Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Smart Health (Amst). Author manuscript; available in PMC 2023 March 01.

Published in final edited form as:

Smart Health (Amst). 2022 March ; 23: . doi:10.1016/j.smhl.2021.100263.

A review of harmonization methods for studying dietary patterns

Venkata Sukumar Gurugubelli[a], **Hua Fang**[a,c,*], **James M Shikany**[b], **Salvador V Balkus**[a], **Joshua Rumbut**[a,c], **Hieu Ngo**[a], **Honggang Wang**[a], **Jeroan J Allison**[c], **Lyn M. Steffen**[d]

[a]University of Massachusetts Dartmouth, 285 Old Westport Rd, North Dartmouth, 02747, Massachusetts, USA

[b]Division of Preventive Medicine, University of Alabama at Birmingham, 1720 University Blvd, Birmingham, 35294, Alabama, USA

[c]Department of Quantitative Health Sciences, University of Massachusetts Medical School, 55 N Lake Ave, Worcester, 01655, Massachusetts, USA

[d]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, 55455, Minnesota, USA

Abstract

Data harmonization is the process by which each of the variables from different research studies are standardized to similar units resulting in comparable datasets. These data may be integrated for more powerful and accurate examination and prediction of outcomes for use in the intelligent and smart electronic health software programs and systems. Prospective harmonization is performed when researchers create guidelines for gathering and managing the data before data collection begins. In contrast, retrospective harmonization is performed by pooling previously collected data from various studies using expert domain knowledge to identify and translate variables. In nutritional epidemiology, dietary data harmonization is often necessary to construct the nutrient and food databases necessary to answer complex research questions and develop effective public health policy. In this paper, we review methods for effective data harmonization, including developing a harmonization plan, which common standards already exist for harmonization, and defining variables needed to harmonize datasets. Currently, several large-scale studies maintain harmonized nutrient databases, especially in Europe, and steps have been proposed to inform the retrospective harmonization process. As an example, data harmonization methods are applied to several U.S longitudinal diet datasets. Based on our review, considerations for future dietary data harmonization include user agreements for sharing private data among participating studies, defining variables and data dictionaries that accurately map variables among studies, and the use of secure data storage servers to maintain privacy. These considerations establish necessary

[*]Corresponding author. Tel.: +0-508-910-6411; hfang2@umassd.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

components of harmonized data for smart health applications which can promote healthier eating and provide greater insights into the effect of dietary patterns on health.

Keywords

Data harmonization; dietary data; diet quality; pattern; longitudinal; intelligent; smart health; randomized controlled trial; observation study

1. Introduction

Healthy eating is one of the foundations of a healthy lifestyle. A proper diet yields numerous benefits, preventing a variety of diseases and helping people stay physically fit and live longer. In fact, many health agencies such as American Heart Association (AHA), American Cancer Society (ACS), and community-based programs recommend a healthy diet as a preventive measure and treatment for obesity, diabetes (type-1, type-2), cardiovascular disease, and various cancers (Medina-Remon, Kirwan, Lamuela-Raventos, & Estruch, 2018; Schwedhelm, Boeing, Hoffmann, Aleksandrova, & Schwingshackl, 2016). As a result, a great deal of current medical research seeks to answer questions regarding how different diets impact health outcomes. Because of the inherent complexity of nutrition and the wide array of possible foods that people may consume on a daily basis, research in this area requires analyzing dietary patterns which reflect overall, habitual, long-term dietary intake across large populations in order to reach statistically robust conclusions. These dietary patterns inform the connection between nutrition and disease prevalence.

Dietary patterns characterize the variations in the population's dietary intake and further help nutritionists understand the relationship between the diet and disease. In nutritional epidemiology, dietary patterns are derived through empirical (a priori) or a posteriori method by analyzing data to identify the risk of prevalence or incidence of a disease or mortality through an individual or population's adherence to the identified patterns. Use of dietary patterns has gained popularity in recent years in order to overcome the conceptual and methodological limitations of analyzing just a single or a few nutrients or foods relative to disease (Hu, 2002). In the future, intelligent machine learning algorithms for smart health applications could even use dietary patterns for individualized intervention recommendations.

However, before dietary patterns can be used in health studies and in intelligent algorithms for smart health, larger datasets are necessary. Despite the popularity and usefulness of dietary pattern analysis, its merits are still being debated in the research community (Hu et al., 1999; Nanri et al., 2012; Roman-Vinas et al., 2009). This is due to challenges related to analysis of usefulness and validity of dietary pattern studies. In addition, collecting and analyzing dietary data in individual studies can be labor-intensive, time-consuming, and expensive. As a result, existing studies primarily rely on small, targeted populations, and the conclusions drawn from these studies are limited. If larger, more general datasets could be collected, the validity and generalizability of research conclusions using dietary patterns could be vastly strengthened.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

How could such datasets be compiled? By combining data from many different studies, researchers can construct larger datasets which support stronger conclusions. Data from multiple sources can be aggregated using a process known as *data harmonization*. Integrating multiple data sources with common features can increase the comparability of research data collected across independent studies, to help find answers to research questions that a single individual study cannot be due to insufficient sample size. When harmonizing datasets from various studies, it is essential to set up harmonization protocols and refer to guidelines set forth before the study commences at participating studies to ensure the accuracy and reproducibility of harmonization results. With the proper procedures in place, data harmonization can help researchers construct datasets that support more robust dietary pattern analysis. This, in turn, will improve research in nutritional epidemiology and allow the development of smart health applications for nutrition.

To support data harmonization efforts in dietary studies, in this review, we identify and outline the existing harmonization approaches and their application to nutritional epidemiology. In section 2, we summarized the outcomes of our review of data harmonization methods and outcomes. Section 3 describes the benefits of harmonization and steps involving the harmonization plan, including examples of harmonization standards, execution, limitations, and challenges. Section 4 details the steps involved in implementing harmonization strategies we learned for longitudinal dietary data obtained from different local and national studies. Finally, we conclude and summarize our findings from the survey.

2. Methods

This review includes a discussion of existing strategies for data harmonization in both observational and randomized controlled trials (RCTs) in epidemiology. In our review, we included published harmonization efforts in the field of nutrition and dietary studies between 2000 and 2017. We used keyword searches “diet” and “harmonization” to identify this literature. We excluded those articles which were not relevant to diet or harmonization in the literature search, as well as those that are not freely available. We performed an extensive search using PubMed and Google Scholar to identify relevant literature and to uncover creative approaches for harmonization. The currently used approaches are summarized in this paper. We also include a section describing a case study of harmonizing dietary data using retrospective harmonization methods.

Data harmonization is the process of pooling data from multiple studies into one useable dataset. It is a strategy to increase the comparability of research data collected across independent studies to help find answers for research questions that a single individual study cannot find. This strategy has gained popularity in recent years in the field of nutritional epidemiology. Extending the usability and validity of methods used for creating dietary patterns using a single harmonized dataset may help researchers identify more accurate associations between diet and various disease outcomes. By pooling the data from multiple-center studies, independent or single-center studies, and limited data sources, researchers can address critical questions regarding the broader impact of medical, psychological, and behavioral research. Consolidating data from similar studies in any chosen field can help researchers make informed decisions about disease incidence or risk factors (Angrisani &

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Jinkook, 2012). Harmonization also plays a vital role in developing standardized indicators to measure the impact of dietary recommendations on different subgroups within a country or globally (Dubois, Girard, & Bergeron, 2007), essential in creating centralized databases (Beer-Borst et al., 2000). These harmonized databases enable researchers to search across disease conditions and dynamically generate charts, maps, and tables (K. Elmore et al., 2014).

There are two types of harmonization – prospective and retrospective. ***Prospective harmonization*** is possible when investigators collectively set up guidelines for collecting, managing, and pooling of the data before initiating the studies. Prospective harmonization can help supply core measures before the data collection, allowing flexibility in data collection on unique characteristics and reducing the efforts required to address the limitations of retrospective harmonization. Prospective harmonization allows more flexibility in study design and minimizes cost and time required, as the technical implementation becomes complex when harmonization is done after designing the study.

On the other hand, ***Retrospective harmonization*** is performed by pooling the data from different studies after data collection and using the knowledge of domain experts to identify and translate study-specific variables to variables with common definitions and units of measure. In both types of harmonization, domain experts should identify and document schematic concepts in the earlier stages, which can be used in later stages of the process. During retrospective harmonization, however, most of the time, data obtained from different studies initially are not harmonized (or have common definitions or units of measure), or documentation regarding the methods used for harmonization may not be provided, leading to unexplained variation in results.

Both types of data harmonization are useful for a wide array of dietary studies. While ongoing or completed independent diet studies collect dietary intake data, data harmonization among dietary studies can provide more powerful and accurate assessment of the relationship between diet and disease in a broader spectrum. Frameworks to estimate average nutrient requirements have been developed in the past by different research groups (King & Garza, 2007). Having research questions, hypotheses, critical data domains, and specific items necessary to address the critical research questions determined before rather than after study completion can serve the objectives of harmonization and reduce the efforts required to integrate data from diverse data sources (Chandler et al., 2015), including international studies.

Current harmonization efforts require planning and cooperation between different researchers. Research projects to standardize data from international studies include – the Biobank Standardization and Harmonization for Research Excellence in European Union (BioSHaRE) (Doiron et al., 2013), European Project on Osteoarthritis (EPOSA) (Schaap et al., 2011), Network of Cohorts in Europe and United States (CHANCES) (Boffetta et al., 2014), Monica Risk, Genetics, Archiving and Monograph (MORGA M) (Evans et al., 2005), the European Prospective Investigation into Cancer and Nutrition (EPIC) study (Orfanos et al., 2007), and Nutrition in Adolescence (HELENA) (Moreno et al., 2008). While developing data harmonization tools and standardized Information systems

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

for centralized databases, these projects also addressed the issues raised when pooling data from individual studies across cohorts in international research projects. For example, the harmonization process involved rigorous study selection criteria, which focused on assessing the comprehensibility of data and access to aggregated data remotely for statistical analysis. This prospective initiative allows the research community to choose harmonized variables required to answer their research question. Periodic workshops will enable investigators to validate the variable definitions. (Schaap et al., 2011) (Boffetta et al., 2014; Evans et al., 2005). While harmonization efforts have been recently promoted and are ongoing, more harmonization initiatives will help expand research opportunities.

