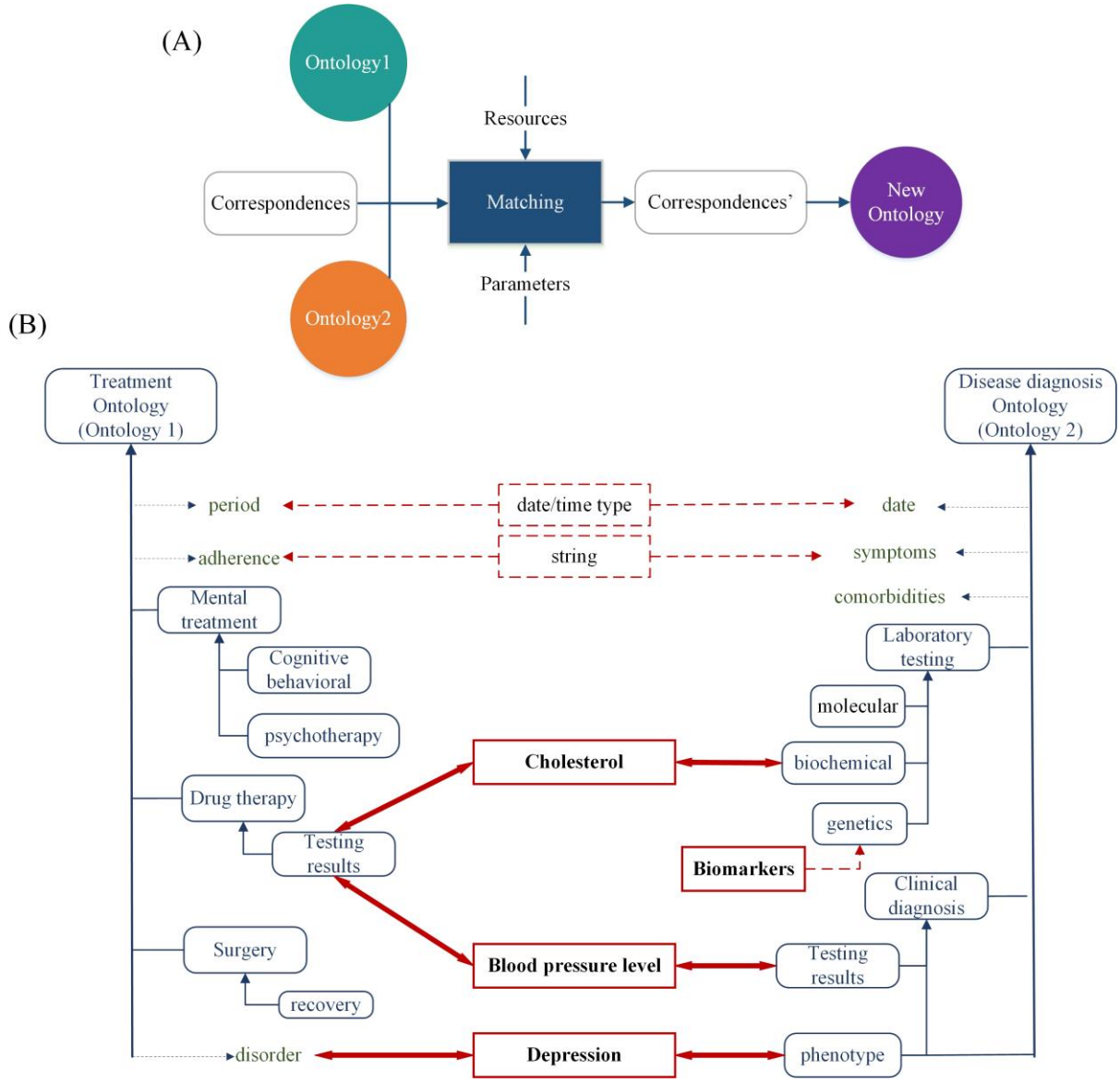

Figure 3. Illustration of the semantic matching procedure. (A) The matching process can be seen as a function which, from a pair of ontologies to match, i.e. Ontology1 and Ontology2, an input alignment with Correspondences between the ontologies, a set of parameters and a set of resources, returns a new set of Correspondences between these ontologies, resulting in a new Ontology [31]. (B) Alignment between the Treatment and Disease diagnosis ontologies. Correspondences are shown as red arrows that connect an entity from one ontology with an entity from another ontology. The entities are annotated with the relation that is expressed by the correspondence.

data into a common data model are provided. Specifically, ATHENA [52] and Usagi [53] are tools that have been built to help with the standardization of vocabularies and the vocabulary mapping, respectively, before the creation of the common data schema.

SNOMED CT [54] is the most comprehensive, multilingual clinical healthcare terminology in the world. It is a resource with comprehensive, scientifically validated clinical content and enables the consistent representation of clinical content in electronic health records. Towards the development of the ATHENA software application and its application design, a stream of concepts (i.e. attributes of existing concepts, new concepts, missing concepts etc.) are extracted from SNOMED international authoritative source in combination with SNOMED UK. Table II presents the integrative tools and services along with their technical descriptions that have been developed and used widely from research groups towards data harmonization process and data integration. Research data integration strategies that foster the sharing of information from individual studies are supported from contemporary Semantic Web technologies [22].

A. Co-analysis of the harmonized dataset(s)

Once the harmonized dataset has been created, according to the methodologies and principles mentioned above, further co-analysis of the results should be performed. Therefore, the quality of the final dataset generated should be estimated and assessed. Specific quality procedures could verify loss of precision and/or key characteristics which are source of bias. Furthermore, defining and implementing a sustainable

Table II. Integrative tools with matching techniques and ontologies that have been developed and used widely from research groups towards data harmonization process and integration cohorts.

