applied sciences

MDPI

Review

Overview of Federated Facility to Harmonize, Analyze and Management of Missing Data in Cohorts

Hema Sekhar Reddy Rajula [1,2,*], Veronika Odintsova [3,4], Mirko Manchia [5,6] and Vassilios Fanos [1]

1 Neonatal Intensive Care Unit, Department of Surgical Sciences, AOU and University of Cagliari, 09042 Cagliari, Italy; vafanos@tin.it
2 Ph.D. student Marie Sklodowska-Curie CAPICE Project, Department of Surgical Sciences, University of Cagliari, 09042 Cagliari, Italy
3 Biological Psychology Department, Vrije Universiteit Amsterdam, 1081 BT Amsterdam, The Netherlands; v.v.odintsova@vu.nl
4 VI Kulakov National Medical Research Center for Obstetrics, Gynecology and Perinatology, Moscow 117198, Russia
5 Section of Psychiatry, Department of Medical Science and Public Health, University of Cagliari, 09125 Cagliari, Italy; mirkomanchia@unica.it
6 Department of Pharmacology, Dalhousie University, Halifax, NS B3H 4R2, Canada
* Correspondence: reddy@unica.it

Received: 10 July 2019; Accepted: 27 September 2019; Published: 1 October 2019

Abstract: Cohorts are instrumental for epidemiologically oriented observational studies. Cohort studies usually observe large groups of individuals for a specific period of time to identify the contributing factors to a specific outcome (for instance an illness) and create associations between risk factors and the outcome under study. In collaborative projects, federated data facilities are meta-database systems that are distributed across multiple locations that permit to analyze, combine, or harmonize data from different sources making them suitable for mega- and meta-analyses. The harmonization of data can increase the statistical power of studies through maximization of sample size, allowing for additional refined statistical analyses, which ultimately lead to answer research questions that could not be addressed while using a single study. Indeed, harmonized data can be analyzed through mega-analysis of raw data or fixed effects meta-analysis. Other types of data might be analyzed by e.g., random-effects meta-analyses or Bayesian evidence synthesis. In this article, we describe some methodological aspects related to the construction of a federated facility to optimize analyses of multiple datasets, the impact of missing data, and some methods for handling missing data in cohort studies.

Keywords: harmonization; meta-analysis; missing data; multiple imputations; information technology; remoteness; cohort studies

1. Introduction

Cohort studies are widely used in epidemiology to measure how the exposure to certain factors influences the risk of a specific disease. The role of large cohort studies is increasing with the development of multi-omics approaches and with the search of methods for the translation of omics findings, especially those that are derived from genome-wide association studies (GWAS) in clinical settings [1]. Many research efforts have been made to link vast amounts of phenotypic data across diverse centers. This procedure concerns molecular information, as well as data regarding environmental factors, such as those recorded in and obtained from health-care databases and epidemiological registers [2]. Cohort studies can be prospective (forward-looking) or retrospective (backward-looking).

Large-scale initiatives of cohort studies have been initiated worldwide, such as the NIG Roadmap Epigenomics Project [3] and the 500 Functional Genomics cohort [4] due to the increasing need of integration of data for genomic analysis. An example of the former is the Nurses' Health Study, a large prospective cohort study that revealed several significant associations between lifestyle choices and health by following up hundreds of thousands of women in North America [5]. Similarly, the National Health and Nutrition Examination Survey (http://www.cdc.gov/nchs/nhanes.htm) discovered the association between cholesterol levels and heart disease. Furthermore, another large prospective cohort study, the Framingham Heart Study (https://www.framinghamheartstudy.org), demonstrated the cardiovascular health risks of tobacco smoking. However, the integration of data from multiple cohort studies faces a variety of challenges [6].

Database federation is a method for data integration that offers constant access to a number of various data sources in which middleware can operate, including interactive database management systems [6]. Individual-level data indicate information about participants, being either contributed by the participants themselves in surveys etc., or collected from registers. In fact, individual-level data assembling of large population-based studies across centers in international collaborations faces several difficulties [7]. On the one hand, merging cohort datasets extends the capability of these studies by allowing research questions of mutual interest to be addressed, by enabling the harmonization of standard measures, and by authorizing the investigation of a range of psychosocial, physical, and contextual factors over time. However, on the other hand, data are often collected in different locations and systems, and they are rarely aggregated into larger datasets, an aspect that limits their utility. Additionally, it is essential to accurately address privacy, legal, and ethical issues that are associated with data usage [8].

Meta-analysis is a common approach to pool results from multiple cohorts and it is consistently used, for example, in GWAS [9] epidemiological studies [10], case-control studies [11] and randomized controlled trials [12]. Several studies are pooled to increase the sample size to increase the statistical power to detect true association signals (decreasing type 1 error). Due to the restrictions of participants confidentiality and the refusal by authors to provide anonymized dataset, individual-level data often cannot be pooled among studies, so meta-analytical approaches are typically used to combine summary statistics across studies, and they have been shown to be as powerful as the integration of individual datasets [9].

In cohort studies, retaining participants with several waves of follow-up is challenging. These waves of data collection give researchers an opportunity to get data regarding deviations in the measures of participants' exposure and outcome over time. The duration of follow-up waves of data can range from one to two years up to 20 to 30 years or even with longer post-baseline assessments. Missingness of the data can be related to study designs in which recurrent measures of exposure and outcome over time are needed. Specifically, the candidates might not available or might be too ill to participate, or they might refuse to respond to specific inquiries or could be deceased [13]. Thus, researchers often face missing data, which might introduce bias in the parameter estimations (for instance of risk calculation) as well as the loss of statistical power and accuracy [13]. Although further research is required to understand both the impact of missing data and the development of effective methodologies for their handling, there are approaches (discussed below) that may decrease the detrimental effect of missing information.