There has been a growing interest in the use of Machine learning algorithms in nutrition in recent years. Studies are starting to take advantage of the power of machine learning algorithms to extract useful information, find patterns and predict diseases. As data obtained during a dietary study can be significantly large from food frequency questionnaires, smartphone/application/ food registry. Supervised machine learning algorithms such as Neural networks, decision tress along with classification and regression are used in dietary data analyses. (Oliveira Chaves et al., 2021) As harmonization enables pools data from different sources, similar deep learning and machine learning methods can be applied to analyze the data to identify patterns or correlations.

Besides cooperation in deciding the study's criteria, the harmonization process also involves technical details to determine how the data will be combined. Variables required to generate target variables (harmonized variables) are identified by assessing the study questionnaires, data collection procedures, data entry and analysis nutrient and food databases, if applicable, and data dictionaries developed by each study. Harmonized variables can be uploaded to secure servers at the host institution to apply processing algorithms implemented for each study to generate the target variables. For secure and easy access to data summaries, password-protected portals should be set up and organized by the host institution server. Funding agencies should recommend that investigators and staff participate in face-to-face meetings to educate them about harmonization protocols ahead of study initiation (Chandler et al., 2015) (Erten-Lyons et al., 2012) to ensure a mutual understanding of these harmonization methods. Now that we have described what harmonization is and how to approach harmonization; in Section 3, we will summarize the benefits and ways to harmonize data, how to develop a plan for harmonization, and execution of the plan. Section 4 describes our experience with data harmonization and the challenges we encountered during this process.

3. A Review of Data Harmonization Methods and Outcomes

3.1. The Benefits of Harmonization

As the breadth and depth of global research output increase, harmonization has become imperative to provide the tools and knowledge needed for more advanced scientific studies. Cancer and genetic studies that implement data harmonization practices improve the ability to answer complex research questions and identify rare outcomes, especially when stratifying data by genotype or other subcategory (Rolland et al., 2015). Worldwide harmonization of nutrient-based dietary standards for micronutrients will significantly

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

benefit health quality and policy development and implementation. Independent studies that follow the same definition for diseases are more straightforward to incorporate in harmonized databases than those that do not have the same specific definitions. These benefits result from the standardization and pooling of data resulting from harmonization practices.

Harmonization initiatives may also allow the construction of food or nutrient databases to be leveraged for scientific and policy purposes. For example, researchers have developed harmonized databases for studying dietary lignans in foods across several countries (Finland, Netherlands, United States, Canada, United Kingdom, Japan, and Spain) (Durazzo et al., 2018; Peterson et al., 2010), which promoted more investigation about the health effects of this compound. The University of Minnesota Nutrient Database, constructed using harmonized data from two case-control dietary studies, allowed researchers to study the association between sugar and starch intake, and the risk of Barrett's esophagus (Li et al., 2017). Furthermore, harmonized food classifications support the development of dietary quality measures; these include Nutri-Score, which measures nutritional quality of foods to provide comprehensive food package labeling for consumers (Dreano-Trecant et al., 2020), and the Dietary Approaches to Stop Hypertension (DASH) index, which provides a single quantitative score of diet quality (Steinberg, Bennett, & Svetkey, 2017) to prevent and control hypertension according to National Heart, Lung, and Blood Institute's guidelines. These quantitative tools, created through data harmonization, promote more dietary research. Table 1 lists all the projects which use harmonization to achieve different objectives. Data harmonization projects create opportunity to conduct joint research and enables debate among subject matter experts to design studies that can focus on identifying health risks across different continents/demographics goals and design better research projects.

Because of the benefits of harmonization in creating joint research and data storage standards, collaborative projects have been formed in Europe to improve and promote harmonization in dietary studies and databases. These efforts include EuroFOODS, EU Cost Action 99, IARC European Nutrient Data Bank, and INFOODS/EuroFIR. These resources promote the harmonization of dietary data across several types of studies. Data sources such as EuroFIR FoodEXplorer provide databases of food nutrients that researchers may link to other data sources. In addition, EuroFIR thesaurus allow the comparison between food composition.

These resources can be used to compose harmonized datasets on the nutritional composition of foods for a broad spectrum of activities in public health nutrition, research, and government policy development and implementation (Egan, Fragodt, Raats, Hodgkins, & Lumbers, 2007). In addition, many studies have adopted standardized definitions for various diseases and their risk factors, allowing researchers to pool data from these studies to build databases to better assess the impact of those risk factors on disease more effectively. For example, many dietary studies have adopted the standard definitions of metabolic syndrome (MetS) as defined by the National Heart, Lung and Blood Institute and International Diabetes Federation in conjunction with the American Heart Association, World Heart Federation, International Atherosclerosis Society, and International Association for the Study of Obesity (Babio et al., 2015; Bellisle, 2014; Fernandez-Montero et al., 2013;

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Kuroki, Kanauchi, & Kanauchi, 2012; Sayon-Orea et al., 2014). A greater standardization across country-specific databases may reduce measurement error (Summer et al., 2013). Harmonization allows exploration of the relationship between chronic disease and dietary patterns. Thus, researchers should develop a strong harmonization plan to enable their data to be applied to a multitude of research questions.

3.2. Developing a Harmonization Plan

Data harmonization is highly beneficial but undergoing such a process is not simple. Therefore, a well-formulated plan is a key to successful data harmonization – especially prospective harmonization efforts. So, how is such a plan developed? During prospective harmonization, as outlined in Elmore et al. 2014 (K. N. Elmore, R.; Gant, Z.; Jeffries, C.; Broeker, L.; Mirabito, M.; Roberts, H., 2014), organizations should expect challenges and unexpected delays in the process. Data harmonization requires informatics skills as well as domain knowledge. Therefore, an organized research team that includes experts in both management and technical aspects is necessary to achieve maximum productivity. Domain experts should be consulted to select the standard data elements to be compared across different datasets. To provide vision and to push the integration effort forward, senior leadership should also be involved.

Furthermore, since the data-sharing policies between research groups play a critical role in harmonizing data, participating studies must discuss and harmonize policies. In addition, results shall be interpreted carefully from the harmonized data. If the outcome of a harmonization effort involves the display and stratification of specific data, some studies might not allow it due to the preexisting confidentiality agreements. Thus, collaboration and effective communication between the research teams seeking to harmonize data is essential.

On the other hand, if the collaborators have already collected the data, harmonization must be performed retrospectively, taking into consideration the differences in measures between the datasets. One approach to retrospective harmonization is the “question-first” paradigm, useful for hypothesis-driven research efforts involving specific variables. This method generally requires a five-step process to complete. First, the research group needs to identify what questions the harmonized data set must answer. Second, the high-level variables which can answer the stated research question should be identified. Third, an assessment should be performed on all the available data sets to check the availability of those variables. Fourth, common data elements need to be developed from the identified high-level variables. Fifth, data points should be mapped and transformed to the shared data elements. In the end, the quality of the harmonized data should be verified and validated (Arriaga et al., 2017; Fortier et al., 2017).

Conversely, organizations can harmonize data retrospectively by reviewing separate but similar datasets (containing similarly defined variables) to align them such that they may be used for comparison or stratification (K. Elmore et al., 2014). This alignment can be accomplished by translating data into a generic format or associating metadata to allow the comparison. Table 2 can be used to identify the level of compatibility between assessment items and target variables for harmonization for study compatibility assessment.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Incompatibility between key variables needed to answer the research question may preclude a study from the harmonization process.

Although the goal is complete harmonization, it can be achieved only when all variables are defined according to previously set data elements in the data sources using the same criteria. While some studies integrate data harmonization practices from the initial stages of a collaborative study (Chandler et al., 2015; K. Elmore et al., 2014), others provide a generic approach for the harmonization process even after data collection (Firnkorn, Ganzinger, Muley, Thomas, & Knaup, 2015). For example, in a research paper on dietary intake in Bangladesh, Karageorgou et al. put forth a seven-step process for retrospective harmonization (Karageorgou et al., 2018). These steps include identifying and retrieving necessary data, identifying unique food items, matching foods to food composition data, standardizing units, classifying foods into food groups, individualizing household consumption, and merging the data into a complete dataset. These steps represent an example of what a retrospective harmonization plan for a dietary study might entail when formally specified.

Both prospective and retrospective harmonization plans should outline proper protocols for future studies. Harmonization standards can help harmonize data across different stages of the longitudinal study when the protocols have changed over time (Rolland et al., 2015). Harmonization rules should be created for each of the variables based on the level of detail available for each variable. If some of the data from studies included detailed information for similar variables, additional variables must be created for increased granularity and should encompass available alternative formats. Categorical variables should be collapsed into the most granular level of detail to incorporate as many studies as possible and make efficient use of study data (Mishra et al., 2016). Furthermore, for dietary studies specifically, variables that are unavailable can be created by compounding outside data. For example, if nutrient data is not available ingredient databases may be used. Defining how to incorporate new data into an already harmonized dataset ensures that future studies can be incorporated more efficiently (Karageorgou et al., 2018).

3.3. Examples of Harmonization Standards

Because harmonization practices allow pooling data from many regional and international journals, they are often used to set regional and international research standards, including dietary and chronic disease patterns. For example, a standard codebook that was created based on the Eurocode 2 Core classification version 99/2 to provide coding for food items consumed worldwide (Goossens et al., 2016) promotes international harmonization for food labeling and food composition tables (Jones, 2014). Policymakers have emphasized the importance of harmonizing food composition data in Europe to improve health, trade (regulation and legislation), agriculture, and the environment (Egan et al., 2007). As a result, many large-scale harmonization studies aim to encourage researchers to include harmonization as their study objectives (Moreno et al., 2008) (Dekker et al., 2013). Although these are a step toward retrospective harmonization initiatives, there still exists a lack of prospective harmonization initiatives.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Prospective harmonization has been promoted by providing initiatives to include a standard format to collect common variables in future randomized controlled trials RCTs for specific research areas. One example of this is OBEsity Diverse Interventions Sharing – focusing on dietary and other interventions (OBEDIS), an expert-led project stating the minimal variables needed in adult obesity interventions. The “blueprint” provided by OBEDIS allows straightforward harmonization of data from obesity studies, including dietary data (Alligier et al., 2020). It recommends using the EPIC-Norfolk food frequency questionnaire for assessing dietary intake, the Dutch Healthy Diet index for assessing dietary quality, the Dutch Eating Behavioral questionnaire for assessing emotional eating, and a 3-day weighed food record. Having a standard set of variables simplifies data harmonization immensely; thus, researchers performing RCTs should implement such protocols before data collection begins.