Tools and Services	Description	Contribution towards Harmonization
BiobankConnect [9]	Software to semi-automatically match desired data elements across biobanks to available elements using ontological and lexical indexing.	Assessing the harmonization potential through: (i) annotation of the desired elements with ontology terms, (ii) expansion of the query with synonyms and sub-class information, (iii) search of available elements for the expanded terms and (iv) sorting of the matches based on matching scores.
DataSHaPER [4, 5]	An integrated approach that generates harmonized variables from questionnaires and physical measures collected in large population studies.	A DataSchema (i.e. a hierarchical structure) is composed of the core variables derived from interviews, questionnaires etc. and provide the information to be harmonized in a specific scientific context. The DataSHaPER harmonization platform enables then the use of a DataSchema in order to integrate data across different studies.
Opal [1]	A software application which enables the management of study data. It includes a feature enabling data harmonization towards studies.	Supporting the development and implementation of processing algorithms which are required to transform the available data into a common harmonized format.
SORTA [49]	A computational approach to rapidly encode original data to widely accepted ontology systems like SNOMED CT [54], etc.	The desired coding system or ontology is indexed for matching searches. The selected dataset is also matched automatically with the index ontology. A list with all the relevant matches is exported.
MOLGENIS/connect [16, 17]	A semi-automatic system which enables the identification, matching and pooling of data elements from different studies.	Extracting the most relevant features from thousands of existing candidates in the available datasets. Ontology-based queries are used to avoid variations in terminology. These attributes are then transformed to common target DataSchema through algorithms.
Query expansion with a medical ontology [40]	A system to expand user's query with medical terms. A medical ontology is used to improve a multimodal retrieval system by expanding the user's query with MeSH [Relationships in medical subject headings] terms selected by the user.	Expanding the queries by adding medical information based on MeSH terms. A set of terms are identified and constitute the bag terms for query expansion
GoPubMed [41]	A web server which allows users to explore PubMed search results through a hierarchically structured vocabulary for molecular biology.	Providing an overview of the literature abstracts by categorizing them according to (GO). General ontology terms are depicted related to the original query while the server enables users to verify its classification.
WordNet [42]	Employment of additional resources, such as the lexical databases, to increase the amount of information which text categorization systems make use of.	Given a set of documents and a set of categories, the categorization systems are able to decide whether any document belongs to any category or not.
FOAM [43]	A framework for ontology alignment and mapping results.	Alignment methods are used and mapped onto a generic alignment process. Similarity assessment and aggregation are also performed for given features indicating the similarities
S-Match [44]	An open source semantic matching framework that tackles the semantic interoperability problem by transforming several data structures into ontologies.	Implementations of basic semantic matching algorithms. This open source semantic matching framework enables the use of ontologies while revealing how they can be used to hold many knowledge organization systems.
SUiS [45]	An information system which utilizes a domain specific ontology for query expansion and translation, for answer generation, and for document analysis domain specific ontology used for query expansion.	Use of two kinds of knowledge sources: (i) a domain specific ontology and (ii) a domain independent date and named entity recognition modules.
ATHENA [6, 52]	Automated Terminology Harmonization, Extraction and Normalization for Analytics.	Design and develop the system for automated or semi-automated vocabulary upload. Process identifies source vocabulary and alter it to standard structure.
Usagi [6, 53]	A software tool created by the OHDSI [6] team and used for mapping codes from a source system into standard terminologies.	Source codes that needs mapping are loaded into the software and a similarity approach is used to connect source codes to Vocabulary concepts.
OBIB [47]	A newly created ontology for biobanking based on the merging of pre-existing ontologies.	Conversion of pre-existing ontologies is applied while the separation of terms defined in the ontologies and other external sources is also conducted. The merging of the overlapping terms identified in the pre-existing ontologies is performed, subsequently.
SAIL [51]	A computational method for addressing the issues of retrospective data harmonization and querying data across biobanks.	Data harmonization within the SAIL method consists of the: (i) creation of a harmonized vocabulary for features of interest, (ii) mapping the harmonized vocabulary to the original biobank variables and (iii) integration of information for each sample.

infrastructure in order to maintain the results of the harmonization process is of great importance. To this end, the last step of the process when all the variables are harmonized, is to co-analyze the dataset(s) aiming to address the questions posed by researchers. There are three types of analysis to co-analyze data, namely (i) pooled analysis, (ii) summary data

meta-analysis and (iii) federated analysis [7]. In pooled analysis, data are integrated in a central location and then analyzed. Since data is centrally stored, flexibility is achieved for further statistical analysis. Nonetheless, important legal and ethical issues rise when pooling data. In the second type, study-specific data analysis is done locally and is followed by a meta-analysis that combines the study-level estimates. Few ethics and data access requirements exist, while this type of analysis is not so flexible because it is limited to summary statistics produced by each study. On the contrary, in federated analysis, the selected studies maintain the complete control of their datasets and the analysis is performed centrally. However, each individual dataset remains on local servers. The DataSHIELD (Data aggregation through anonymous summary-statistics from harmonized individual level databases) method provides powerful tools and functions that permit the federated statistical analyses of pooled datasets of several collaborating studies [55, 56]. It enables a fully efficient integrated analysis of biomedical data even if ethical and/or legal considerations do not allow the spread of individual-level data to third parties. More specifically, Fig. 4 illustrates the basic IT infrastructure underlying the DataSHIELD method, where individual datasets remain on data computers locally, whereas, an additional computer is identified as the analysis computer, centrally. Additionally, the ESPRESSO (Estimating Sample-size and Power in R by Exploring Simulated Study Outcomes) tool [57, 58] enables researchers to define the statistical power concerning a given set or target sample size. It also allows one to calculate the sample size required to achieve specific statistical results. With ESPRESSO and its functions, the assessment of errors in power calculation under various biomedical scenarios can also be achieved. Fig. 5 presents an overview of the ESPESSO algorithm [57].

CIRCE (cohort definition and syntax compiler tool), ACHILLES (Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems) and CALYPSO (Criteria Assessment Logic for Your Population Study in Observational data) are large-scale analytic tools that support the cohort definition, the patient record profiling and the study feasibility assessment, respectively, once patient-level datasets have been integrated into a common data format [6]. The characterization of the extracted dataset through a comprehensive description and the validation of the data quality can be achieved. Moreover, evaluation of the impact of the study is enabled through the CALYPSO tool. In addition, open-source large scale analytic tools, through R [59] and scalable to big data have been developed by [6] for an end-to-end analysis from common data formats through evidence.

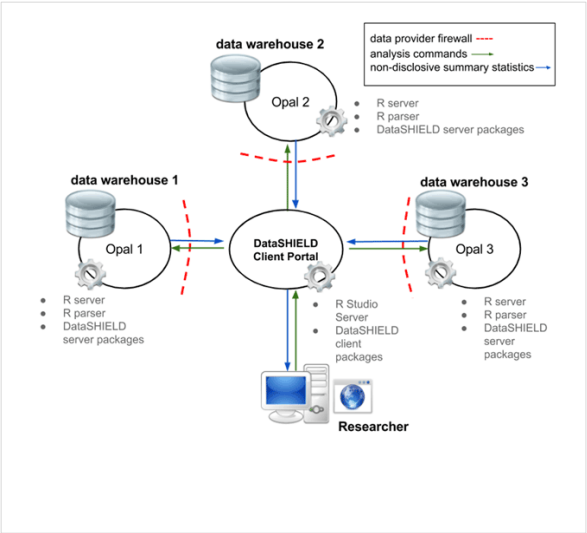

Figure 4. The basic IT infrastructure underlying the DataSHIELD distributed approach (adapted from [56]).

IV. LEGAL AND ETHICAL ISSUES

A. *Data Sharing and Data Protection*

One of the major issues to consider when sharing and harmonizing data from different cohorts is the protection of ethical and legal aspects of the shared and harmonized data. Towards this several different aspects and methodologies/policies have been identified by either the clinical centers participating in data sharing or expert legal offices that have been assigned the role of solving the legal constraints to share data between different countries with different data protection laws and regulations. The innate sensitivity of biomedical information has led to a set of principles and rules for safeguarding protection and privacy of personal data in all stages of data manipulation starting from data collection protocols to data analysis infrastructures.