There are several ways of storing and integrating data, such as warehouses, federations, data hotels, etc. For example, the Dutch Techcentre for Life sciences (DTL) is an innovative solution for data hotels. DTL keenly supports FAIR (Findable, Accessible, Interoperable, and Reusable) data stewardship of life science data, within its partnership and in close collaboration with its international partners [14]. The principles of FAIR data serve as an international guideline for high quality data stewardship. A federated facility is a meta-database system that is distributed across multiple locations. It allows making the data from different sources comparable and useful for analyses [15]. Its main difference

between federated facility and registries and warehouses is that data management is carried out via a remote distributed request from one federated server (or database manager) to multiple sources.

A federated facility allows for researchers to receive analytical insight from the pooled information of diverse datasets without having to move all the data to the main location, thus reducing the extent of data movement in the distribution of intermediate results, and maximizing the security of the local data in the distributed sources [16,17]. Most of the data are analyzed close to where they are produced in a federated analytical model. To enable collaboration at scale, federated analytics permits the integration of intermediate outcomes of data analytics while the raw data remains in its locked-down site. When the integrated results are pooled and explored, a substantial amount of knowledge is acquired, and researchers managing a single-center database have the ability to compare their results with the findings that were derived by the analyses of federated pooled data.

In this paper, we describe some methodological aspects that are related to the federated facility. Specifically, we review the methods of harmonization and analysis of multi-center datasets, focusing on the impact of missing data (as well as different approaches to deal with them) in cohorts. Thus, firstly, we aim to suggest a few examples of cohort studies and a data collection procedure for a cohort study, and, secondly, to offer approaches of harmonization and integrative data analysis over cohorts. To this end, we present different methods for handling missing data, such as complete case-analysis and multiple imputations. Finally, we offer a perspective on the future directions of this research area.

2. Examples of Cohort Studies and Integration of Cohorts

Cohort studies allow for one to answer different epidemiological questions regarding the association between an exposure factor and a disease, such as whether exposure to smoking is associated with the manifestation of lung cancer. The British Doctors Study, which started in 1951 (and continued until 2001), was a cohort study that comprised both smokers (the exposed group) and non-smokers (the unexposed group) [18]. The study delivered substantial evidence of the association of smoking with the prevalence of lung cancer by 1956. In a cohort study, the groups are selected in terms of many other variables (i.e., general health and economic status), such that the effect of the variable being evaluated, i.e., smoking (independent variable), is the only one that could be associated with lung cancer (dependent variable). In this study, a statistically significant increase in the prevalence of lung cancer in the smoking group when compared to the non-smoking group rejects the null hypothesis of the absence of a relationship between risk factor and outcome.

Another example is the Avon Longitudinal Study of Parents and Children (ALSPAC), a prospective observational study that examines the impacts on health and development across the life course [19]. ALSPAC is renowned for investigating how genetic and environmental factors affect health and growth in parents and children [20]. This study has examined multiple biological, (epi) genetic, psychological, social, and environmental factors that are associated with a series of health, social, and developmental outcomes. Enrollment sought to register pregnant women in the Bristol area in the UK during 1990–92. This was prolonged to comprise additional children that are eligible up to the age of 18 years. In 1990–92, the children from 14,541 pregnancies were enrolled, which increased the number of participants enrolled to include 15,247 pregnancies by the age of 18 years. The follow-up comprised 59 questionnaires (four weeks–18 years of age) and nine clinical assessment visits (7–17 years of age) [19]. Genetic (the DNA of 11,343 children, genome-wide data for 8365 children, complete genome sequencing for 2000 children) and epigenetic (methylation sampling of 1000 children) data were collected during this study [19].

The federated model is more often used in multi-center studies, large national biobanks, such as the UK Biobank [21], and meta-analyses projects combining data from different registries or databases. It requires new methods and systems to handle large data collection and storing.

One example is the Cross Study funded by the National Institutes of Health (NIH). In this project, data are combined from three current longitudinal studies of adolescent development with a specific emphasis on recognizing evolving pathways that are prominent in substance use and disorder [22].

All three studies oversampled offspring who had at least one biological parent affected by alcohol use disorder and comprised a matched sample of healthy control offspring of unaffected parents. The Michigan Longitudinal Study [23] is the first study that has collected a comprehensive dataset in a large sample of 2–5 year olds subjects who were evaluated via four waves of surveys up to early adulthood. The Adolescent and Family Developmental Project [24] is the second study that recruited families of adolescents aged 11–15, with the surveys being distributed well into adulthood. The Alcohol, Health, and Behavior Project [25] is the third study to include intensive assessments of college freshmen, who, up to their thirties, participated in more than six waves of surveys. Collectively, these three studies span the first four decades of life, mapping the phases when early risk factors for later substance outcomes first emerge (childhood), substance use initiation typically occurs (adolescence), top rates of substance use disorders are evident (young adulthood), and deceleration in substance involvement is evident (adulthood). One potential cause might be that conducting such analyses can be an extremely complex and challenging task. Key practical issues that are associated with data acquisition and data management are often exceed by a multitude of difficulties that arise from at times substantial study-to-study differences [22].