Researchers can design questionnaires or study materials in dietary studies ready for harmonization using previously validated global coding manuals and food descriptions. During the harmonization of seven independent population-based surveys in six European countries to create a centralized database, the EURope ALIMentation (EURALIM) coding manual (Beer-Borst, 2000) was used to compare dietary measures across populations (Beer-Borst et al., 2000). Dietary assessment methods may vary across studies, which is a critical challenge to expect during retrospective harmonization. Converting fruit and vegetable intake quantities to frequencies (categorical variables) can help overcome this problem without considering different serving sizes across participating studies. Researchers should take careful measures to ensure interpretability and avoid loss of information after harmonization (Beer-Borst et al., 2000).

Several large-scale studies have already relied on harmonization practices and may serve as future models. The Environmental Determinants of Diabetes in the Young (TEDDY) study is a prospective, multi-center international study investigating associations between diet and other environmental factors and diabetes (Joslowski et al., 2017). The harmonized acrylamide database was compiled by using values from the E.U. monitoring database of acrylamide concentrations in foods maintained by the Institute for Reference Materials and Measurements (IRMM), which consists of a broad range of food commodities, all analyzed in the lab (Freisling et al., 2013). The national food, nutrition, and physical activity survey were designed to assess Portugal’s population diet and physical activity using data harmonized according to EU-MENU and European Food Safety Authority guidelines (Lopes et al., 2018). Harmonization has also been used to study the relationship between linoleic acid, arachidonic acid, and cardiovascular disease, and also protein and risk of diabetes (Marklund et al., 2019; Sluik et al., 2019).

By making basic information such as data summaries and descriptive statistics available to the collaborators, the coordinating center will control the data and monitor data access. When there was a need to perform more complex analysis in BioSHaRE, the DataSHIELD (Wolfson et al., 2010) method was employed in the R software environment (Doiron et al., 2013) for data transmission through a secure layer. BioSHaRE emphasized the importance of high-level collaboration between different parties to achieve the goals of harmonization. The BioSHaRE project also required the active involvement of study investigators

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and research center staff throughout the project. Mica-Opal federated framework and DataSHIELD R-package played vital roles in enabling a secured infrastructure that could tolerate computationally extensive tasks. The algorithms used for generating the required variables will vary depending on the agreement between locally available data and the response variables (Schaap et al., 2011).

3.4. Executing the Harmonization Plan: Variable Definition

Defining the variables needed to combine datasets from different studies is one of the most critical challenges in both prospective and retrospective harmonization. Therefore, all variables must be documented clearly and concisely (Olstad, Poirier, Naylor, Shearer, & Kirk, 2015). In the harmonization of nutrient profiling systems, for example, it is essential to avoid inconsistencies in variable definitions and results between the different studies (Pavlovic, Prentice, Thorsdottir, Wolfram, & Branca, 2007). For example, when defining “average requirement,” it is essential to mention the average requirement for a defined group of individuals or across different populations. Failure to do this could lead to misinterpretation and errors in analysis. Variations in this information could lead to severe health consequences in that regional population. Therefore, documenting variable definitions is so crucial.

To define variables properly, a systematic review may be necessary. Systematic reviews play a significant role in the process of harmonization by evaluating the strength and quality of evidence under review. These reviews are essential for deciding which variables are necessary to pool data. Furthermore, a recently published consensus study (“Harmonization of Approaches to Nutrient Reference Values,” 2018) outlines assessing the risk of bias as a significant challenge. Researchers may use a qualitative approach or perform a bias-adjusted meta-analysis to evaluate the risk of bias. In the process of systematic reviews, evidence mapping may be used to mitigate bias or uncertainty. Systematic reviews can also help overcome non-methodological challenges, such as constraints in policy, funding availability, and lack of expertise, by providing justification for variable definition choices to other professionals in the specific area being researched.

Global harmonization variable definitions will improve the objectivity and transparency of values derived by diverse regional, national, and international groups. While providing an everyday basis for experts, this process will also permit developing countries to convoke groups of experts to identify how to modify existing standards for a population’s specific requirements, objectives, and national policies. The harmonization task involves convening an ad hoc committee to review and assess methodological approaches to develop recommendations using experts’ evidence and discussions. The framework will consider impacts and trade-offs in consideration of methodological approaches for developing intake recommendations globally. Workshops and conferences should be conducted to reach a global agreement on methodologies. For establishing the recommended nutrient intake, either an existing systematic review is updated, or a new review is initiated. Each stage of the process must be documented, including the limitations in the data and methods to enable transparency in the harmonization process. All the uncertainties must be considered.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The report outlines the detailed process of deriving nutrient values, starting with identifying nutrient values to review until the selected nutrient values are accepted, revised, or derived.

Disparate datasets may define variables differently; thus, it is important to document these definitions when performing harmonization. After assessing the availability of the data on each research area, a common wiki website to document all the standard variable definitions may serve as a reference for the researchers who wish to perform analyses in the future. (Evans et al., 2005) After identifying the variables from which target variables of the parent project can be derived, new standard variables can be generated based on the data available from the cohorts. As a wiki website can help in attaining an agreement among the cohort investigators on standard variables and variable definitions via discussion boards, this approach can also serve as an alternative to workshops for those involved in harmonization. Finally, the data's availability, comparability, and quality for variables from each cohort can be assessed and documented in wikis. When harmonizing nutrients and international exchange of food data, standardizing and evaluating nutrient components such as energy and saturated fats are helpful based on the methodological objectives, and the data from different cohorts (Olstad et al., 2015; Orfanos et al., 2007; Pavlovic et al., 2007)

Finally, for prospective harmonization, one of the most effective ways of ensuring that different studies collect comparable variables is to develop standardized data collection procedures before data collection begins, but during the study design phase. Some recent studies have expounded specific criteria for measuring certain variables with harmonized data in dietary studies – for example, in diet diversity data collection for allergy and asthma studies; Venter et al. provided recommendations such as defining portion size quantitatively and measuring the frequency of specific food or food group exposure (Venter et al., 2020). In a study using harmonized data to measure the effect of workplace health promotion, Coenen et al. conclude that including the effectiveness and compliance of such studies is also necessary (Coenen et al., 2020). These allow the harmonization of data from multiple studies to obtain all necessary information for comparison. Furthermore, nutrition care terminology harmonization (Gabler et al., 2018) and pre-defined ontology frameworks for dietary studies – including Output of Nutritional Epidemiology (ONE), which defines a set of terms and the relationships between them in nutritional epidemiology (Yang et al., 2019) – can help ensure descriptions for measures are comparable across studies and that studies meet the criteria for interoperability. If data from multiple studies is already comparable, harmonization becomes highly simplified.

3.5. Limitations and Challenges during Harmonization:

Beyond potential missing values, the data harmonization process may involve both statistical and methodological challenges. Although data harmonization is more straightforward when performing comparative standardized studies based on a uniform core protocol or meta-analysis (Beer-Borst, 2000), researchers should expect critical challenges in the process of harmonizing data from independent studies. This is because slight measurement errors can bias the intended comparisons. Comparatively, it is cheaper to pool and analyze data from previous studies than starting a new study to incorporate harmonization standards. However, subject matter experts should evaluate the process in each study center using

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

standard evaluation questionnaires before integrating these data into an international system of risk factor surveillance which is used for identifying higher risk of chronic health conditions. (Beer-Borst, 2000). Data from different studies are not always comparable either; concessions made during harmonization may lead to data loss (Coenen et al., 2020), so consultations with experts may be necessary to determine whether harmonization between two studies is feasible to answer the desired research question.

Studies use different protocols depending on the data required for their objective, which raises another challenge—variations in the data available from cohorts present novel challenges in harmonization flow. Lack of harmonized information on core characteristics (e.g., occupational exposures) relevant to the project may limit the usefulness of the data for the intended purpose. For instance, when harmonizing dietary data, finding an optimal value for dietary intake is challenging for two reasons (Beaton, 2009). First, the requirements of dietary intake may vary among individuals based on individual metabolic capacities. Second, where there is unknown measurement error, moving the cutoff point can only change the proportion of error among two categories; however, it is impossible to avoid measurement error. This becomes challenging in the process of either prospective or retrospective harmonization when different studies use improper tools and elaborate definitions of common data elements. Thus, comparing studies may be difficult.

Researchers should also expect challenges specific to harmonizing dietary data. The enormous number of foods, constantly changing nutrient composition of foods, and possible missing data values make maintaining food databases difficult (Kapsokefalou et al., 2019). In addition, some food data are challenging to measure - for example; it is challenging to assess intakes of red and processed meats separately because of disparities in how these foods are typically prepared and processed in different food cultures (Mendez & Kogevinas, 2011). Also, since food groupings in the food composition databases differ among nations, to overcome categorization difficulties, the Langua aLimentaria method (Ireland & Møller, 2013) was used to provide a standardized language to describe foods in a systematic way (Joslowski et al., 2017). Having different food composition tables can lead to errors in estimating dietary intake when comparing different countries (Kovalskys et al., 2015). Lacking standardized dietary methods for deriving dietary or nutrient patterns and standard nutrient databases, disparities in data collection, analysis, and interpretation of dietary data make identifying their association with the disease more challenging. (Moskal et al., 2014)

4. Implementation Of Harmonization for Longitudinal Dietary Data

In the field of smart health, researchers may seek to build intelligent algorithms or software systems to monitor or predict patient health outcomes. This often requires harmonizing data from multiple sources. However, as data harmonization is challenging, harmonizing specific data such as dietary data uncovers intricacies in implementing principles. Harmonizing dietary data requires large-scale collaborative work with the participating studies to achieve harmonization goals. This section describes a retrospective harmonization approach and the challenges encountered while implementing principles of harmonization as part of an ongoing project (Fang, 2019). As part of the retrospective harmonization process, this study selected four local studies (Merriam et al., 2009; Ockene et al., 2012; Schneider et al.,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2008; M. L. Wang et al., 2015) and two national-level studies ("Design of the Women's Health Initiative Clinical Trial and Observational Study," 1998; Friedman et al., 1988) which collected dietary data as part of their study design. The key to successful harmonization is a knowledge transfer between participating studies and the coordinating center during every step of the process. Documentation and inventory of all data, code, manuscripts, and other related materials associated with the process itself have been cataloged during every step of harmonization. In addition, this catalog will help in documentation and knowledge transfer for the collaborators who will work on the project at later stages. The complete process we followed is depicted in Figure 1. The steps shown in this figure are described in the following paragraphs.