The evolution of big data and their multi-purpose utilization towards new knowledge mining, which is enabled by innovative developments in information technologies (IT), pose new challenges that need to be addressed in the context of informed consent, privacy of data, ownership, and epistemology in assessing big data ethics and objectivity of big data [60]. For instance, a comprehensive essay centering on issues regarding consent obtaining in biobank studies recognized the need for a defensible, sustainable and conceptually coherent consent policy [61]. Adapting freely given, specific and informed consent along with anonymization mechanisms such that: (i) facilitating the process of dynamic re-consent through the use of Information Technology (IT) providing a transparency level between individuals and their data and, (ii) balancing the need for irreversible anonymization and data linkage and continuing data update, are key issues in the protection of individuals with regard to processing of personal data [60-65].

B. *Harmonizing Data Protection Laws*

1) European Union General Data Protection Regulation (GDPR)

The intensified interest on big data sharing, aggregation, linkage and analysis yielded to the forthcoming replacement of the European Union (EU) Data Protection Directive (DPD) 95/46/EC by the General Data Protection Regulation (GDPR). At the heart of the EU DPD and GDPR lie the principles of fairness and lawfulness assuring the openness and legality of the use of personal sensitive data. The GDPR (http://www.eugdpr.org/eugdpr.org.html) mainly overhauls

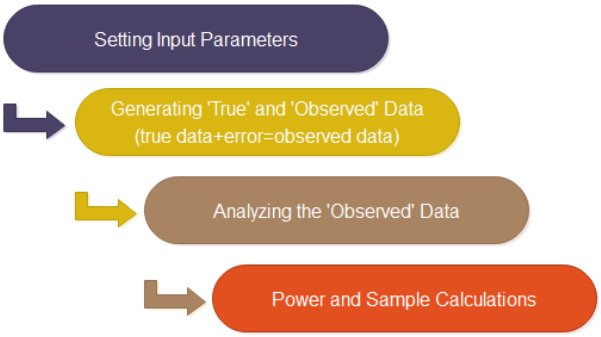

Figure 5. An overview of the ESPRESSO algorithm [57-58].

the EU Directive 95/46/EC with respect to rights of the data subject by introducing or strengthening the rights to: (i) access to data, (ii) rectification and erasure ('right to be forgotten'), (iii) data portability, and (iv) notification for a personal data breach. In addition, the concept of privacy by design calls for the effective implementation of appropriate technical and organizational measures (e.g. pseudo-anonymization) from the design phase of a system in order to ensure non-attribution to an identified or identifiable natural person and meet the requirements of GDPR. The conditions for consent have been also strengthened requesting clarity of the information provided to the individuals.

2) Framework for Responsible Sharing of Genomic and Health-Related Data

The EU BioSHARE Project has developed, under the aegis of the Global Alliance for Genomics and Health, the Framework for Responsible Sharing of Genomic and Health-Related Data [66-69]. This Framework has established a set of foundational principles for responsible sharing of genomic and health-related data: (i) respect individuals, families and communities, (ii) advance research and scientific knowledge, (iii) promote health, wellbeing and the fair distribution of benefits; and (iv) foster trust, integrity and reciprocity. In addition, it has set out ten core elements complementing the interpretation of the aforementioned principles: (i) transparency, (ii) accountability, (iii) engagement, (iv) data quality and security, (v) privacy, data protection and confidentiality, (vi) risk-benefit analysis, (vii) recognition and attribution, (viii) sustainability, (ix) education and training, and (x) accessibility and dissemination.

C. Data Protection Technological Solutions

1) DataSHIELD: An Ethically Robust Solution to Multiple-Site Individual-Level Data Analysis

An exemplar technological solution to preventing re-identification of an individual has been proposed within the DataSHIELD initiative by "taking the analysis to the data, not the data to the analysis", which confines the control researchers retain over the data. In particular, DataSHIELD "enables the co-analysis of individual-level data from multiple studies or sources without physically transferring or sharing the data and without providing any direct access to individual-level data" [70, 71]. The latter feature contributes significantly to properly addressing several ethics-related concerns pertaining to the privacy and confidentiality of the data, the protection of the research participant's rights, and the post-data-sharing concerns. In addition to standard technical and administrative data protection mechanisms, DataSHIELD includes: (i) a systematic three-level validation process of each DataSHIELD command for risks of disclosure, (ii) output restrictions to impede disclosure of possibly identifiable objects, (iii) automatic generation of new subject's identifiers by Opal; original subject's identifiers are stored securely in a distinct database in Opal, (iv) protection mechanisms from potential external network attacks, and (v) encrypted and secured internet communications. It should be noted that ethico-legal and data access approvals as well anonymization of the data constitute necessary preliminary setup steps required for DataSHIELD-based analysis.

V. CASE STUDIES

The presented case studies are mainly comprised of novel worldwide initiatives that involve medical data harmonization, with emphasis in chronic diseases. The ensuing case studies are mainly based on the data harmonization strategy that has been presented in Fig. 2, with differences in the implementation of the harmonization approach according to the context of each study.

In an aging study [72], the authors extracted 26 harmonized variables (across 1768 records) for different attributes, such as socio-demographic, physical and social activity, mental health, etc., based on longitudinal data obtained from older people across two cohorts (one Dutch and one British). According to the harmonization procedure, several study-related values were merged across studies, under certain assumptions, and then new variables were created for representing identical variables, based on appropriately pre-defined research questions. Furthermore, standardization procedures were adopted to facilitate score comparisons across the two cohorts by minimizing age and period effects across the studies. Finally, differences in sampling, study design, measurement instruments, response rates and selective attrition were presented as the fundamental methodological challenges for cross-national studies.

In [73], guidelines for phenotype harmonization are provided by the authors in order to deal with the phenotype heterogeneity across several genome-wide association study (GWAS) consortia using the Gene Environment Association Studies (GENEVA) multi-site GWAS consortium. Their main objective was to create a centralized database for all phenotype and genotype data, therefore enabling cross-study analyses. A context-specific harmonization procedure was adopted. More specifically, a common set of covariates was first identified (a process referred to as uniform covariate selection) across all cohorts with proper adjustments being made wherever possible. Otherwise, stratifications or even exclusions were applied. Finally, the study described methodological challenges on phenotype harmonization that may appear useful for similar studies using GWAS cohorts. Guidelines for stepwise data harmonization procedure, in cancer epidemiology, are also described in [10] where the harmonization procedure took place in the Fred Hutchinson

Cancer Research Center (Fred Hutch) using data from four large consortia.

The potential of harmonizing physical capabilities variables as well as testing for the effect of age and gender interactions across 8 UK cohorts (~ 40000 individuals, under the HALCyon - Healthy Ageing across the Life Course research programme) was assessed in [74], indicating that data harmonization is possible indeed, with (statistically) significant gender differences in two physical measures. The data harmonization procedure is followed by several examples concerning the creation of harmonized variables for a variety of attributes, such as grip strength, walking speed, standing balance, and other related measures. Adjustments and standardizations were further applied on several physical measures to ensure measure-related compatibility across the cohorts. The effects of age and gender were also investigated per physical capability measure. Moreover, new variables were constructed, under certain assumptions, to facilitate harmonization.