Similarly, the European Union-funded ongoing project Childhood and Adolescence Psychopathology: unraveling the complex etiology by a large Interdisciplinary Collaboration in Europe (CAPICE—https://www.capice-project.eu/) [26] is currently working to create a facility for federated analyses. This requires the databases to have a common structure. CAPICE brings together data from eight population-based birth and childhood (twin) cohorts to focus on the causes of individual differences in childhood and adolescent psychopathology and its course. However, different cohorts use a different measure to assess childhood and adolescent mental health. These different instruments assess the same dimensions of child psychopathology, but they phrase questions in different ways and use different response categories. Comparing and combining the results across cohorts is most efficient when a common unit of measurement is used.

Another project, the Biobank Standardisation and Harmonization for Research Excellence in the European Union (BioSHare) study, built the federated facility using the Mica-Opal federated framework aiming at building a cooperative group of researchers and developing tools for data harmonization, database integration, and federated data analyses [7]. New database management systems and web-based networking technologies are at the limelight of providing solutions to federated facility [7]. Furthermore, the GenomeEUtwin is a large-scale biobank-based research project that integrates massive amounts of genotypic and phenotypic data from distinct data sources that are located in specific European countries and Australia [2]. The federated system is a network called TwinNET used to exchange and pool analyses. The system pools data from over 600,000 twin pairs, and genotype information from a part of those with the goal to detect genetic variants related to common diseases. The network architecture of TwinNET consists of the Hub (the integration node) and Spokes (data-providing centers, such as twin registers). Data-providers initiate connections while using virtual private network tunnels that provide security. This approach also allows for the storage and combining of two databases: the genotypic and the phenotypic database, which are often stored in different locations [27]. The development of Genome EUtwin facility started from the integration of the limited number of variables that appear simple and non-controversial and it is intended to include more variables standard for the world twin community. Most of the European twin registries do not have genotypic or phenotypic information from non-twin individuals. But some do, and GenomEUtwin will want to take advantage of those samples. The advantage with this structure consists in the possibility to store completely new variables as soon as they emerge without changing the database structure. By applying the same variable names and value formats to variables in common to all databases, several advantages will be accomplished [27]. Here, we describe the process of building a federated facility, divided into separate steps (see Table 1).

Table 1. Main Steps of Federated Facility Process.

Step	Description
Data collection in cohort studies	Study data is obtained from self-completed paper-based/online questionnaires, biosample analysis, clinical assessments, linkage to administrative records, etc.
Integration on cohorts	Remote access to aggregated data for statistical analysis is provided and data collected in multiple studies is integrated with the use of harmonization data tools (if needed)
Mega-analyses, meta-analyses or integrative data analyses	Statistical tools for analysis of combined data are applied

3. Data Collection Procedure for a Cohort Study

Several sets of data might be collected in the context of a cohort study. These might include clinical, biological, and imaging data. Data from clinical assessments are comprised of physiological, cognitive, structured or semi-structured interviews measures and/or computer-based questionnaires. Genetic, transcriptomic, proteomic, metabolomic, epigenetic, biochemical, and environmental exposure data can be obtained from the analysis of biological samples [28]. Imaging data can be collected as a part of routine clinical assessment (including magnetic resonance imaging, computer tomography scans, dual-energy x-ray absorptiometry, retinal scan, peripheral quantitative computed tomography, and three-dimensional (3D) Face and body shape). Data are obtained through administrative records comprised of maternity and birth records, child health records, electronic health records, primary and secondary health care records, and social network channels. In the presence of applicable data formats, this information might be transferred while using innovative tools that are becoming increasingly available and that are now robust enough to allow for digital continuity. There will be meticulous management and stewardship of the valuable digital resources, to the benefit of the entire academic community [14].

4. Data Integration

Built-in security features of database management systems can limit access to the whole dataset of a federated facility, and security can be increased while using encryption. Some solutions can be applied, such as establishing a common variable format and standard, creating a unique identifier for all individuals in the cohorts, implementing security access to data and integrity constraints in the database management system, and making automated integration algorithms in the core module to synchronize or federate multiple heterogeneous data sources, to facilitate the integration of different datasets among various cohorts.

There are three steps in data integration: (1) extraction of data and harmonization into a common format at a data provider site; (2) transfer of harmonized data to a data-collecting center for checking; and, (3) data load into a common database [2].

Harmonization is a systematic approach, which allows the integration of data collected in multiple studies or multiple sources. Sample size can be increased by pooling data from different cohort studies. Conversely, individual datasets may be comprised of variables that measure the same hypothesis in different ways, which hinders the efficacy of pooled datasets [29]. Variable harmonization can help to handle this problem.

The federation facility might also be created without the need for harmonization (i.e., any cohorts that have some data (e.g., genotyping data) in one place and some data (e.g., phenotypic) in another or have a connection to national registries, etc.). For example, in the Genome of the Netherlands Project (http://www.nlgenome.nl/), nine large Dutch biobanks (~35,000 samples) were imputed with the population-specific reference panel, an approach that led to the identification of a variant within the ABCA6 gene that is associated with cholesterol levels [30].