Study identification and material acquisition:

Before the Harmonization process began, data and variables required to answer research questions were identified using * **Compatibility class can be used to determine whether the study can be included in the harmonization process.** (Fortier et al., 2010)

. All candidate studies were categorized into three classes, and those studies with complete and partial compatibility were included as part of the harmonization process. Those candidate studies which did not contain enough information to construct necessary variables were categorized under the impossible criteria (compatibility class can be used to determine whether the study can be included in the harmonization process) and were excluded (Fortier et al., 2010). In the next step, selected study sites were contacted through organizers or points of contact who could provide access to the necessary data. If the study sites required a proposal for data use, a document describing the purpose of harmonization was provided. This document clearly outlined the project's goals, how the coordinating center intended to use the data, a list of collaborators who would have access to the data, and the extent of use. This document also included a request for additional material related to data such as derived dietary variables, data dictionaries, and any documentation required to interpret or understand the data structure.

The coordinating center obtained data in available formats by working with participating study centers. The process started with obtaining signatures on data use agreements (DUAs) from all participating collaborators involved in the harmonization process who need access to work with the data closely. DUA is an agreement between the study center and principal investigators and collaborators, which outlines data usage and sharing policies. Upon data exchange, obtained data were stored on a secure server with controlled network access, limited to the collaborators who need to work with data. All new collaborators who need access were approved via the principal investigator to access data. When using the data, the collaborators must adhere to the data-sharing policies outlined by each study center per contractual agreements, which ensures data privacy.

Aggregation of data from multiple sites:

The beginning of the harmonization process was a thorough inventory of all data, code, manuscripts, and other material related to both local and national studies. This step allowed the team to track decision-making and avoid duplication of effort in the process. In addition,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

both local and national study files were organized in a consistent format for readability and accessibility. During all the stages of the harmonization process, the coordinating center has preserved communications and detailed meeting notes in secure storage.

Create harmonized data dictionary:

Although all participating study centers provided variable descriptions, a data dictionary was created at the coordinating center for each local and national study in a unified format to allow for comparison. The variables of interest for dietary pattern recognition were identified and put into a separate dictionary. Each data dictionary contained the equivalent variables from each local study (if available) along with any notes regarding discrepancies in units [e.g., kilograms, pounds], scales [e.g., variables produced from Likert scale questions, as in 0=never, 1=rarely, 2=sometimes]. For example, some might have a 5-level scale, others seven, others just yes/no, or available responses. In contrast, another might have several options].

Construction of harmonized dataset:

Dietary recall data, such as from 24-hour dietary recalls, food frequency questionnaires, and diet histories, present a harmonization challenge across studies. Food pattern and nutrient databases allow for converting the recall/history data to a set of nutrients that may be used to calculate dietary quality indices or derive dietary patterns. In the United States, the foundation of many of these databases was the USDA National Nutrient Database for Standard Reference (SR). A subset of the food and beverage items and their corresponding nutrients from the SR formed the basis for the Food and Nutrient Database for Dietary Studies (FNDDS). The Food Patterns Equivalents Database (FPED, formerly the MyPyramid Equivalents Database or MPED) converts FNDDS items into food groups based on USDA dietary guidance (Ahuja, Moshfegh, Holden, & Harris, 2013). A similar database, Food Intakes Converted to Retail Commodities Databases (FICRCD), translates FNDDS items into retail commodities. In 2019, these components were integrated into the FoodData Central system, which combined the above databases and the Branded Food Products Database (BFPDB), Experimental Foods, and an Application Programming Interface (API) for automated access (“A Consumer Food Data System for 2030 and Beyond,” 2020; “U.S. Department of Agriculture, Agricultural Research Service,” 2019). In addition, SR data was also used in developing other nutrient analysis systems, such as the University of Minnesota Nutrition Coordinating Center’s Nutrition Data System for Research (NDSR) and international food composition databases (Bowman SA et al., 2011; Heisey, King, Rubenstein, Bucks, & Welsh, 2012; Schakel, Sievert, & Buzzard, 1988).

Dietary quality indices such as Healthy Eating Index (HEI) have often been derived from FPED (Krebs-Smith et al., 2018), MPED (Guenther, Reedy, & Krebs-Smith, 2008), and NDSR outputs. These variables include food groups, subgroups, and nutrients; they will be created based on the collected original food item data and nutrient values computed from the NDSR. While most studies use NDSR food variables/ groups, some used alternative MPED, which presented a challenge in harmonizing dietary data further. New methods of calculating the same score from different input data must be evaluated to ensure equivalency. For example, the Geisinger Rural Aging Study (GRAS) compared the HEI-2005 scores

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of GRAS participants to age- and race-matched samples from the National Health and Nutrition Examination Survey (NHANES) 2001–2002, FNDDS and MPED databases to calculate HEI-2005 scores. Different versions of dietary quality indices do not necessarily have the same components. Each update of the food group databases such as FPED may create distinct groups, and harmonizing different versions requires careful evaluation. The National Cancer Institute's Automated Self-Administered 24-Hour Dietary Assessment Tool (ASA24) has released SAS code to translate between one version of the MPED database and a later FPED database [10], but other versions would require new efforts.

To identify disparities in the interpretation of the previously established food standardization protocol among four different studies (Mendez & Kogevinas, 2011), each collaborator has discussed the adjustments they have made to the initial groupings with the coordinator, which can help the coordinators to evaluate the standardization protocol. Collaborators expect that analyses performed to identify relevant health effects using standardized data may be strongly related to specific food subgroups rather than broader subgroups (Slimani et al., 2007). By supporting post hoc harmonization of intakes of selected food groups, this project serves as a basis for developing a harmonized database that can facilitate pooled analyses of prospective relationships between dietary intakes and health outcomes in pregnancy and offspring using eligible data from multiple cohorts. As nutrient analysis tends to be complex, many studies focus more on selected food groups to derive dietary patterns.

Storage and Infrastructure for Smart Health:

Having a shared data source for all sites will ensure prompt delivery and validation during the data collection stage. A central data source, or a system of distributed data sources, can enable investigators and coordinators to change the variable definitions/questionnaires only once to update them across sites. Having such infrastructure can help gather, store, and analyze data from a single place by enabling coordinators to control secure data access (Joslowski et al., 2017). Furthermore, such database infrastructure enables researchers to use the harmonized data in smart health applications – for example, the construction of a system for intelligent, real-time dietary pattern analysis or precision nutrition health.

The back-end system requires security, efficiency, and scalability of the database. The security aspect of the application guarantees that only researchers with appropriate permission have access to sensitive medical data or the permission to add/remove data. One of the essential aspects of designing our back-end database is a database management system (DBMS). A DBMS is software that communicates with the database, applications, and user interface to perform data entry, report generation, validation, and security maintenance (Nourani, Ayatollahi, & Dodaran, 2019). In table 1, we first find and compare the most used DBMS. There are two main types of DBMSs: relational and non-relational (X. Wang, Williams, Liu, & Croghan, 2019). The relational DBMSs use the Structured Query Language (SQL), and their data appear as tables of data with rows, columns, and a strict structure and explicit dependencies. With this integrated data storage structure, the relational schemas rely on the clean separation between data structure and data value. Additionally, one of the main advantages of this type of database is its maturity in various levels of data control, including granular security enforcement. The relational databases

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

are often used for information integration, ad hoc analysis, and reporting multiparameter and longitudinal data tracking. However, one of the disadvantages of relational databases is their querying performance in large-scale dataset as well as their scalability. Additionally, PostgreSQL, being an open-source advanced database with active development, provides the tool we need to develop an interactive and intelligent database.

On the other hand, the non-relational databases are not limited to a table structure. They can also store non-structured data such as articles, documents, photos, etc. Hence, as opposed to a relational DBMS, the non-relational databases are exceptional in scalability across multiple servers. They also require minimal pre-deployment preparations and can make quick updates to the data structure easier. However, the non-relational databases are limited in joining related data, and they lack data standardization. Both of these systems have been used to build database tools in practice (Gabetta et al., 2015; Harris et al., 2009; Ohno-Machado et al., 2017; Wade, Hum, & Murphy, 2011; X. Wang et al., 2009).

Due to the nature of the harmonized database containing personal health information, it is important that the chosen DBMS can maintain the harmonized database and personal health information. For this reason, relational databases are the better options. However, we need to overcome the disadvantages that come with relational databases. The first one is their querying performance in large-scale datasets. To solve this problem, we chose PostgreSQL, an open-source relation DBMS, that have been optimized to deal with a large amount of Spatio-temporal data. PostgreSQL is shown to be four times faster in response time in most cases compared to MongoDB, one of the most representative non-relation DBMS (Makris, Tserpes, Spiliopoulos, Zissis, & Anagnostopoulos, 2020). PostgreSQL also outperforms a time-series database (database optimized for time-stamped or time series data), InfluxDB, in querying speed. PostgreSQL has 83.52%-89.42% shorter query time compared to using InfluxDB database using large-scale dataset. The second problem is the scalability of the relational database. The main problem with relational DBMS is scalability is that they can only scale vertically, not horizontally. To increase the performance, we need to increase the load on a single server by increasing RAM, SSD, or CPU. Contrary, non-relational DBMS can scale horizontally by adding more servers to the database. PostgreSQL provides good scalability, allowing multiple technical options for scaling, both vertically and horizontally. Especially for read scalability, PostgreSQL has an efficient built-in replication system that can be utilized for scalability. Based on these findings, we found that PostgreSQL can satisfy the requirements for our harmonized database. Aside from the security of relational databases, PostgreSQL also provides fast querying time and scalability. Most importantly, it is open-source and community-driven, which means the technology will continue to advance and improve quality, enabling us to develop interactive and intelligent harmonized databases.