In [75], the authors constructed a harmonized database including data from 10107 persons using longitudinal data from five European cohorts on osteoarthritis (OA), based on existing guidelines from the CLESA (Comparison of Longitudinal European Studies on Aging) project [76]. The proposed harmonization procedure was developed under a sequentially-based algorithm by taking into consideration (a) the availability of a variable across at least three out of five cohorts, (b) the measuring concept, and (c) the overlapping of the response categories of that variable. This scheme was repeated for all variables of interest. An extra weighting procedure was further applied after merging the datasets in order to adjust for age, sex and sample size variations across the datasets.

Apart from the context-specific harmonization methods, several attempts have been made towards establishing a generalized harmonization platform. An example of such a platform is presented in [5] along with an illustrative example of the pairing rules developed for one variable, under the task of (successfully) assessing the potential of harmonizing 50 large population-based studies on diabetes. The platform, which is known in the related literature as the DataSHaPER is a dynamically evolving entity comprised by two major platforms; the DataSchema Platform and the Harmonization Platform.

The DataSHaPER method has been also applied in [4] for assessing the potential of harmonization across 53 of the world's largest longitudinal population-based epidemiological studies (6.9 million participants in total) containing a large variety of variables related to different pathologies, such as cancer, stroke, diabetes, etc. A 36% compatibility for creating a completely harmonized dataset was found. This percentage was higher (i.e., 62%) if only the 'essential' variables participated during the harmonization procedure. According to the harmonization procedure, all variables were first categorized based on their importance for broad-based epidemiological studies as 'essential', 'important' and 'useful'. The DataSchema variables were initially identified (148 variables in total; 38 essential, 45 important and 65 useful) and then the potential of generating each DataSchema variable was further evaluated for each study individually, through the development of context-specific sets of pairing rules per variable. These pairing rules were also examined in order to control for possible pairing errors. In a similar study [1, 48], the authors developed a complete stepwise data harmonization approach based on the DataSHaPER and on tools developed by OBiBa (i.e., Mica software for developing web portals for studies and Opal software for data harmonization and integration [1, 48]) for the Healthy Obese Project (HOP) with the purpose of identifying metabolically healthy obese individuals across eight large cohorts (229534 subjects in total). A 73% compatibility for creating a harmonized HOP database was discovered. The potential of generating the DataSchema variables from each individual study was explored by recruiting and testing the appropriateness of any study-specific data related to the DataSchema variables. The latter were selected from each dataset individually, according to appropriately defined research questions, leading to a set of 96 target variables in total. This set was finally employed as a template for data harmonization.

An additional very important initiative in the domain is the European Project entitled Electronic Health Records for Clinical Research (EHR4CR) funded through the Innovative Medicine Initiative of the European Commission [77]. The EHR4CR platform is an open IT platform that unlocks the information stored in Electronic Health Records for improving clinical research while fully respecting patient privacy and ensuring a high level of security. Data harmonization and controlled sharing are central elements of the project. The platform developed enables efficient communication between sponsors and investigators, speeding up clinical trial protocol design and patient recruitment.

Statistical methods such as item response theory, factor analysis, regression models and semiautomatic approaches such as lexical and semantic matching, have been also used for data harmonization. For example, the authors in [78] confirmed the suitability of Item Response Theory (IRT) peculiarly in neuroticism and extraversion phenotypes harmonization by analyzing personality data from 160671 individuals across 23 GPC (Genetics of Personality Consortium) cohorts (6 twin cohorts) from Europe, USA and Australia.

The fact that their approach was able to successfully identify a genetic variant associated with personality reveals an increase in the statistical power of the IRT in item-based behavioral measures harmonization. The proposed harmonization method was based on an un-biased personality score estimation process, which is known in IRT as 'test-linking'. This procedure was applied on each individual cohort and was conducted by IRT models which were appropriately fitted to the items. In addition, the IRT analysis revealed that the estimated neuroticism and extraversion scores were strongly independent of the corresponding inventory and heritable as well. Moreover, the unidimensionality of the items was assessed by plotting the test information curves (TIC) for combinations of inventories as well as for each inventory separately whereas the variance of these items was assessed by a proposed Bayesian method

for quantifying the non-invariance of the items across all cohorts.

Furthermore, the authors in [79] extracted harmonized child personality factors from two cohorts based on a five factor model that was originally developed for harmonizing child personality factors across the two studies. Regression models were developed in order to control for age and sex interactions across samples based on best-fitting model estimations. The prediction model was tested on each sample to assess the structure of the five factor model. Correlations were also computed between the predicted personality factors and similar factors from existing studies. The fact that the strong correlations between the generated harmonized variables and similar constructs from previous studies were strong confirms that the child personality factors were harmonized, thus enabling comparisons across both studies.

In [9], the BiobankConnect software was tested across 6 biobanks under the EU-BioSHaRE Healthy Obese Project (HOP) [1, 48] on an integrated schema of 32 desired data elements across 6 biobanks from the BioSHaRE HOP which were marked as relevant or not for 5 out of 6 biobanks. Out of 41184 matches, 420 were classified as relevant. An average precision of 0.75 was found at rank 1 and recall of 0.74, 0.82, and 0.88 at ranks 10, 20, and 50, respectively. According to the harmonization process, the potential of harmonizing a dataset was explored by means of manually matching the variables of interest to the target variables.

The SMART (Statistical Modeling of Aging and Risk) project [80] aims on studying the clinical phenotypes and risk factors of different neuropathologies, such as Mild Cognitive Impairment (MCI) and dementia as well as their progress and related combinations. It is comprised by 11 cohorts which combine longitudinal data on cognition and aging. The studies comprising the SMART project are well-established allowing for the application of various data mining strategies and statistical analysis procedures on the field of dementia and relative neuropathologies as well. The SMART dataset is already standardized due to the fact that all research centers had similar data templates in general. As a matter of fact, data harmonization was applied only for classifying individuals into two classes; impaired and non-impaired, according to various neurophysiological test instruments. In order for the authors to deal with variations between scores, the latter were first predicted for various demographic factors (e.g., sex, age) using linear modeling. The predicted scores were then subtracted from the original ones and the normalized scores were used for classification. The ensuing results of the harmonization process were preliminary, representing a small portion of the SMART database. Table III highlights indicative case studies involving cohort harmonization in epidemiology which have further revealed important results related to successful harmonization outcomes with an additional high medical impact. These studies include tools that have been already presented in Table II (e.g., the BiobankConnect software, DataSHaPER, Opal), as well as, other similar tools. Moreover, the presented studies include valuable information about their open source data accessibility. For example, the DataSHaPER platform provides access to existing biobanks.