The potential of harmonization can be evaluated with the studies' questionnaires, data dictionaries, and standard operating techniques. Harmonized datasets available on a server in every single research centers across Europe can be interlocked through a federated system to allow for integration and statistical analysis [7]. To co-analyze harmonized datasets, the Opal [31], Mica software [7], and the Data SHIELD package within the R environment are used to generate a federated infrastructure that allows for investigators to mutually analyze harmonized data while recollecting individual-level records within their corresponding host organizations [7]. The idea is to generate harmonized datasets on local servers in every single host organization that can be securely connected while using encrypted remote connections. Using a strong collaborative association among contributing centers, this approach can lead to effortless collateral analyses using globally harmonized research databases while permitting each study to maintain complete control over individual-level data [7].

Data harmonization is implemented in light of several factors, for instance, detailed or partial variable matching about the question asked/responded, the answer noted (value definition, value level, data type), the rate of measurement, the period of measurement, and missing values [29].

For instance, in the context of the CAPICE project, the important variables are those concerning demographics (i.e., sex, family structure, parental educational attainment, parental employment, socio-economic status (SES), individual's school achievements, mental health measures (both for psychopathology as well as for wellbeing and quality of life) by various raters (mother, father, self-report, teacher)), pregnancy/the perinatal period (i.e., alcohol and substance use during pregnancy, birth weight, parental mental health, breast feeding), general health (i.e., height, weight, physical conditions, medication, life events), family (i.e., divorce, family climate, parenting, parental mental health), and biomarkers (genomics, epigenetics, metabolomics, microbiome data, etc.). All of these pieces of data gathered in children and parents are harmonized while using various procedures.

Variable manipulation is not essential if the query asked/responded and the answer noted in both datasets is the same [29]. If the response verified is not the same, the response is re-categorized/re-organized to improve the comparability of records from both of the datasets. Missing values are generated for each subsequent unmatched variable and are switched by multiple imputations if the same pattern is calculated in both datasets, even if using different methods/scales. A scale that is applied in both datasets is recognized as a reference standard [29]. If the variables are calculated several times and/or in distinct periods, these are harmonized by gestation trimesters data. Lastly, the harmonized datasets are assembled into a single dataset.

5. Meta-analysis and mega-analysis

Researchers are currently analyzing large datasets to clarify the biological underpinnings of diseases that, particularly in complex disorders, remain obscure. However, due to privacy concerns and legal complexities, data hosted in different centers cannot always be directly shared. In practice, data sharing is also hindered by the administrative burden that is associated with the need to transfer huge volumes of data. This situation made researchers to look for an analytical solution within meta-analysis or federated learning paradigms. In the federated setting, a model is fitted without sharing individual data across centers, but only using model parameters. Meta-analysis instead performs statistical testing by combining results from several independent analyses, for instance, by sharing p-values, effect sizes, and/or standard errors across centers [32].

Lu and co-authors have recently proposed two additions to the splitting approach for meta-analysis: splitting in a cohort and splitting cohorts [9]. The first method implies that data for each cohort is divided, monitored by a choice of variables on one subset and calculation of p-values on the other subset, and by meta-analysis across all cohorts [9]. This is a typical addition of the data splitting approach that can be applied to numerous cohorts. The second method comprises splitting cohorts as an alternative. Cohorts are divided into two groups, one group is used for variable selection and the other is used for attaining p-values as well as meta-analysis. This is a more applicable method, since it simplifies the analysis burden for each study and decreases the possibility of errors [9].

As the focus of a meta-analysis is on the creation of summary statistics obtained from several studies, this method is most efficient when the original individual records used in prior analyses are not accessible or no longer collected [22]. The individual-level information can be pooled into a single harmonized dataset upon which mega-analyses are carried out [33]. The increased flexibility in handling confounders at the individual patient level and assessing the impact of missing data are substantial benefits of a mega-analytical method [34]. Mega-analyses have also been endorsed to evade the assumptions of within-study normality and recognize the within-study variances, which are particularly challenging with small samples. In spite of these benefits, mega-analysis requires homogeneous data sets and the creation of a shared centralized database [34].

Meta-analysis has several disadvantages, including the presence of high level of heterogeneity [35], unmeasured confounders [36], limitation by ecological fallacy [37]. In addition, most of the primary studies included in meta-analysis are conducted in developed or western countries [38]. However, there are numerous benefits to directly fitting models straight into the original raw data instead of creating the applicable summary statistics. Current technological developments (such as a superior capacity for data sharing and wide opportunities for electronic data storage and retrieval) have increased the feasibility of retrieving original individual records for secondary analysis. This gives new opportunities for the progress of different approaches to integrate results across studies by using original individual records to overcome some of the inevitable limitations of meta-analysis [22].

Here, we focus on approaches of integrative data analysis within the psychological sciences, as approaches to collecting current data can differ across disciplines.

6. Integrative Data Analysis

Integrative data analysis (IDA) is the statistical analysis of a dataset that contains two or more separate samples that have been combined into one [22]. The characteristics of a sample that allow for considering it as a separate entity can be defined on a case-by-case basis. In specific situations, there may be differences in the design of the studies from which samples were due to recruited participants. For instance, separate samples might be collected in a multi-site employing single-site strategy in which key design characteristics remain constant (e.g., recruitment, procedures, and measurement). However, in other situations, each study is designed in a distinct setting (e.g., distinct hospitals or regions of the country) or across distinct time periods (e.g., as recruitment moves across different birth cohorts or school years). These separate samples are combined for analysis or cohort differences [22].