Data management is a critical challenge in biomedical research. The proposed harmonized database raised the demand for an enhanced data management capacity. The DBMS needs to provide efficient data access, data integrity and security, data administration, concurrent access, crash recovery, and reduced application development time. Most importantly, the DBMS needs to ensure the secure data access of the nutrition data such that only the investigators and coordinators can quickly access and update the central data source; the basic information such as data summaries and descriptive statistics are made available to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

researchers. Thus, the DBMS can serve as an analysis tool and a testing framework for diet-related research. Based on these requirements, we designed a database consisting of two main parts: the front-end user interface and the back-end DBMS. The user interface provides a fast and reliable framework that allows users to search and query the data from the database quickly and easily. It also allows the users to run our proposed machine learning methods for reference and test their methods using harmonized data for validation.

5. Discussion & Conclusion:

Data harmonization of diet data enables researchers to create databases that can increase the breadth and depth of global research productivity in smart health. Large-scale dietary data harmonization can be used for global risk surveillance. Regardless of perspective or retrospective harmonization, the process, usability, and validity of methods used for deriving dietary patterns from single dietary data set to derive patterns from harmonized dietary data will open opportunities for innovation in method design and evaluation. Selecting the studies which align with the goals of harmonization with clear criteria will reduce the magnitude of resources required to harmonize the data in the later steps. Data dictionaries which are provided by participant studies and those generated by the coordinating center play a crucial role in knowledge transfer and in achieving the goals of harmonization.

As data harmonization becomes more prevalent, distributed computing technologies can aid in its deployment. In future studies, rather than storing harmonized data in a centralized location, data sources can be distributed across multiple sites. In this way, large quantities of data can be stored and accessed by other users more quickly and easily. This also promotes increased reliability and easier expansion of data storage capabilities. Hence, distributed data storage can provide greater efficiency than centralized databases.

Furthermore, an advanced source of harmonized dietary data, such as the case study presented in this survey, will enable the development of intelligent technologies for health and nutrition. With extensive databases of dietary information, researchers can use machine learning models and pattern analysis tools to examine diet quality in real time. In addition, the data can be used in the development of smart health devices that monitor and improve diet quality. Large, complex datasets are often important component of smart health and intelligent systems, and data harmonization procedures provide researchers these quality datasets that are necessary to develop such systems.

The success of harmonization hinges on the concrete planning and communication between the coordinating center and the participating research centers and mutual understanding of the goals of the project. A clear understanding of the process, continuous cooperation, and communication between the participants allows progression and reduces friction in the process. As the process moves forward, the responsibility of the coordinating center centers on reaching the goals with the help of participant groups. As dietary patterns facilitate identifying associations between diet and disease, funding agencies should encourage researchers to use standardized definitions so that future studies can easily harmonize to compile risk factor databases [31] further. The European Human Biomonitoring Initiative (HBM4EU) study encouraged others to provide documentation containing clear definitions

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of common data elements to aid prospective harmonization initiatives (Berman, Goldsmith, Levine, & Grotto, 2017). There is a need for countries' political and administrative bodies to encourage researchers' efforts in data harmonization to understand diet/disease associations (Aubert et al., 2019; Dreano-Trecant et al., 2020; Durazzo et al., 2018; Li et al., 2017; Rolland et al., 2015) (Kapsokefalou et al., 2019). Dietary patterns derived from harmonized databases will allow epidemiologists to identify the causal relationship between aspects of diet and medical conditions across different populations with interpretable and accurate results.

It is clear that harmonized dietary datasets are useful across many different applications – studying the link between diet and disease, creating intelligent health systems, and improving dietary quality. In this paper, we have reviewed methods for effective data harmonization, including current standards, steps for developing a harmonization plan, and methods for defining common variables. By following the system of steps outlined in this work to combine disparate datasets into one common source, researchers will unlock more powerful insights and gain the ability to answer far more complex research questions than previously possible.

Acknowledgments

This research was partly supported by NIH 1R56DK114514-01A1 and NIH R01DK129432 to Dr. Fang.

References

Ahuja JK, Moshfegh AJ, Holden JM, & Harris E (2013). USDA food and nutrient databases provide the infrastructure for food and nutrition research, policy, and practice. J Nutr, 143(2), 241S–249S. doi:10.3945/jn.112.170043 [PubMed: 23269654]

Alligier M, Barres R, Blaak EE, Boirie Y, Bouwman J, Brunault P, … Laville M (2020). OBEDIS Core Variables Project: European Expert Guidelines on a Minimal Core Set of Variables to Include in Randomized, Controlled Clinical Trials of Obesity Interventions. Obes Facts, 13(1), 1–28. doi:10.1159/000505342 [PubMed: 31945762]

Angrisani M, & Jinkook L (2012). Harmonization of Cross-National Studies of Aging to the Health and Retirement Study: Income Measures. Retrieved from https://www.rand.org/pubs/working_papers/WR861z5.html

Arriaga ME, Vajdic CM, Canfell K, MacInnis R, Hull P, Magliano DJ, … Laaksonen MA (2017). The burden of cancer attributable to modifiable risk factors: the Australian cancer-PAF cohort consortium. BMJ Open, 7(6), e016178. doi:10.1136/bmjopen-2017-016178

Aubert AM, Forhan A, de Lauzon-Guillain B, Chen LW, Polanska K, Hanke W, … Bernard JY (2019). Deriving the Dietary Approaches to Stop Hypertension (DASH) Score in Women from Seven Pregnancy Cohorts from the European ALPHABET Consortium. Nutrients, 11(11). doi:10.3390/nu11112706

Babio N, Becerra-Tomas N, Martinez-Gonzalez MA, Corella D, Estmch R, Ros E, … Investigators, P. (2015). Consumption of Yogurt, Low-Fat Milk, and Other Low-Fat Dairy Products Is Associated with Lower Risk of Metabolic Syndrome Incidence in an Elderly Mediterranean Population. J Nutr, 145(10), 2308–2316. doi:10.3945/jn.115.214593 [PubMed: 26290009]

Beaton GH (2009). 1986 E.V. McCOLLUM INTERNATIONAL LECTURESHIP IN NUTRITION. Toward Harmonization of Dietary, Biochemical, and Clinical Assessments: The Meanings of Nutritional Status and Requirements. Nutrition Reviews, 44(11), 349–358. doi:10.1111/j.1753-4887.1986.tb07570.x

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Beer-Borst S (2000). Obesity and other health determinants across Europe: The EURALIM Project. Journal of Epidemiology & Community Health, 54(6), 424–430. doi:10.1136/jech.54.6.424 [PubMed: 10818117]

Beer-Borst S, Hercberg S, Morabia A, Bernstein MS, Galan P, Galasso R, … Northridge ME (2000). Dietary patterns in six European populations: results from EURALIM, a collaborative European data harmonization and information campaign. European Journal of Clinical Nutrition, 54(3), 253–262. doi:10.1038/sj.ejcn.1600934 [PubMed: 10713749]

Bellisle F (2014). Meals and snacking, diet quality and energy balance. Physiol Behav, 134, 38–43. doi:10.1016/j.physbeh.2014.03.010 [PubMed: 24657181]

Berman T, Goldsmith R, Levine H, & Grotto I (2017). Human biomonitoring in Israel: Recent results and lessons learned. Int J Hyg Environ Health, 220(2 Pt A), 6–12. doi:10.1016/j.ijheh.2016.09.008 [PubMed: 27663636]

Boffetta P, Bobak M, Borsch-Supan A, Brenner H, Eriksson S, Grodstein F, … Trichopoulou A (2014). The Consortium on Health and Ageing: Network of Cohorts in Europe and the United States (CHANCES) project--design, population and data harmonization of a large-scale, international study. Eur J Epidemiol, 29(12), 929–936. doi:10.1007/s10654-014-9977-1 [PubMed: 25504016]

Bowman SA, Martin CL, Friday JE, Clemens J, Moshfegh AJ, Lin B, & HF W. (2011). Retail food commodity intakes: mean amounts of retail commodities per individual, 2001-2002. Retrieved from http://www.ars.usda.gov/ba/bhnrc/fsrg

Chandler RK, Kahana SY, Fletcher B, Jones D, Finger MS, Aklin WM, … Webb C (2015). Data Collection and Harmonization in HIV Research: The Seek, Test, Treat, and Retain Initiative at the National Institute on Drug Abuse. American Journal of Public Health, 105(12), 2416–2422. doi:10.2105/Ajph.2015.302788 [PubMed: 26469642]

Coenen P, Robroek SJW, van der Beek AJ, Boot CRL, van Lenthe FJ, Burdorf A, & Oude Hengel KM (2020). Socioeconomic inequalities in effectiveness of and compliance to workplace health promotion programs: an individual participant data (IPD) meta-analysis. Int J Behav Nutr Phys Act, 17(1), 112. doi:10.1186/s12966-020-01002-w [PubMed: 32887617]

. A Consumer Food Data System for 2030 and Beyond. (2020). In A Consumer Food Data System for 2030 and Beyond. Washington (DC).