Several EU-funded studies on data harmonization were recently initiated by the European Commission under the PM04: Networking and optimizing the use of population and patient cohorts at EU level, HORIZON 2020 EU Research and Innovation programme. The HarmonicSS project [81] is an ongoing initiative that aims to harmonize regional, national and international longitudinal cohorts of patients diagnosed with primary Sjögrens Syndrome (pSS) by taking into consideration ethical, legal and privacy issues to construct an integrative cloud-based cohort. On the latter, data mining, data governance and visual analytics then will be developed as well as tools for clinical trial patient selection. Another EU initiative is the EUROlinkCAT (Establishing a linked European Cohort of Children with Congenital Anomalies) [82] which aims to enrich the existing EuroCHILD Cohort Network by bringing together pregnancy and child cohorts as well as biobanks to provide a shared data-management platform and harmonization strategies. The LifeBrain project [83] focuses on the integration, harmonization and enrichment of major neuroimaging studies to obtain brain imaging, cognitive and mental health measures of more than 6000 individuals in order to provide novel information regarding the brain deficits and diagnosis of brain disorders and therefore construct preventive and therapeutic strategies. The ESCAPE-NET (European Sudden Cardiac Arrest network: towards Prevention, Education and NEw Treatment) [84] is another ongoing project where European scientific teams have been gathered in order to design SCA (Sudden Cardiac Arrest) prevention and treatment strategies by combining existing European databases.

VI. DISCUSSION AND FUTURE DIRECTIONS

Data harmonization comprises a fundamental procedure prior to any data analysis across longitudinal cohorts and has been adopted by a variety of epidemiological studies, some of which are presented in the sequel. These studies mainly aim to investigate the potential of harmonizing large epidemiological datasets in order to generate a sustainable and robust dataset with induced variable heterogeneity that is capable of increasing the statistical power of the studies. Afterwards, data pooling combined with straightforward data mining methods can be applied on the harmonized dataset to better comprehend the origins and the progress of the disease under examination, thus enabling the design of accurate treatment strategies.

Data harmonization and integrative analysis of synthesized datasets have become increasingly important in the last decade. The invaluable sources of social, environmental, lifestyle factors and genetic interactions and determinants, with reference to the disease onset, progression and classification of different phenotypes, have given rise to data harmonization procedure. Moreover, this approach of data sharing facilitates the integration of the patient cohorts and the clinical picture and outcome during the disease management. Data harmonization can address the unmet needs of chronic diseases including the stratification of patients to distinct

Table III. Summary of the highlights from selected harmonization case studies.

Type of study	Case study	Scope	Harmonization method	Results
Medium scale cohort harmonization studies	**[72]**	Create a harmonized dataset for aging studies across 2 cohorts (1768 records)	Context-specific with appropriately pre-defined research questions and standardizations for dealing with variables heterogeneity among the datasets	A harmonized dataset with 26 harmonized variables
	[75]	Construct a harmonized database on osteoarthritis (OA) across 5 European longitudinal cohorts (10107 persons)	A sequentially-based algorithm using guidelines from the CLESA project [76] as well as adjusting for age, sex and sample size variations	A harmonized database was successfully constructed
	[9]	Harmonize data from the EU-BioSHaRE HOP [1] on an integrated schema of 32 desired data elements across 6 biobanks which were marked as relevant or not for 5 out of 6 biobanks	The BiobankConnect software [9]	Out of 41184 matches, 420 were classified as relevant with an average precision of 0.75 at rank 1 and recall of 0.74, 0.82, and 0.88 at ranks 10, 20, and 50, respectively
	[74]	Investigate the potential of harmonizing physical capabilities values across 8 UK cohorts (~ 40000 individuals)	A context-specific harmonization approach was adopted with further adjustments and standardizations to ensure compatibility as well as testing for gender and age interactions among studies	Data harmonization is possible with (statistically) significant gender differences in two physical measures
	[5]	Develop a retrospective harmonization data approach for the Healthy Obese Project (HOP) with the purpose of identifying metabolically healthy obese individuals across 8 large cohorts (229534 in total)	A complete step-wise harmonization approach is presented based on the DataSHaPER approach [5] as well as on tools developed by OBiBa [1]	A 73% compatibility of creating a harmonized HOP database
	[80]	Study the clinical phenotypes and risk factors of different neuropathologies (e.g., MCI, dementia) across 11 longitudinal cohorts	Simple due to already standardized datasets; Classification of individuals into two classes (impaired and non-impaired) based on neurothapological test instruments and prediction models	Preliminary results of the overall SMART database harmonization process
Large scale cohort harmonization studies	**[78]**	Explore the suitableness of Item Response Theory (IRT) in neuroticism and extraversion phenotypes harmonization across 23 GPC cohorts from Europe, USA and Australia	An un-biased scores estimation process (known as 'test-linking') based on appropriately fitted IRT models	The estimated neuroticism and extraversion scores were strongly independent of the corresponding inventory and heritable
	[4]	Assess the potential of retrospective harmonization on a large variety of epidemiological-related variables (e.g. cancer, stroke, diabetes, etc.) across 53 of the world's largest longitudinal population-based epidemiological studies (~ 6.9 million participants)	The DataSHaPER approach [5]	A 36% compatibility for creating a harmonized database (62% if only the 'essential' variables are taken into consideration)

subgroups, the selection of new targets for therapy as well as the validation or identification of novel biomarkers.

Several key services and tools have been developed aiming to support data harmonization and co-analysis. Achieving harmonization of disparate studies as a rigorous scientific process improves the knowledge management and further the extraction of new scientific knowledge. When conducting data harmonization the definition of the research problem and the data quality, as well as the selection of the studies and the targeted variables constitute the basic steps. In addition, the processing of the final harmonized datasets and the estimation of their quality compose also key steps during the scientific procedure. In order to overcome the semantic heterogeneity obstacle among the different meanings of data elements, semantic technologies and matching techniques have been proposed and developed. Towards this, such approaches can help in data integration, when heterogeneity exists within different datasets and cohorts or when new relationships among data elements can be discovered. Besides that, ontologies and vocabularies can be used to organize knowledge regarding the power of linked data.

In addition, one of the major issues to consider when sharing and harmonizing data from different cohorts is the protection of ethical and legal aspects of the shared data. Towards this, several policies have been identified by either the clinical centers participating in data integration or expert legal offices that have been assigned the role of solving the legal constraints to share heterogeneous sources of information from different countries with different data protection laws and regulations. Concerning the protection laws across Europe when harmonizing different cohorts, the GDPR overhauls the EU Data Protection Directive aiming to

assure the availability and legality of the use of personal sensitive data. Specific rights of the data have been introduced with reference to the concept of privacy and the conditions of personal consent. Furthermore, the Framework for Responsible Sharing of Genomic and Health-Related Data has been developed, during the EU BioSHARE Project, and facilitates the responsible sharing of genomic and health-related patient data. In terms of the technological solutions for data protection, DataSHIELD initiative enables the co-analysis of data from heterogeneous sources, while addressing any ethics-related concern and data access approval.

In most of the described case studies, data harmonization is usually performed in a context-specific manner. More specifically, researchers define questions based on the type of epidemiology under examination. These research questions are combined with pairing rules in order to construct harmonized databases. In fact, the potential of harmonizing each individual dataset is explored by matching each study-related variable with those derived by a pre-defined (common) template based on the pairing rules. This process is a step-wise harmonization approach and is the most preferable way for harmonization, according to the existing literature. Moreover, the semantic interlinking mechanisms are easier to be implemented in this way since it is a study-related approach.