However, though IDA may be applied to a variety of designs, the emphasis here is unambiguously on the later situation, and namely where numerous samples are drawn from independent current studies and assembled into a dataset for follow up analyses. This was exactly what the authors experienced in the project called Cross Study, in which their attention was on data collected from three independent studies where participants were different from one to another in both hypothetically and methodologically meaningful ways [22]. The investigators were confident that the greatest potential for upcoming applications of IDA in psychological research comes from the combination of data from two or more studies [22].

7. Different Approaches to Dealing with Missing Data in Cohorts

Different types of missing data in phenotypic and genotypic databases can appear in cohorts: these include, but are not limited to, the following: irrelevant non-response, responses of participants that were excluded from the study in the follow up, irrelevant missing structural data (when data are irrelevant in the context), responses of participants such as “do not know” responses, no answer to a specific item, and missing data due to error codes.

There are several approaches for dealing with missing data: complete-case analyses, last observational carried forward (LOCF) method, mean value substitution method, missing indicator method, and multiple imputations (MI).

The complete case-analysis only is comprised of applicants with full data in all waves of data collection, thus possibly decreasing the accuracy of the estimates of the exposure-outcome relations [13]. To be effective, complete case analyses should assume that applicants with missing data can be considered to be a random sample of those that were intended to be observed (generally referred as missing completely at random (MCAR)) [13], or at least that the probability of data being missing does not dependent on the observed value [39]. Further, LOCF is a method of imputing missing data in longitudinal studies with the non-missing value from the previously completed time-point for the same individual, since the imputed values are unrelated to a subject's other measurements. The mean value substitution replaces the missing value with the average value available from the other individual time-points of a longitudinal study [40]. The missing indicator method comprises an additional category for the analysis created for applicants with missing data [41].

Missing data can be handled by means of multiple imputation [42–45]. MI methods are used to address missing data and its assumptions are more flexible than those of complete case analysis [46]. The principle of MI is to substitute missing observations with plausible values multiple times and generate complete data sets [47]. Every single complete data set is individually analyzed and the effects of the analyses are then assembled. MI results are consistent when the missing mechanism satisfies the MCAR and the missing at random (MAR) assumptions [48,49]. Multiple imputation consists of three steps. The first step is (1) to determine which variables are used for imputation. The variables used for imputation should be selected on the basis of the presentation of their missing information as MAR [43], that is, whether or not a score that is missing depend on the missing value [42]. The variables that cause the missingness are unknown to the researcher unless missingness is, to some extent, expected. Practically, variables are selected in a way that expected to be good predictors containing missing values. One can choose the number of variables and which variables to use, but there is no alternate way to assess whether MAR is achieved, and MAR is an assumption. Binary or ordinal variables may be imputed under a normality assumption and then rounded off to discrete values. If a variable is right skewed, it might be modeled on a logarithmic scale and then transformed back to the original scale after imputation [42]. We should impute variables that are functions of other (incomplete) variables. Several data sets consist of transformed variables, sum scores, interaction variables, ratio's, and so on. It can be useful to integrate the transformed variables into the multiple imputation algorithms [50]. The second step is (2) to generate imputed data matrices. One of the tools that can be used is the R package Multiple Imputation by Chained Equations (MICE) [50], which uses an iterative procedure in which each variable is sequentially imputed and restricted on the real and imputed values of the other variables. The third step of the multiple imputation procedure is (3) to analyze each imputed data set as desired and pool the results [51].

8. Discussion and Future Perspective

The current narrative review focuses on the rationale of federated facility, as well as the challenges and solutions developed to when attempting to maximize the advantages that are obtained from the federated facility of cohort studies. Assembling individual-level data can be a useful, particularly when the results of interest are relevant. There are several benefits of federated facility and harmonizing cohorts: integrating harmonized data allows for an increase in sample sizes, improves the generalizability of results [1,7,52], ensures the validity of comparative research, creates opportunities to compare different population groups by filling the gaps in the distribution (different age groups, nationality, ethnicity etc.), facilitates extra proficient secondary use of data, and offers opportunities for collaborative and consortium research.

Data pooling of different cohort studies faces many hurdles, including interoperability, shared access, and ethical issues, when cohorts working under different national regulations are integrated.

It is essential that strong collaboration among different parties exists to effectively implement database federation and data harmonization [7]. A federated framework allows investigators to analyze data safely and remotely (i.e., produce summary statistics, contingency tables, logistic regressions)

facilitating their accessibility and decreasing actual time restrictions without the burden of filing several data access requests at various research centers, thereby saving principal investigators and study managers time and resources [7].

An important aspect of our review is to provide insights into large samples that result from merging the datasets. Meta-analysis, or mega-analysis of studies, might lead to a more robust estimate of the magnitude of the associations ultimately increasingly the generalizability of findings [33]. As progressively thorough computations can be accomplished in a mega-analysis, some researchers reckon that mega-analysis of individual-participant data can be more efficient than meta-analysis of aggregated data [34]. The mega-analytical framework appears to be the more robust methodology due to the relatively high amount of variation detectable among cohorts in multi-center studies.

In cohort studies, several methods are used to deal with missing data in the exposure and outcome analyses. The most common method is to perform a complete case analysis, an approach that might generate biased consequences if the missing data are not followed by the assumptions of missing completely at random (MCAR). The complete-case analysis allows for consistent results only when the missing data probabilities do not depend on the disease and exposure status simultaneously. Nowadays, researchers are using advanced statistical modeling procedures (for example, *MI* and Bayesian) to handle missing data. Combining studies by Bayesian enable us to quantify the relative evidence with respect to multiple hypotheses using the information from multiple cohorts [44]. Missingness is a typical problem in cohort studies and it is likely to introduce substantial bias into the results. We highlighted how the unpredictable recording of missing data in cohort studies, if not dealt with properly, and the ongoing use of inappropriate approaches to handle missing data in the analysis can substantially affect the study findings leading to inaccurate estimates of associations [13]. Increasing the quality of the study design and phenotyping should be a priority to decrease the amount and impact of missing data. Robust and adequate study designs minimize the additional requests on participants and clinicians beyond routine clinical care, an aspect that encourages the implementation of pragmatic trial design [53].