Dekker LH, Boer JM, Stricker MD, Busschers WB, Snijder MB, Nicolaou M, & Verschuren WM (2013). Dietary patterns within a population are more reproducible than those of individuals. J Nutr, 143(11), 1728–1735. doi:10.3945/jn.113.177477 [PubMed: 24027185]

Design of the Women's Health Initiative Clinical Trial and Observational Study. (1998). Controlled Clinical Trials, 19(1), 61–109. doi:10.1016/s0197-2456(97)00078-0 [PubMed: 9492970]

Doiron D, Burton P, Marcon Y, Gaye A, Wolffenbuttel BHR, Perola M, … Fortier I (2013). Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg Themes Epidemiol, 10(1), 12. doi:10.1186/1742-7622-10-12 [PubMed: 24257327]

Dreano-Trecant L, Egnell M, Hercberg S, Galan P, Soudon J, Fialon M, … Julia C (2020). Performance of the Front-of-Pack Nutrition Label Nutri-Score to Discriminate the Nutritional Quality of Foods Products: A Comparative Study across 8 European Countries. Nutrients, 12(5). doi:10.3390/nu12051303

Dubois L, Girard M, & Bergeron N (2007). The choice of a diet quality indicator to evaluate the nutritional health of populations. Public Health Nutrition, 3(03). doi:10.1017/s1368980000000409

Durazzo A, Lucarini M, Camilli E, Marconi S, Gabrielli P, Lisciani S, … Marletta L (2018). Dietary Lignans: Definition, Description and Research Trends in Databases Development. Molecules, 23(12). doi:10.3390/molecules23123251

Egan MB, Fragodt A, Raats MM, Hodgkins C, & Lumbers M (2007). The importance of harmonizing food composition data across Europe. Eur J Clin Nutr, 61(7), 813–821. doi:10.1038/sj.ejcn.1602823 [PubMed: 17554245]

Elmore K, Nelson R, Gant Z, Jeffries C, Broeker L, Mirabito M, & Roberts H (2014). Data harmonization process for creating the National Center for HTV/AIDS, Viral Hepatitis, STD, and TB Prevention Atlas. Public Health Rep, 129 Suppl 1, 63–69. doi:10.1177/00333549141291S110 [PubMed: 24385651]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Elmore KN,R; Gant Z; Jeffries C; Broeker L; Mirabito M; Roberts H (2014). Data harmonization process for creating the National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention Atlas. Public Health Rep, 129 Suppl 1, 63–69. doi:10.1177/00333549141291S110 [PubMed: 24385651]

Erten-Lyons D, Sherbakov LO, Piccinin AM, Hofer SM, Dodge HH, Quinn JF, … Kaye JA (2012). Review of selected databases of longitudinal aging studies. Alzheimers Dement, 8(6), 584–589. doi:10.1016/j.jalz.2011.09.232 [PubMed: 23102128]

Evans A, Salomaa V, Kulathinal S, Asplund K, Cambien F, Ferrario M, … Project, M. (2005). MORGAM (an international pooling of cardiovascular cohorts). Int J Epidemiol, 34(1), 21–27. doi:10.1093/ije/dyh327 [PubMed: 15561751]

Fang H (2019). VIP: Visual-valid dietary behavior pattern recognition for local national trials. In. University of Massachusetts - Dartmouth: National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

Fernandez-Montero A, Bes-Rastrollo M, Beunza JJ, Barrio-Lopez MT, de la Fuente-Arrillaga C, Moreno-Galarraga L, & Martinez-Gonzalez MA (2013). Nut consumption and incidence of metabolic syndrome after 6-year follow-up: the SUN (Seguimiento Universidad de Navarra, University of Navarra Follow-up) cohort. Public Health Nutr, 16(11), 2064–2072. doi:10.1017/S1368980012004442 [PubMed: 23092760]

Firnkorn D, Ganzinger M, Muley T, Thomas M, & Knaup P (2015). A Generic Data Harmonization Process for Cross-linked Research and Network Interaction. Construction and Application for the Lung Cancer Phenotype Database of the German Center for Lung Research. Methods Inf Med, 54(5), 455–460. doi:10.3414/ME14-02-0030 [PubMed: 26394900]

Fortier I, Burton PR, Robson PJ, Ferretti V, Little J, L'Heureux F, … Hudson TJ (2010). Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. Int J Epidemiol, 39(5), 1383–1393. doi:10.1093/ije/dyq139 [PubMed: 20813861]

Fortier I, Raina P, Van den Heuvel ER, Griffith LE, Craig C, Saliba M, … Burton P (2017). Maelstrom Research guidelines for rigorous retrospective data harmonization. Int J Epidemiol, 46(1), 103–105. doi:10.1093/ije/dyw075 [PubMed: 27272186]

Freisling H, Moskal A, Ferrari P, Nicolas G, Knaze V, Clavel-Chapelon F, … Slimani N (2013). Dietary acrylamide intake of adults in the European Prospective Investigation into Cancer and Nutrition differs greatly according to geographical region. Eur J Nutr, 52(4), 1369–1380. doi:10.1007/s00394-012-0446-x [PubMed: 23238529]

Friedman GD, Cutter GR, Donahue RP, Hughes GH, Hulley SB, Jacobs DR, … Savage PJ (1988). Cardia: study design, recruitment, and some characteristics of the examined subjects. Journal of Clinical Epidemiology, 41(11), 1105–1116. doi:10.1016/0895-4356(88)90080-7 [PubMed: 3204420]

Gabetta M, Limongelli I, Rizzo E, Riva A, Segagni D, & Bellazzi R (2015). BigQ: a NoSQL based framework to handle genomic variants in i2b2. BMC Bioinformatics, 16, 415. doi:10.1186/s12859-015-0861-0 [PubMed: 26714792]

Gabler GJ, Coenen M, Bolleurs C, Visser WK, Runia S, Heerkens YF, & Stamm TA (2018). Toward Harmonization of the Nutrition Care Process Terminology and the International Classification of Functioning, Disability and Health-Dietetics: Results of a Mapping Exercise and Implications for Nutrition and Dietetics Practice and Research. J Acad Nutr Diet, 118(1), 13–20 e13. doi:10.1016/j.jand.2016.12.002 [PubMed: 28169211]

Goossens ME, Isa F, Brinkman M, Mak D, Reulen R, Wesselius A, … Zeegers MP (2016). International pooled study on diet and bladder cancer: the bladder cancer, epidemiology and nutritional determinants (BLEND) study: design and baseline characteristics. Arch Public Health, 74, 30. doi:10.1186/s13690-016-0140-1 [PubMed: 27386115]

Guenther PM, Reedy J, & Krebs-Smith SM (2008). Development of the Healthy Eating Index-2005. J Am Diet Assoc, 108(11), 1896–1901. doi:10.1016/j.jada.2008.08.016 [PubMed: 18954580]

. Harmonization of Approaches to Nutrient Reference Values. (2018). In Harmonization of Approaches to Nutrient Reference Values: Applications to Young Children and Women of Reproductive Age. Washington (DC).

Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, & Conde JG (2009). Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

translational research informatics support. J Biomed Inform, 42(2), 377–381. doi:10.1016/j.jbi.2008.08.010 [PubMed: 18929686]

Heisey PW, King JL, Rubenstein KD, Bucks DA, & Welsh R (2012). Assessing the Benefits of Public Research Within an Economic Framework: The Case of USDA's Agricultural Research Service: U.S. Dept. of Agr.

Hu FB (2002). Dietary pattern analysis: a new direction in nutritional epidemiology. Curr Opin Lipidol, 13(1), 3–9. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11790957 [PubMed: 11790957]

Hu FB, Rimm E, Smith-Warner SA, Feskanich D, Stampfer MJ, Ascherio A, … Willett WC (1999). Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. Am J Clin Nutr, 69(2), 243–249. doi:10.1093/ajcn/69.2.243 [PubMed: 9989687]

Ireland J, & Møller A (2013). What's New in LanguaL™? Procedia Food Science, 2, 117–121. doi:10.1016/j.profoo.2013.04.018

Jones JM (2014). CODEX-aligned dietary fiber definitions help to bridge the 'fiber gap'. Nutr J, 13, 34. doi:10.1186/1475-2891-13-34 [PubMed: 24725724]

Joslowski G, Yang J, Aronsson CA, Ahonen S, Butterworth M, Rautanen J, … Group, T. S. (2017). Development of a harmonized food grouping system for between-country comparisons in the TEDDY Study. J Food Compost Anal, 63, 79–88. doi:10.1016/j.jfca.2017.07.037 [PubMed: 29151672]

Kapsokefalou M, Roe M, Turrini A, Costa HS, Martinez-Victoria E, Marletta L, … Finglas P (2019). Food Composition at Present: New Challenges. Nutrients, 11(8). doi:10.3390/nu11081714

Karageorgou D, Imamura F, Zhang J, Shi P, Mozaffarian D, & Micha R (2018). Assessing dietary intakes from household budget surveys: A national analysis in Bangladesh. PFoS One, 13(8), e0202831. doi:10.1371/journal.pone.0202831

King JC, & Garza C (2007). Harmonization of nutrient intake values. Food Nutr Bull, 28(1 Suppl International), S3–12. doi:10.1177/15648265070281S101 [PubMed: 17521115]

Kovalskys I, Fisberg M, Gomez G, Rigotti A, Cortes LY, Yepez MC, … Group, E. S. (2015). Standardization of the Food Composition Database Used in the Latin American Nutrition and Health Study (ELANS). Nutrients, 7(9), 7914–7924. doi:10.3390/nu7095373 [PubMed: 26389952]

Krebs-Smith SM, Pannucci TE, Subar AF, Kirkpatrick SI, Lerman JL, Tooze JA, … Reedy J (2018). Update of the Healthy Eating Index: HEI-2015. J Acad Nutr Diet, 118(9), 1591–1602. doi:10.1016/j.jand.2018.05.021 [PubMed: 30146071]

Kuroki Y, Kanauchi K, & Kanauchi M (2012). Adherence index to the American Heart Association Diet and Lifestyle Recommendation is associated with the metabolic syndrome in Japanese male workers. Eur J Intern Med, 23(8), e199–203. doi:10.1016/j.ejim.2012.08.002 [PubMed: 22951435]

Li N, Petrick JL, Steck SE, Bradshaw PT, McClain KM, Niehoff NM, … Gammon MD (2017). Dietary sugar/starches intake and Barrett's esophagus: a pooled analysis. Eur J Epidemiol, 32(11), 1007–1017. doi:10.1007/s10654-017-0301-8 [PubMed: 28864851]

Lopes C, Torres D, Oliveira A, Severe M, Guiomar S, Alarcao V, … Consortium, I.-A. (2018). National Food, Nutrition, and Physical Activity Survey of the Portuguese General Population (2015-2016): Protocol for Design and Development. JMIR Res Protoc, 7(2), e42. doi:10.2196/resprot.8990 [PubMed: 29449204]