Noteworthy attempts have been recently accomplished towards establishing a semi-automatic data harmonization platform. An example of such an attempt is the DataSHaPER tool which is able to generate a complete harmonized dataset mainly for cancer epidemiology studies. In addition, the tools that were developed under the EU FP7 BioSHaRE project (i.e., the BiobankConnect software) constitute the fundamental basis for data harmonization and have been used in several studies which are described in the current paper. The Opal and Mica software combined with the DataSchema variable template (from the DataSHaPER method) comprise the two main mechanisms used for data harmonization.

Statistical methods and Item Response Theory aspects comprise an alternative strategy for data harmonization. Towards this direction, regression models have been used for predicting measures (scores) in various epidemiological studies based on ‘best-fitting’ models as well as to control for population (e.g., age, sex) differences among the cohorts under examination. Then, the predicted measures are compared with the original ones in order to evaluate the prediction model for data harmonization. Such approaches are less time-consuming than context-specific procedures but they are often complex.

Summarizing, harmonization is a challenging field with crucial methodological challenges yet to be met. These challenges mainly focus on barriers introduced by language differentiation and variables variation across the datasets, as well as by the definition and the context of the research questions. Harmonization is crucial prior to any meta-analysis in order to generate a sustainable and robust dataset with induced variable heterogeneity that will be capable of increasing the statistical power of the studies. However, in order to extend and better understand data harmonization, it is necessary to promote the description of existing harmonization procedures in epidemiology.

Harmonization is a rapidly evolving research field which has recently gained attention since it constitutes an emerging and promising approach for ensuring homogeneity across longitudinal epidemiological studies. Its importance lies on the innovation it offers for heterogeneous data integration. This evidence combined with the recent advances in machine learning, reinforcement learning, data mining strategies and artificial intelligence can lead to the development of new data harmonization strategies yielding higher statistical power and harmonization quality, with minimum loss. The current review can enhance relevant knowledge to such initiatives and thus provide a good impact in the scientific community.

REFERENCES

[1] D. Doiron, P. Burton, Y. Marcon, A. Gaye, B. H. R. Wolffenbuttel, M. Perola, R. P. Stolk, L. Foco, C. Minelli, M. Waldenberger, R. Holle, K. Kvaloy, H. L. Hillege, A. M. Tasse, V. Ferretti, and I. Fortier, "Data harmonization and federated analysis of population-based studies: the BioSHaRE project," *Emerg Themes Epidemiol*, vol. 10, p. 12, Nov 21 2013.

[2] C. Tenopir, S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame, "Data sharing by scientists: practices and perceptions," *PLoS ONE*, vol. 6, p. e21101, 2011.

[3] D. L. Longo and J. M. Drazen, "Data Sharing," *New England Journal of Medicine*, vol. 374, pp. 276-277, 2016.

[4] I. Fortier, D. Doiron, J. Little, V. Ferretti, F. L’Heureux, R. P. Stolk, B. M. Knoppers, T. J. Hudson, and P. R. Burton, "Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies," *International journal of epidemiology*, vol. 40, pp. 1314-1328, 2011.

[5] I. Fortier, P. R. Burton, P. J. Robson, V. Ferretti, J. Little, F. L'Heureux, M. Deschenes, B. M. Knoppers, D. Doiron, J. C. Keers, P. Linksted, J. R. Harris, G. Lachance, C. Boileau, N. L. Pedersen, C. M. Hamilton, K. Hveem, M. J. Borugian, R. P. Gallagher, J. McLaughlin, L. Parker, J. D. Potter, J. Gallacher, R. Kaaks, B. Liu, T. Sprosen, A. Vilain, S. A. Atkinson, A. Rengifo, R. Morton, A. Metspalu, H. E. Wichmann, M. Tremblay, R. L. Chisholm, A. Garcia-Montero, H. Hillege, J. E. Litton, L. J. Palmer, M. Perola, B. H. Wolffenbuttel, L. Peltonen, and T. J. Hudson, "Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies," *Int J Epidemiol*, vol. 39, pp. 1383-93, Oct 2010.

[6] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, and P. R. Rijnbeek, "Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers," *Studies*

in health technology and informatics, vol. 216, p. 574, 2015.

[7] I. Fortier, P. Raina, E. R. Van den Heuvel, L. E. Griffith, C. Craig, M. Saliba, D. Doiron, R. P. Stolk, B. M. Knoppers, and V. Ferretti, "Maelstrom Research guidelines for rigorous retrospective data harmonization," *International journal of epidemiology*, vol. 46, pp. 103-105, 2017.

[8] C. Pardo, F. J. Pino, F. García, M. Piattini, and M. T. Baldassarre, "An ontology for the harmonization of multiple standards and models," *Computer Standards & Interfaces*, vol. 34, pp. 48-59, 2012.

[9] C. Pang, D. Hendriksen, M. Dijkstra, K. J. van der Velde, J. Kuiper, H. L. Hillege, and M. A. Swertz, "BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing," *Journal of the American Medical Informatics Association*, vol. 22, pp. 65-75, 2014.

[10] B. Rolland, S. Reid, D. Stelling, G. Warnick, M. Thornquist, Z. Feng, and J. D. Potter, "Toward rigorous data harmonization in cancer epidemiology research: one approach," *American journal of epidemiology*, vol. 182, pp. 1033-1038, 2015.

[11] P. Dalerba, D. Sahoo, S. Paik, X. Guo, G. Yothers, N. Song, N. Wilcox-Fogel, E. Forgó, P. S. Rajendran, and S. P. Miranda, "CDX2 as a prognostic biomarker in stage II and stage III colon cancer," *New England Journal of Medicine*, vol. 374, pp. 211-222, 2016.

[12] C. Lynch, "Big data: How do your data grow?," *Nature*, vol. 455, pp. 28-29, 2008.

[13] M. Grootveld, "What you need to know to prepare a Data Management Plan (DMP). Training session on writing a Data Management Plan (DMP)," *10442/15536*, p. 00: 26: 23, 2017.

[14] O. Lancaster, T. Beck, D. Atlan, M. Swertz, D. Thangavelu, C. Veal, R. Dalgleish, and A. J. Brookes, "Cafe Variome: General-Purpose Software for Making Genotype–Phenotype Data Discoverable in Restricted or Open Access Contexts," *Human mutation*, vol. 36, pp. 957-964, 2015.

[15] M. A. Swertz, M. Dijkstra, T. Adamusiak, J. K. van der Velde, A. Kanterakis, E. T. Roos, J. Lops, G. A. Thorisson, D. Arends, and G. Byelas, "The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button," *BMC bioinformatics*, vol. 11, p. S12, 2010.

[16] C. Pang, D. van Enckevort, M. de Haan, F. Kelpin, J. Jetten, D. Hendriksen, T. de Boer, B. Charbon, E. Winder, and K. J. van der Velde, "MOLGENIS/connect: a system for semi-automatic integration of heterogeneous phenotype data with applications in biobanks," *Bioinformatics*, vol. 32, pp. 2176-2183, 2016.

[17] M. Swertz, D. van Enckevort, and C. Pang, "MOLGENIS catalogue," *Journal of Clinical Bioinformatics*, vol. 5, p. S8, 2015.

[18] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, pp. 28-37, 2001.