An organization of databases can facilitate the use of innovative exploratory tools based on machine learning and data mining for data analyses due to data-harmonization techniques. In this narrative review, we proposed several approaches of data integration over cohorts, meta-analysis, mega-analysis in a framework for federated system, and various methods to handle missing data. Further developments of these studies will extend the proposed analysis, from multi-center facility to large-scale cohort data, such as in the context of the CAPICE project.

9. Conclusions

In our review, we highlighted the relevance of setting up reliable database management systems and innovative internet-based networking technologies to provide the resources to support collaborative, multi-center studies in a proficient and secure manner. Variable harmonization remains an essential feature for conducting research using several datasets and permits to increase the statistical power of a study capitalizing on sample size, allowing for more advanced statistical analyses, and answering research questions that might not be addressed by a single study. Future research in this area is needed to develop novel methods to handle missing data, which can substantially impact in very large scale analysis.

Author Contributions: Conceptualization, writing—original draft preparation, H.S.R.R.; added part of the content to the draft—review and editing, V.O.; review and editing, M.M.; supervision, V.F.

Funding: This research was funded by European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 721567, CAPICE project - Childhood and Adolescence Psychopathology: unravelling the complex etiology by a large Interdisciplinary Collaboration in Europe under Grant Agreement 721567. The APC was funded by the CAPICE project.

Acknowledgments: We acknowledge the support from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 721567, CAPICE project -

Childhood and Adolescence Psychopathology: unravelling the complex etiology by a large Interdisciplinary Collaboration in Europe under Grant Agreement 721567.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wijmenga, C.; Zhernakova, A. The importance of cohort studies in the post-GWAS era. *Nat. Genet.* **2018**, *50*, 322–328. [CrossRef] [PubMed]
2. Muilu, J.; Peltonen, L.; Litton, J.-E. The federated database – a basis for biobank-based post-genome studies, integrating phenome and genome data from 600 000 twin pairs in Europe. *Eur. J. Hum. Genet.* **2007**, *15*, 718–723. [CrossRef] [PubMed]
3. Bernstein, B.E.; Stamatoyannopoulos, J.A.; Costello, J.F.; Ren, B.; Milosavljevic, A.; Meissner, A.; Kellis, M.; Marra, M.A.; Beaudet, A.L.; Ecker, J.R.; et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **2010**, *28*, 1045–1048. [CrossRef] [PubMed]
4. Bakker, O.B.; Aguirre-Gamboa, R.; Sanna, S.; Oosting, M.; Smeekens, S.P.; Jaeger, M.; Zorro, M.; Võsa, U.; Withoff, S.; Netea-Maier, R.T.; et al. Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses. *Nat. Immunol.* **2018**, *19*, 776–786. [CrossRef]
5. Colditz, G.A.; Philpott, S.E.; Hankinson, S.E. The Impact of the Nurses' Health Study on Population Health: Prevention, Translation, and Control. *Am. J. Public Health* **2016**, *106*, 1540–1545. [CrossRef]
6. Haas, L.M.; Lin, E.T.; Roth, M.A. Data integration through database federation. *IBM Syst. J.* **2010**, *41*, 578–596. [CrossRef]
7. Doiron, D.; Burton, P.; Marcon, Y.; Gaye, A.; Wolffenbuttel, B.H.R.; Perola, M.; Stolk, R.P.; Foco, L.; Minelli, C.; Waldenberger, M.; et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerg. Themes Epidemiol.* **2013**, *10*, 12. [CrossRef]
8. Haynes, C.L.; Cook, G.A.; Jones, M.A. Legal and ethical considerations in processing patient-identifiable data without patient consent: lessons learnt from developing a disease register. *J. Med. Ethics* **2007**, *33*, 302–307. [CrossRef]
9. Lu, C.; O'Connor, G.T.; Dupuis, J.; Kolaczyk, E.D. Meta-analysis for penalized regression methods with multi-cohort Genome-wide Association Studies. *Hum. Hered.* **2016**, *81*, 142. [CrossRef]
10. Lim, G.Y.; Tam, W.W.; Lu, Y.; Ho, C.S.; Zhang, M.W.; Ho, R.C. Prevalence of Depression in the Community from 30 Countries between 1994 and 2014. *Sci. Rep.* **2018**, *8*. [CrossRef]
11. Ng, A.; Tam, W.W.; Zhang, M.W.; Ho, C.S.; Husain, S.F.; McIntyre, R.S.; Ho, R.C. IL-1β, IL-6, TNF- α and CRP in Elderly Patients with Depression or Alzheimer's disease: Systematic Review and Meta-Analysis. *Sci. Rep.* **2018**, *8*, 12050. [CrossRef] [PubMed]
12. Ng, J.H.; Ho, R.C.M.; Cheong, C.S.J.; Ng, A.; Yuen, H.W.; Ngo, R.Y.S. Intratympanic steroids as a salvage treatment for sudden sensorineural hearing loss? A meta-analysis. *Eur. Arch. Oto-Rhino-Laryngology* **2015**, *272*, 2777–2782. [CrossRef] [PubMed]
13. Karahalios, A.; Baglietto, L.; Carlin, J.B.; English, D.R.; Simpson, J.A. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC Med. Res. Methodol.* **2012**, *12*, 96. [CrossRef] [PubMed]
14. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
15. Wade, T.D. Traits and types of health data repositories. *Heal. Inf. Sci. Syst.* **2014**, *2*, 4. [CrossRef] [PubMed]
16. Thomas, G.; Thompson, G.R.; Chung, C.-W.; Barkmeyer, E.; Carter, F.; Templeton, M.; Fox, S.; Hartman, B. Heterogeneous distributed database systems for production use. *ACM Comput. Surv.* **1990**, *22*, 237–266. [CrossRef]
17. Herscovitz, E. Secure virtual private networks: the future of data communications. *Int. J. Netw. Manag.* **1999**, *9*, 213–220. [CrossRef]
18. Di Cicco, M.E.; Ragazzo, V.; Jacinto, T. Mortality in relation to smoking: the British Doctors Study. *Breathe (Sheffield, England)* **2016**, *12*, 275–276. [CrossRef] [PubMed]