Makris A, Tserpes K, Spiliopoulos G, Zissis D, & Anagnostopoulos D (2020). MongoDB Vs PostgreSQL: A comparative study on performance aspects. GeoInformatica, 25(2), 243–268. doi:10.1007/s10707-020-00407-w

Marklund M, Wu JHY, Imamura F, Del Gobbo LC, Fretts A, de Goede J, … Outcomes Research, C. (2019). Biomarkers of Dietary Omega-6 Fatty Acids and Incident Cardiovascular Disease and Mortality. Circulation, 139(21), 2422–2436. doi:10.1161/CIRCULATIONAHA.118.038908 [PubMed: 30971107]

Medina-Remon A, Kirwan R, Lamuela-Raventos RM, & Estruch R (2018). Dietary patterns and the risk of obesity, type 2 diabetes mellitus, cardiovascular diseases, asthma, and neurodegenerative diseases. Crit Rev Food Sci Nutr, 58(2), 262–296. doi:10.1080/10408398.2016.1158690 [PubMed: 27127938]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Mendez MA, & Kogevinas M (2011). A comparative analysis of dietary intakes during pregnancy in Europe: a planned pooled analysis of birth cohort studies. Am J Clin Nutr, 94(6 Suppl), 1993S–1999S. doi:10.3945/ajcn.110.001164 [PubMed: 21974890]

Merriam PA, Ma Y, Olendzki BC, Schneider KL, Li W, Ockene IS, & Pagoto SL (2009). Design and methods for testing a simple dietary message to improve weight loss and dietary quality. BMC Med Res Methodol, 9, 87. doi:10.1186/1471-2288-9-87 [PubMed: 20042092]

Mishra GD, Chung HF, Pandeya N, Dobson AJ, Jones L, Avis NE, … Anderson D (2016). The InterLACE study: Design, data harmonization and characteristics across 20 studies on women's health. Maturitas, 92, 176–185. doi:10.1016/j.maturitas.2016.07.021 [PubMed: 27621257]

Moreno LA, Gonzalez-Gross M, Kersting M, Molnar D, de Henauw S, Beghin L, … Group, H. S. (2008). Assessing, understanding and modifying nutritional status, eating habits and physical activity in European adolescents: the HELENA (Healthy Lifestyle in Europe by Nutrition in Adolescence) Study. Public Health Nutr, 11(3), 288–299. doi:10.1017/S1368980007000535 [PubMed: 17617932]

Moskal A, Pisa PT, Ferrari P, Byrnes G, Freisling H, Boutron-Ruault MC, … Slimani N (2014). Nutrient patterns and their food sources in an International Study Setting: report from the EPIC study. PFoS One, 9(6), e98647. doi:10.1371/journal.pone.0098647

Nanri A, Shimazu T, Ishihara J, Takachi R, Mizoue T, Inoue M, … Group, J. F. V. S. (2012). Reproducibility and validity of dietary patterns assessed by a food frequency questionnaire used in the 5-year follow-up survey of the Japan Public Health Center-Based Prospective Study. J Epidemiol, 22(3), 205–215. doi:10.2188/jea.je20110087 [PubMed: 22343330]

Nourani A, Ayatollahi H, & Dodaran MS (2019). A Review of Clinical Data Management Systems Used in Clinical Trials. Rev Recent Clin Trials, 14(1), 10–23. doi:10.2174/1574887113666180924165230 [PubMed: 30251611]

Ockene IS, Tellez TL, Rosal MC, Reed GW, Mordes J, Merriam PA, … Ma Y (2012). Outcomes of a Latino community-based intervention for the prevention of diabetes: the Lawrence Latino Diabetes Prevention Project. American Journal of Public Health, 102(2), 336–342. doi:10.2105/AJPH.2011.300357 [PubMed: 22390448]

Ohno-Machado L, Sansone SA, Alter G, Fore I, Grethe J, Xu H, … Kim HE (2017). Finding useful data across multiple biomedical data repositories using DataMed. Nat Genet, 49(6), 816–819. doi:10.1038/ng.3864 [PubMed: 28546571]

Oliveira Chaves L, Gomes Domingos AL, Louzada Fernandes D, Ribeiro Cerqueira F, Siqueira-Batista R, & Bressan J (2021). Applicability of machine learning techniques in food intake assessment: A systematic review. Crit Rev Food Sci Nutr, 1–18. doi:10.1080/10408398.2021.1956425

Olstad DL, Poirier K, Naylor PJ, Shearer C, & Kirk SF (2015). Policy outcomes of applying different nutrient profiling systems in recreational sports settings: the case for national harmonization in Canada. Public Health Nutr, 18(12), 2251–2262. doi:10.1017/S1368980014002754 [PubMed: 25471048]

Orfanos P, Naska A, Trichopoulos D, Slimani N, Ferrari P, van Bakel M, … Trichopoulou A (2007). Eating out of home and its correlates in 10 European countries. The European Prospective Investigation into Cancer and Nutrition (EPIC) study. Public Health Nutr, 10(12), 1515–1525. doi:10.1017/S1368980007000171 [PubMed: 17582244]

Pavlovic M, Prentice A, Thorsdottir I, Wolfram G, & Branca F (2007). Challenges in harmonizing energy and nutrient recommendations in Europe. Ann Nutr Metab, 51(2), 108–114. doi:10.1159/000102458 [PubMed: 17489023]

Peterson J, Dwyer J, Adlercreutz H, Scalbert A, Jacques P, & McCullough ML (2010). Dietary lignans: physiology and potential for cardiovascular disease risk reduction. Nutr Rev, 68(10), 571–603. doi:10.1111/j.1753-4887.2010.00319.x [PubMed: 20883417]

Rolland B, Reid S, Stelling D, Warnick G, Thornquist M, Feng Z, & Potter JD (2015). Toward Rigorous Data Harmonization in Cancer Epidemiology Research: One Approach. Am J Epidemiol, 182(12), 1033–1038. doi:10.1093/aje/kwv133 [PubMed: 26589709]

Roman-Vinas B, Ribas Barba L, Ngo J, Martinez-Gonzalez MA, Wijnhoven TM, & Serra-Majem L (2009). Validity of dietary patterns to assess nutrient intake adequacy. Br J Nutr, 101 Suppl 2, S12–20. doi:10.1017/S0007114509990547 [PubMed: 19594960]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Sayon-Orea C, Martinez-Gonzalez MA, Gea A, Flores-Gomez E, Basterra-Gortari FJ, & Bes-Rastrollo M (2014). Consumption of fried foods and risk of metabolic syndrome: the SUN cohort study. Clin Nutr, 33(3), 545–549. doi:10.1016/j.clnu.2013.07.014 [PubMed: 23954218]

Schaap LA, Peeters GM, Dennison EM, Zambon S, Nikolaus T, Sanchez-Martinez M, … group, E. r. (2011). European Project on OSteoArthritis (EPOSA): methodological challenges in harmonization of existing data from five European population-based cohorts on aging. BMC Musculoskelet Disord, 12, 272. doi:10.1186/1471-2474-12-272 [PubMed: 22122831]

Schakel SF, Sievert YA, & Buzzard IM (1988). Sources of data for developing and maintaining a nutrient database. J Am Diet Assoc, 88(10), 1268–1271. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/3171020 [PubMed: 3171020]

Schneider KL, Bodenlos JS, Ma Y, Olendzki B, Oleski J, Merriam P, … Pagoto SL (2008). Design and methods for a randomized clinical trial treating comorbid obesity and major depressive disorder. BMC Psychiatry, 8, 77. doi:10.1186/1471-244X-8-77 [PubMed: 18793398]

Schwedhelm C, Boeing H, Hoffmann G, Aleksandrova K, & Schwingshackl L (2016). Effect of diet on mortality and cancer recurrence among cancer survivors: a systematic review and meta-analysis of cohort studies. Nutr Rev, 74(12), 737–748. doi:10.1093/nutrit/nuw045 [PubMed: 27864535]

Slimani N, Deharveng G, Unwin I, Southgate DA, Vignat J, Skeie G, … Riboli E (2007). The EPIC nutrient database project (ENDB):a first attempt to standardize nutrient databases across the 10 European countries participating in the EPIC study. Eur J Clin Nutr, 61(9), 1037–1056. doi:10.1038/sj.ejcn.1602679 [PubMed: 17375121]

Sluik D, Brouwer-Brolsma EM, Berendsen AAM, Mikkila V, Poppitt SD, Silvestre MP, … Feskens EJM (2019). Protein intake and the incidence of pre-diabetes and diabetes in 4 population-based studies: the PREVIEW project. Am J Clin Nutr, 109(5), 1310–1318. doi:10.1093/ajcn/nqy388 [PubMed: 31051510]

Steinberg D, Bennett GG, & Svetkey L (2017). The DASH Diet, 20 Years Later. JAMA, 317(15), 1529–1530. doi:10.1001/jama.2017.1628 [PubMed: 28278326]

Summer SS, Ollberding NJ, Guy T, Setchell KD, Brown N, & Kalkwarf HJ (2013). Cross-border use of food databases: equivalence of US and Australian databases for macronutrients. J Acad Nutr Diet, 113(10), 1340–1345. doi:10.1016/j.jand.2013.05.021 [PubMed: 23871108]

U.S. Department of Agriculture, Agricultural Research Service. (2019). Retrieved from http://fdc.nal.usda.gov

Venter C, Greenhawt M, Meyer RW, Agostoni C, Reese I, du Toit G, … O'Mahony L (2020). EAACI position paper on diet diversity in pregnancy, infancy and childhood: Novel concepts and implications for studies in allergy and asthma. Allergy, 75(3), 497–523. doi:10.1111/all.14051 [PubMed: 31520486]

Wade TD, Hum RC, & Murphy JR (2011). A Dimensional Bus model for integrating clinical and research data. J Am Med Inform Assoc, 18 Suppl 1, i96–102. doi:10.1136/amiajnl-2011-000339 [PubMed: 21856687]

Wang ML, Gellar L, Nathanson BH, Pbert L, Ma Y, Ockene I, & Rosal MC (2015). Decrease in Glycemic Index Associated with Improved Glycemic Control among Latinos with Type 2 Diabetes. J Acad Nutr Diet, 115(6), 898–906. doi:10.1016/j.jand.2014.10.012 [PubMed: 25547339]