[19] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "The role of ontologies in biological and biomedical research: a functional perspective," *Briefings in bioinformatics*, vol. 16, pp. 1069-1080, 2015.

[20] E. Blomqvist, "The use of Semantic Web technologies for decision support–a survey," *Semantic Web*, vol. 5, pp. 177-201, 2014.

[21] S. Bratt, "Semantic web and other W3C technologies to watch," *Talks at W3C, January*, 2007.

[22] O. Lassila and R. R. Swick, "Resource description framework (RDF) model and syntax specification," 1999.

[23] F. Manola, E. Miller, and B. McBride, "RDF primer," *W3C recommendation*, vol. 10, p. 6, 2004.

[24] E. Prud and A. Seaborne, "SPARQL query language for RDF," 2006.

[25] D. L. McGuinness and F. Van Harmelen, "OWL web ontology language overview," *W3C recommendation*, vol. 10, p. 2004, 2004.

[26] S. Bechhofer, "OWL: Web ontology language," in *Encyclopedia of database systems*, ed: Springer, 2009, pp. 2008-2009.

[27] A. Miles, B. Matthews, M. Wilson, and D. Brickley, "SKOS core: simple knowledge organisation for the web," in *International Conference on Dublin Core and Metadata Applications*, 2005, pp. 3-10.

[28] M. Brochhausen, A. D. Spear, C. Cocos, G. Weiler, L. Martín, A. Anguita, H. Stenzhorn, E. Daskalaki, F. Schera, and U. Schwarz, "The ACGT Master Ontology and its applications–Towards an ontology-driven cancer research and management system," *Journal of biomedical informatics*, vol. 44, pp. 8-25, 2011.

[29] M. Kifer, "Rule Interchange Format: The Framework," *RR*, vol. 8, pp. 1-11, 2008.

[30] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, and C. J. Mungall, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature biotechnology*, vol. 25, p. 1251, 2007.

[31] J. Euzenat and P. Shvaiko, *Ontology matching* vol. 18: Springer, 2007.

[32] L. Otero-Cerdeira, F. J. Rodríguez-Martínez, and A. Gómez-Rodríguez, "Ontology matching: A literature review," *Expert Systems with Applications*, vol. 42, pp. 949-971, 2015.

[33] P. Lambrix and H. Tan, "SAMBO—a system for aligning and merging biomedical ontologies," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 4, pp. 196-206, 2006.

[34] W. Hu, Y. Qu, and G. Cheng, "Matching large ontologies: A divide-and-conquer approach," *Data & Knowledge Engineering*, vol. 67, pp. 140-160, 2008.

[35] M. Nagy, M. Vargas-Vera, and E. Motta, "DSSim: managing uncertainty on the semantic Web," in

Proceedings of the 2nd International Conference on Ontology Matching-Volume 304, 2007, pp. 160-169.

[36] J. Li, J. Tang, Y. Li, and Q. Luo, "RiMOM: A dynamic multistrategy ontology alignment framework," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 1218-1232, 2009.

[37] Y. R. Jean-Mary, E. P. Shironoshita, and M. R. Kabuka, "Ontology matching with semantic verification," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, pp. 235-251, 2009.

[38] M. H. Seddiqui and M. Aono, "An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, pp. 344-356, 2009.

[39] I. F. Cruz, C. Stroe, F. Caimi, A. Fabiani, C. Pesquita, F. M. Couto, and M. Palmonari, "Using AgreementMaker to align ontologies for OAEI 2011," in *Proceedings of the 6th International Conference on Ontology Matching-Volume 814*, 2011, pp. 114-121.

[40] M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. Ureña-López, "Query expansion with a medical ontology to improve a multimodal information retrieval system," *Computers in biology and medicine*, vol. 39, pp. 396-403, 2009.

[41] A. Doms and M. Schroeder, "GoPubMed: exploring PubMed with the gene ontology," *Nucleic acids research*, vol. 33, pp. W783-W786, 2005.

[42] M. Rodriguez, J. Hidalgo, and B. Agudo, "Using WordNet to complement training information in text categorization," in *Proceedings of 2nd International Conference on Recent Advances in Natural Language Processing II: Selected Papers from RANLP*, 2000, pp. 353-364.

[43] M. Ehrig and Y. Sure, "Foam–framework for ontology alignment and mapping results of the ontology alignment evaluation initiative," in *Integrating Ontologies Workshop Proceedings*, 2005.

[44] F. Giunchiglia, A. Autayeu, and J. Pane, "S-Match: an open source framework for matching lightweight ontologies," *Semantic Web*, vol. 3, pp. 307-317, 2012.

[45] K. Nilsson, H. Hjelm, and H. Oxhammar, "SUiS–cross-language ontology-driven information retrieval in a restricted domain," in *Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)*, 2006, pp. 139-145.

[46] J. Kuiper, I. J. Marshall, B. C. Wallace, and M. A. Swertz, "Spá: A web-based viewer for text mining in evidence based medicine," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2014, pp. 452-455.

[47] M. Brochhausen, J. Zheng, D. Birtwell, H. Williams, A. M. Masci, H. J. Ellis, and C. J. Stoeckert, "OBIB-a novel ontology for biobanking," *Journal of biomedical semantics*, vol. 7, p. 23, 2016.

[48] "Harmonisation for Research Excellence in the European Union (BioSHaRE-EU)," *URL: http://www.p3g.org/bioshare*, 2015.

[49] C. Pang, A. Sollie, A. Sijtsma, D. Hendriksen, B. Charbon, M. de Haan, T. de Boer, F. Kelpin, J. Jetten, and J. K. van der Velde, "SORTA: a system for ontology-based re-coding and technical annotation of biomedical phenotype data," *Database*, vol. 2015, 2015.

[50] R. Thompson, L. Johnston, D. Taruscio, L. Monaco, C. Béroud, I. G. Gut, M. G. Hansson, A. Peter-Bram, G. P. Patrinos, and H. Dawkins, "RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research," *Journal of general internal medicine*, vol. 29, pp. 780-787, 2014.

[51] M. Gostev, J. Fernandez-Banet, J. Rung, J. Dietrich, I. Prokopenko, S. Ripatti, M. I. McCarthy, A. Brazma, and M. Krestyaninova, "SAIL—a software system for sample and phenotype availability across biobanks and cohorts," *Bioinformatics*, vol. 27, pp. 589-591, 2010.

[52] *OHDSI ATHENA standardized vocabularies.* Available:http://www.ohdsi.org/analytic-tools/athena-standardized-vocabularies/

[53] *Usagi, an application to help create mappings between coding systems and the Vocabulary standard concepts.* Available: http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:usagi

[54] *SNOMED CT International Edition*. Available: https://www.nlm.nih.gov/healthit/snomedct/international.html

[55] M. Wolfson, S. E. Wallace, N. Masca, G. Rowe, N. A. Sheehan, V. Ferretti, P. LaFlamme, M. D. Tobin, J. Macleod, J. Little, I. Fortier, B. M. Knoppers, and P. R. Burton, "DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data," *International Journal of Epidemiology*, vol. 39, pp. 1372-1382, 2010.