19. Boyd, A.; Golding, J.; Macleod, J.; Lawlor, D.A.; Fraser, A.; Henderson, J.; Molloy, L.; Ness, A.; Ring, S.; Davey Smith, G. Cohort Profile: the 'children of the 90s'–the index offspring of the Avon Longitudinal Study of Parents and Children. *Int. J. Epidemiol.* **2013**, *42*, 111–127. [CrossRef]
20. Fraser, A.; Macdonald-Wallis, C.; Tilling, K.; Boyd, A.; Golding, J.; Davey Smith, G.; Henderson, J.; Macleod, J.; Molloy, L.; Ness, A.; et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int. J. Epidemiol.* **2013**, *42*, 97–110. [CrossRef]
21. Manolio, T.A.; Weis, B.K.; Cowie, C.C.; Hoover, R.N.; Hudson, K.; Kramer, B.S.; Berg, C.; Collins, R.; Ewart, W.; Gaziano, J.M.; et al. New models for large prospective studies: is there a better way? *Am. J. Epidemiol.* **2012**, *175*, 859–866. [CrossRef]
22. Curran, P.J.; Hussong, A.M. Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychol. Methods* **2009**, *14*, 81–100. [CrossRef] [PubMed]
23. Zucker, R.A.; Fitzgerald, H.E.; Refior, S.K.; Puttler, L.I.; Pallas, D.M.; Ellis, D.A.; Fitzgerald, H.E.; Refior, S.K.; Puttler, L.I.; Pallas, D.M.; et al. The Clinical and Social Ecology of Childhood for Children of Alcoholics: Description of a Study and Implications for a Differentiated Social Policy. In *Children of Addiction*; Routledge: Ann Arbor, MI, USA, 2002; pp. 125–158.
24. Chassin, L.; Rogosch, F.; Barrera, M. Substance use and symptomatology among adolescent children of alcoholics. *J. Abnorm. Psychol.* **1991**, *100*, 449–463. [CrossRef] [PubMed]
25. Sher, K.J.; Walitzer, K.S.; Wood, P.K.; Brent, E.E. Characteristics of children of alcoholics: putative risk factors, substance use and abuse, and psychopathology. *J. Abnorm. Psychol.* **1991**, *100*, 427–448. [CrossRef] [PubMed]
26. Revolution, T.H.E.; Microbiomics, O.F. Selected Abstracts of the 14 th International Workshop on Neonatology THE REVOLUTION OF MICROBIOMICS NUTRITION, BACTERIA AND PROBIOTICS IN PERINATAL AND PEDIATRIC HEALTH CAGLIARI (ITALY). *J Pediatr Neonat Individual Med.* **2018**, *7*, 1–66.
27. Litton, J.E.; Muilu, J.; Björklund, A.; Leinonen, A.; Pedersen, N.L. Data Modeling and Data Communication in GenomEUtwin. *Twin Res.* **2003**, *6*, 383–390. [CrossRef]
28. Rajula, H.S.R.; Mauri, M.; Fanos, V. Scale-free networks in metabolomics. *Bioinformation* **2018**, *14*, 140–144. [CrossRef]
29. Patel, A.; Patten, S.; Giesbrecht, G.; Williamson, T.; Tough, S.; Dahal, K.A.; Letourneau, N.; Premji, S. Harmonization of data from cohort studies– potential challenges and opportunities. *Int. J. Popul. Data Sci.* **2018**, *3*, 23889.
30. van Leeuwen, E.M.; Karssen, L.C.; Deelen, J.; Isaacs, A.; Medina-Gomez, C.; Mbarek, H.; Kanterakis, A.; Trompet, S.; Postmus, I.; Verweij, N.; et al. Genome of the Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels. *Nat. Commun.* **2015**, *6*, 6065. [CrossRef]
31. Open-source software for biobankers | BBMRI-ERIC: Making New Treatments Possible. Available online: http://www.bbmri-eric.eu/news-events/open-source-software-for-biobankers/ (accessed on 5 June 2019).
32. Silva, S.; Gutman, B.A.; Romero, E.; Thompson, P.M.; Altmann, A.; Lorenzi, M. Federated Learning in Distributed Medical Databases: Meta-Analysis of Large-Scale Subcortical Brain Data. *aeXiv* **2019**, arXiv:1810.08553.
33. Singh, A.; Babyak, M.A.; Brummett, B.H.; Kraus, W.E.; Siegler, I.C.; Hauser, E.R.; Williams, R.B. Developing a synthetic psychosocial stress measure and harmonizing CVD-risk data: a way forward to GxE meta- and mega-analyses. *BMC Res. Notes* **2018**, *11*, 504. [CrossRef]
34. Boedhoe, P.S.W.; Heymans, M.W.; Schmaal, L.; Abe, Y.; Alonso, P.; Ameis, S.H.; Anticevic, A.; Arnold, P.D.; Batistuzzo, M.C.; Benedetti, F.; et al. An Empirical Comparison of Meta- and Mega-Analysis With Data From the ENIGMA Obsessive-Compulsive Disorder Working Group. *Front. Neuroinform.* **2018**, *12*, 102. [CrossRef] [PubMed]
35. Abraham, N.; Buvanaswari, P.; Rathakrishnan, R.; Tran, B.X.; Thu, G.V.; Nguyen, L.H.; Ho, C.S.; Ho, R.C. A Meta-Analysis of the Rates of Suicide Ideation, Attempts and Deaths in People with Epilepsy. *Int. J. Environ. Res. Public Health* **2019**, *16*, 1451. [CrossRef] [PubMed]
36. Low, Z.X.; Yeo, K.A.; Sharma, V.K.; Leung, G.K.; McIntyre, R.S.; Guerrero, A.; Lu, B.; Sin Fai Lam, C.C.; Tran, B.X.; Nguyen, L.H.; et al. Prevalence of Burnout in Medical and Surgical Residents: A Meta-Analysis. *Int. J. Environ. Res. Public Health* **2019**, *16*, 1479. [CrossRef] [PubMed]
37. Foo, S.Q.; Tam, W.W.; Ho, C.S.; Tran, B.X.; Nguyen, L.H.; McIntyre, R.S.; Ho, R.C. Prevalence of Depression among Migrants: A Systematic Review and Meta-Analysis. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1986. [CrossRef] [PubMed]