Wang X, Liu L, Fackenthal J, Cummings S, Cook M, Hope K, … Olopade OI (2009). Translational integrity and continuity: personalized biomedical data integration. J Biomed Inform, 42(1), 100–112. doi:10.1016/j.jbi.2008.08.002 [PubMed: 18760382]

Wang X, Williams C, Liu ZH, & Croghan J (2019). Big data management challenges in health research-a literature review. Brief Bioinform, 20(1), 156–167. doi:10.1093/bib/bbx086 [PubMed: 28968677]

Wolfson M, Wallace SE, Masca N, Rowe G, Sheehan NA, Ferretti V, … Burton PR (2010). DataSHIELD: resolving a conflict in contemporary bioscience--performing a pooled analysis of individual-level data without sharing the data. Int J Epidemiol, 39(5), 1372–1382. doi:10.1093/ije/dyq111 [PubMed: 20630989]

Yang C, Ambayo H, Baets B, Kolsteren P, Thanintorn N, Hawwash D, … Lachat C (2019). An Ontology to Standardize Research Output of Nutritional Epidemiology: From Paper-Based Standards to Linked Content. Nutrients, 11(6). doi:10.3390/nu11061300

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

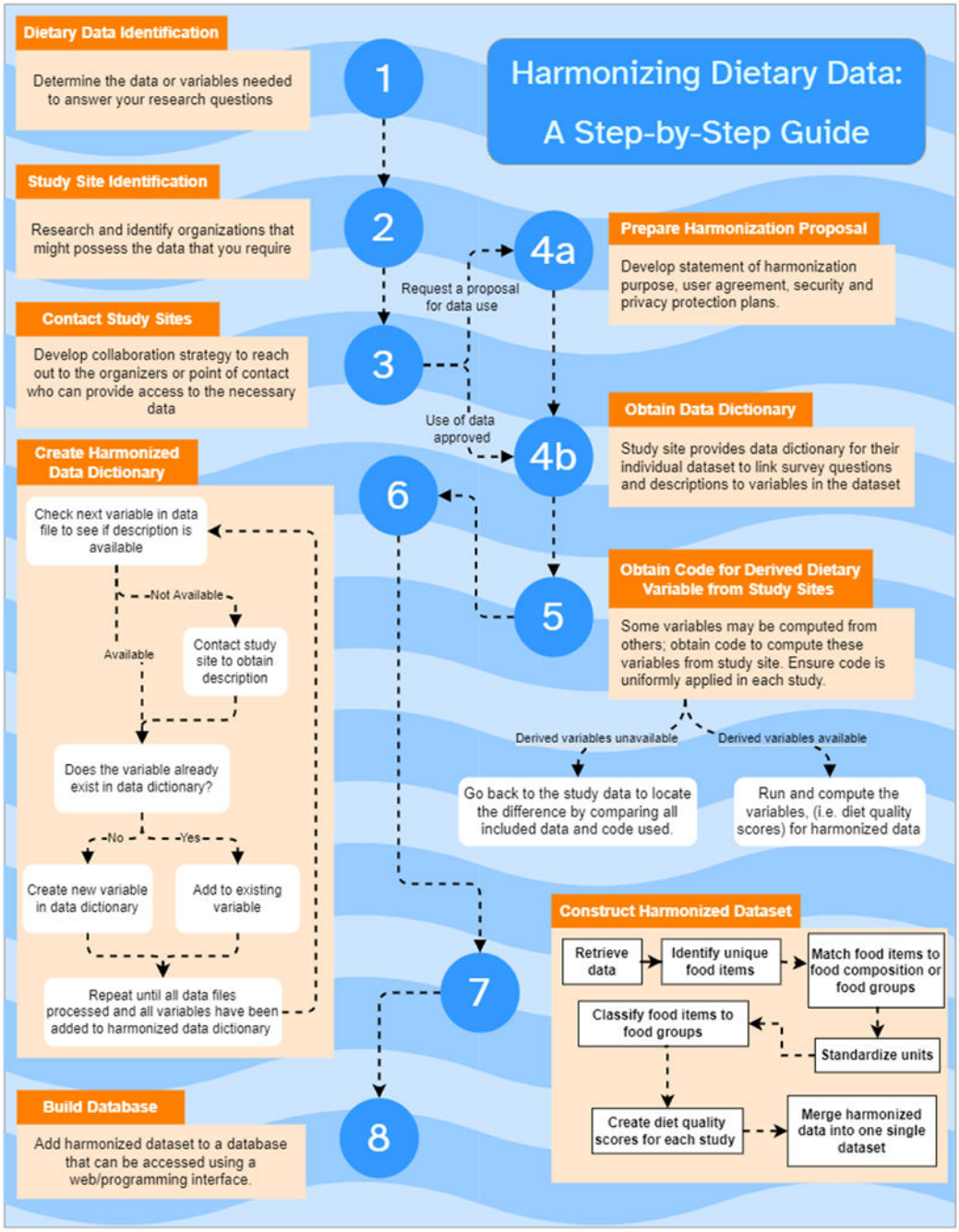

Fig. 1.

Data harmonization process. In steps 1-3, researchers determine where to obtain the data which they need. In step 4, they communicate with study sites to obtain available data. Steps 5-6 describe how to construct a data dictionary, and steps 7--8 describe how to use the data dictionary to create a database of fully harmonized data.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Collaborative harmonization projects for dietary data.

Project*	Country	Description
EuroFIR FoodEXplorer	Finland, Netherlands, United States, Canada, United Kingdom, Japan, and Spain	Extensive, up-to-date database tool which provides detailed data on all common foods and nutrients. Users can browse data from numerous countries through a web browser or download data using the software. This data is suitable for merging with data collected in nutrition studies to analyze the health impact of different nutrients or dietary patterns. FoodEXplorer is extremely useful for data harmonization.
EuroFIR Thesauri	Finland, Netherlands, United States, Canada, United Kingdom, Japan, and Spain	Set of precise definitions for all common foods and their components. Provides a standardized vocabulary for harmonizing food data, allowing researchers to ensure their data collected is compatible with other food datasets.
INFOODS/ EuroFIR	Finland, Netherlands, United States, Canada, United Kingdom, Japan, and Spain	Organization to manage networks of food composition data; ensures data quality and availability across the globe. Performs harmonization of food data from different countries.
COST Action 99	European countries**	Network of compatible food composition databases; ensures the quality of data and compatibility of nutrition data across 32 member countries. Serves to perform harmonization of food composition data.
IARC European Nutrient Data Bank	European countries**	A database of harmonized food and nutrient data across 10 European countries created in 2002. Initial data included 100 nutrients across 1000 foods per country. Compiled using a standard food classification procedure.
EuroFOODS	European countries**	Workshop formed in 1983 to ensure harmonized data across nutrition databases in Europe. Works outline guidelines for nutrient database harmonization.
TEDDY	United States	Study consortium of six clinical centers in the United States and Europe and a data coordinating center to identify environmental factors predisposing to, or protecting against, islet autoimmunity and type 1 diabetes.

* Harmonization project title

** Finland, Netherlands, United Kingdom, and Spain

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Compatibility class is determined variable-by-variable for studies which are candidates to be included in the harmonization process.

Compatibility Class*	**Pairing rules**
Complete	The definition, format, and data collection procedures allow the construction of the variable as described.
Partial	The meaning and the format of the question or questions included in the questionnaire could allow the construction of the variable as described but with an unavoidable loss of information.
Impossible	There is no information or insufficient information in the questionnaire to allow the construction of the variable as described.

*Compatibility class can be used to determine whether the study can be included in the harmonization process. (Fortier et al., 2010)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Recommended harmonization process by Karageorgou et al.

Step *	**Description**
Step 1: Data Retrieval	Identification and retrieval of relevant dietary and sociodemographic variables.
Step 2: Unique food item identification and description	Identification of unique food items (single-ingredient or disaggregated ingredient) across the diet assessment methods by matching their available food description, further accounting for food consumed away from home.
Step 3: Food matching	Matching food items to available food composition data for nutrient profiling.
Step 4: Unit standardization	Accounting for non-edible portions and cooking alterations using yield factors, converting, and reporting in standardized metrics.
Step 5: Food classification	Classifying unique food items to food groups using previously established methods.
Step 6: Individualization of household consumption	Household food and nutrient consumption individualized by the adult male equivalent (AME) [36] and the per capita (P.C.) [37] approach.
Step 7: Final dataset preparation	Merging and creating a complete dataset including individual-level dietary and sociodemographic information.

* Each step is part of the retrospective harmonization procedure for 24-hour dietary recall and household datasets in Karageorgou et al.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4

Comparison of prospective and retrospective harmonization.

	Prospective	**Retrospective**
Guideline Creation	Before data is collected	After data is collected
Schema Purpose	Provides a plan; schema guides data collection to ensure disparate sources are easily combined Supplies core measures for interviews, surveys, and questionnaires	Provides a translation; allows data from different sources to be transformed so that pooling is possible Variables are documented and translated to answer a specific research question
Advantages	More flexible study design, allowing different ways to collect core measures Minimize cost and time due to lower complexity	Allows pooling of data already collected; faster and cheaper than collecting new data Can combine results of disparate studies that share a common purpose
Disadvantages	Cannot incorporate the same measures collected previously or in other studies More expensive and time-consuming to collect additional data instead of combining previous datasets	More complex requires expert domain knowledge to pool data Some data may not be comparable due to variations in available variables across cohorts or heterogeneity in measures If data are not comparable, the risk of data loss
Pertinent Resources	Following collaborative project guidelines such as EuroFOODS or OBEDIS allows the combination of data from different studies Global coding manuals provide insight into optimal questionnaire construction	Wiki websites allow documentation of how variables are defined in different studies Databases and informatics tools allow data to be easily stored and managed
Other Requirements	Use informatics and domain knowledge to select which elements are required to be collected in which manner May seek to follow criteria from other similar studies for maximum comparability	The process must identify relevant questions, variables needed, data available, and possible mappings of data points between sets Can also translate all data into generic format if possible
Example for Dietary Data	The use of standardized disease definitions allow disparate questionnaires to gather common dietary risk factors for disease	Food definitions constantly change over time; can use common food definitions to pool data from different time periods after data collection