[56] R. Wilson, O. Butters, D. Avraam, J. Baker, J. Tedds, A. Turner, M. Murtagh, and P. Burton, "DataSHIELD–new directions and dimensions," *Data Science Journal*, vol. 16, 2017.

[57] A. Gaye, T. W. Burton, and P. R. Burton, "ESPRESSO: taking into account assessment errors on outcome and exposures in power analysis for association studies," *Bioinformatics*, vol. 31, pp. 2691-2696, 2015.

[58] P. R. Burton, A. L. Hansell, I. Fortier, T. A. Manolio, M. J. Khoury, J. Little, and P. Elliott, "Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology," *International journal of epidemiology*, vol. 38, pp. 263-273, 2008.

[59] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier,

Y. Ge, and J. Gentry, "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology*, vol. 5, p. R80, 2004.

[60] B. D. Mittelstadt and L. Floridi, "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts," *Sci Eng Ethics*, vol. 22, pp. 303-41, Apr 2016.

[61] T. Caulfield and B. Murdoch, "Genes, cells, and biobanks: Yes, there's still a consent problem," *PLoS Biol*, vol. 15, p. e2002654, Jul 2017.

[62] M. Mostert, A. L. Bredenoord, M. C. Biesaart, and J. J. van Delden, "Big Data in medical research and EU data protection law: challenges to the consent or anonymise approach," *Eur J Hum Genet*, vol. 24, pp. 956-60, Jul 2016.

[63] M. Capocasa, P. Anagnostou, F. D'Abramo, G. Matteucci, V. Dominici, G. Destro Bisol, and F. Rufo, "Samples and data accessibility in research biobanks: an explorative survey," *PeerJ*, vol. 4, p. e1613, 2016.

[64] S. Bull, N. Roberts, and M. Parker, "Views of Ethical Best Practices in Sharing Individual-Level Data From Medical and Public Health Research: A Systematic Scoping Review," *J Empir Res Hum Res Ethics*, vol. 10, pp. 225-38, Jul 2015.

[65] D. Hallinan and M. Friedewald, "Open consent, biobanking and data protection law: can open consent be 'informed' under the forthcoming data protection regulation?," *Life Sci Soc Policy*, vol. 11, p. 1, 2015.

[66] B. M. Knoppers, "International ethics harmonization and the global alliance for genomics and health," *Genome Medicine*, vol. 6, pp. 13-13, 02/27 2014.

[67] B. M. Knoppers, J. R. Harris, I. Budin-Ljosne, and E. S. Dove, "A human rights approach to an international code of conduct for genomic and clinical data sharing," *Hum Genet*, vol. 133, pp. 895-903, Jul 2014.

[68] B. M. Knoppers, "Framework for responsible sharing of genomic and health-related data," *Hugo J*, vol. 8, p. 3, Dec 2014.

[69] V. Rahimzadeh, S. O. Dyke, and B. M. Knoppers, "An International Framework for Data Sharing: Moving Forward with the Global Alliance for Genomics and Health," *Biopreserv Biobank*, vol. 14, pp. 256-9, Jun 2016.

[70] I. Budin-Ljosne, P. Burton, J. Isaeva, A. Gaye, A. Turner, M. J. Murtagh, S. Wallace, V. Ferretti, and J. R. Harris, "DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis," *Public Health Genomics*, vol. 18, pp. 87-96, 2015.

[71] S. E. Wallace, A. Gaye, O. Shoush, and P. R. Burton, "Protecting personal data in epidemiological research: DataSHIELD and UK law," *Public Health Genomics*, vol. 17, pp. 149-57, 2014.

[72] P. A. Bath, D. Deeg, and J. Poppelaars, "The harmonisation of longitudinal data: a case study using data from cohort studies in The Netherlands and the United Kingdom," *Ageing & Society*, vol. 30, pp. 1419-1437, 2010.

[73] S. N. Bennett, N. Caporaso, A. L. Fitzpatrick, A. Agrawal, K. Barnes, H. A. Boyd, M. C. Cornelis, N. N. Hansel, G. Heiss, and J. A. Heit, "Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience," *Genetic epidemiology*, vol. 35, pp. 159-173, 2011.

[74] R. Cooper, R. Hardy, A. A. Sayer, Y. Ben-Shlomo, K. Birnie, C. Cooper, L. Craig, I. J. Deary, P. Demakakos, and J. Gallacher, "Age and gender differences in physical capability levels from mid-life onwards: the harmonisation and meta-analysis of data from eight UK cohort studies," *PLoS ONE*, vol. 6, p. e27899, 2011.

[75] L. A. Schaap, G. M. Peeters, E. M. Dennison, S. Zambon, T. Nikolaus, M. Sanchez-Martinez, E. Musacchio, N. M. van Schoor, and D. J. Deeg, "European Project on OSteoArthritis (EPOSA): methodological challenges in harmonization of existing data from five European population-based cohorts on aging," *BMC musculoskeletal disorders*, vol. 12, p. 272, 2011.

[76] N. Minicuci, M. Noale, C. Bardage, T. Blumstein, D. J. Deeg, J. Gindin, M. Jylhä, S. Nikula, A. Otero, and N. L. Pedersen, "Cross-national determinants of quality of life from six longitudinal studies on aging: the CLESA project," *Aging clinical and experimental research*, vol. 15, pp. 187-202, 2003.

[77] EHR4CR. *Electronic Health Records for Clinical Research - (EHR4CR)*. Available: http://www.ehr4cr.eu/views/solutions/platform.cfm

[78] S. M. Van den Berg, M. H. De Moor, M. McGue, E. Pettersson, A. Terracciano, K. J. Verweij, N. Amin, J. Derringer, T. Esko, and G. Van Grootheest, "Harmonization of Neuroticism and Extraversion phenotypes across inventories and cohorts in the Genetics of Personality Consortium: an application of Item Response Theory," *Behavior genetics*, vol. 44, pp. 295-313, 2014.

[79] M. L. Kern, S. E. Hampson, L. R. Goldberg, and H. S. Friedman, "Integrating prospective longitudinal data: Modeling personality and health in the Terman Life Cycle and Hawaii Longitudinal Studies," *Developmental psychology*, vol. 50, p. 1390, 2014.

[80] E. L. Abner, F. Schmitt, P. Nelson, W. Lou, L. Wan, R. Gauriglia, H. Dodge, R. Woltjer, L. Yu, and D. Bennett, "The Statistical Modeling of Aging and Risk of Transition Project: Data collection and harmonization across 11 longitudinal cohort studies of aging, cognition, and dementia," *Observational studies*, vol. 1, p. 56, 2015.

[81] HarmonicSS. (2017). *HARMONIzation and integrative analysis of regional, national and international Cohorts on primary Sjögren's Syndrome (pSS) towards improved stratification, treatment and health policy making*. Available: harmonicss.eu/

[82] EUROlinkCAT, "Establishing a linked European Cohort of Children with Congenital Anomalies," 2017.
[83] LifeBrain, "Healthy minds 0-100 years: Optimising the use of European brain imaging cohorts," 2017.
[84] ESCAPE-NET, "European Sudden Cardiac Arrest network: towards Prevention, Education and NEw Treatment," 2017.