38. Ng, T.K.S.; Ho, C.S.H.; Tam, W.W.S.; Kua, E.H.; Ho, R.C.-M. Decreased Serum Brain-Derived Neurotrophic Factor (BDNF) Levels in Patients with Alzheimer's Disease (AD): A Systematic Review and Meta-Analysis. *Int. J. Mol. Sci.* **2019**, *20*, 257. [CrossRef] [PubMed]
39. White, I.R.; Carlin, J.B. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat. Med.* **2010**, *29*, 2920–2931. [CrossRef] [PubMed]
40. Molenberghs, G. Analyzing incomplete longitudinal clinical trial data. *Biostatistics* **2004**, *5*, 445–464. [CrossRef] [PubMed]
41. Greenland, S.; Finkle, W.D. A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *Am. J. Epidemiol.* **1995**, *142*, 1255–1264. [CrossRef]
42. Schafer, J.L.; Graham, J.W. Missing data: Our view of the state of the art. *Psychol. Methods* **2002**, *7*, 147–177. [CrossRef] [PubMed]
43. Demirtas, H. Flexible Imputation of Missing Data. *J. Stat. Softw.* **2018**, *85*. [CrossRef]
44. *Multiple Imputation for Nonresponse in Surveys*; Rubin, D.B. (Ed.) Wiley Series in Probability and Statistics; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1987; ISBN 9780470316696.
45. Sterne, J.A.C.; White, I.R.; Carlin, J.B.; Spratt, M.; Royston, P.; Kenward, M.G.; Wood, A.M.; Carpenter, J.R. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **2009**, *338*, b2393. [CrossRef] [PubMed]
46. Montez-Rath, M.E.; Winkelmayer, W.C.; Desai, M. Addressing Missing Data in Clinical Studies of Kidney Diseases. *Clin. J. Am. Soc. Nephrol.* **2014**, *9*, 1328. [CrossRef] [PubMed]
47. Nooraee, N.; Molenberghs, G.; Ormel, J.; Van den Heuvel, E.R. Strategies for handling missing data in longitudinal studies with questionnaires. *J. Stat. Comput. Simul.* **2018**, *88*, 3415–3436. [CrossRef]
48. Ebrahim, G.J. Missing Data in Clinical Studies Molenberghs G. and Kenward M. G. *J. Trop. Pediatr.* **2007**, *53*, 294. [CrossRef]
49. Carpenter, J.R.; Kenward, M.G. *Multiple imputation and its application*; John Wiley & Sons: London, UK, 2013; ISBN 9780470740521.
50. van Buuren, S.; Groothuis-Oudshoorn, K. **mice**: Multivariate Imputation by Chained Equations in *R. J. Stat. Softw.* **2011**, *45*. [CrossRef]
51. Zondervan-Zwijnenburg, M.A.J.; Veldkamp, S.A.M. Parental age and offspring childhood mental health: a multi-cohort, population-based investigation. *Child Dev.*. (in Press)
52. Thompson, A. Thinking big: Large-scale collaborative research in observational epidemiology. *Eur. J. Epidemiol.* **2009**, *24*, 727–731. [CrossRef]
53. Ford, I.; Norrie, J. Pragmatic Trials. *N. Engl. J. Med.* **2016**, *375*, 454–463. [CrossRef] [PubMed]