Environ Health Perspect, DOI: 10.1289/ehp15410, Accepted 4/15/2025

Unlocking the Power of Data Harmonization in Environmental Health Sciences: A Comprehensive Exploration of Significance, Use Cases, and Recommendations for Standardization Efforts

Authors

Jeanette A. Stingone,[1] HC Bledsoe,[2] Grace Cooney,[2] Mireya Diaz-Insua,[3] Elaine Faustman,[4] Karamarie Fecho,[5,6] Ramkiran Gouripeddi,[7] Philip Holmes,[8] David Kaeli,[9] Oswaldo Lozoya,[10] Anna Maria Masci,[11] Hina Narayan,[12] Charles Schmitt,[13] Maria Shatz,[13] Wren Tracy[2]

[1]Department of Epidemiology, Columbia University Mailman School of Public Health, New York, New York, USA
[2]ICF, Reston, Virginia, USA
[3]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, USA
[4]Department of Environmental & Occupational Health Sciences, University of Washington, Seattle, Washington, USA
[5]Copperline Professional Solutions, LLC, Pittsboro, North Carolina, USA
[6]Renaissance Computing Institute, University of North Carolina, Chapel Hill, North Carolina, USA
[7]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA
[8]Department of Physics, Villanova University, Villanova, Pennsylvania, USA
[9]Department of Electrical and Computer Engineering, Northeastern University, Boston, Massachusetts, USA
[10]RTI International, Research Triangle Park, North Carolina, USA
[11]Department of Data Impact and Governance, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA
[12]Milken Institute School of Public Health, The George Washington University, Washington, District of Columbia, USA
[13]Office of Data Science, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA

Address correspondence to Jeanette Stingone, Department of Epidemiology, Columbia University Mailman School of Public Health, 722 West 168th Street, New York, NY 10032 USA. Email: js5406@cumc.columbia.edu

The authors declare they have nothing to disclose.

Abstract

Background: The field of environmental health sciences increasingly demands comprehensive and diverse datasets, particularly in response to emerging research areas such as climate change, mixtures, and exposomics. The data needed to address the complexity of environmental health research questions often extend beyond the boundaries of a single study or data resource. Traditional data management approaches struggle to harmonize the ever-expanding and heterogeneous data sources needed for research in the environmental health sciences. Harmonization may help address this issue as it involves aligning and standardizing various elements of data to allow comprehensive analysis, data pooling and interpretation across studies.

Objectives: The primary objective is to inform researchers about the transformative potential of embracing harmonization methodologies and to motivate contributions to ongoing efforts, thereby fostering advancements.

Methods: Using the Environmental Health Language Collaborative's Data Harmonization Use Case, we provide a practical illustration of existing data harmonization approaches, identify gaps, and emphasize future research and application directions. We selected two publicly available environmental epidemiology studies on the topic of childhood asthma and three studies on the topic of biomarkers of metals exposure during pregnancy and birth outcomes and applied several existing harmonization approaches to assess interoperability.

Discussion: Our process revealed the potential limitations of many existing harmonization approaches, with notable failures to identify common variables across independent datasets and lack of agreement between human and computer-based approaches. This use case identified various challenges with existing approaches, including reliance on often incomplete data documentation and large amounts of manual effort. To address these challenges, we recommend the continued advancement and dissemination of community data standards, the development of software and tools to facilitate harmonization through automation, and strategic efforts to promote engagement in data harmonization within the environmental health sciences community. Collaborative science is needed to advance our understanding of environmental contributors to health, and realizing the harmonization potential of our scientific data is a step toward improved collaboration.

1. Introduction

The advent and promise of "big data" in recent decades have increased the emphasis on collaborative research, particularly the importance of intra- and inter-research area data sharing and using existing datasets to their full potential with meta-analyses.[1] Environmental health data generated from multiple studies and organizations, such as healthcare, environmental research organizations, and regulatory agencies, represent a valuable resource for researchers to use to address complex environmental health challenges.[2] These diverse data streams offer larger sample sizes, greater exposure variability, and concomitant statistical power.

The integration of bigger, more varied datasets from environmental and public health fields of study can enable robust evaluations of diverse environmental exposures, identification of high-risk population subgroups, building better predictive models, and comparative geospatial analyses.[3]

Data harmonization is the practice of "reconciling various types, levels and sources of data in formats that are compatible and comparable, and thus useful for better decision-making" or analysis.[4] It involves aligning and standardizing various elements of data, including methods, measurements, units, data elements, and other critical aspects necessary for comprehensive data analysis and interpretation. Using analytic tools on these harmonized datasets may yield advantages over analyzing individual studies or data silos, such as increased speed, accuracy, and generalizability of analyses.[5] Harmonization is a transformation that can be applied to existing datasets of individual studies or to data at its source of generation.

This transformation can enable subsequent integration, either of data across studies or of data sources, such as electronic health record data across hospitals. In addition to consideration of data structures, the ethical, socio-legal, and cultural contexts associated with individual datasets and sources further affect their potential to be shared or pooled for secondary analyses and thus must be represented and harmonized to common terms for comparability.[6] Data, referring to the raw information collected or generated during research, encompasses a wide array of variables and parameters essential for scientific inquiry within the environmental health sciences (EHS). Metadata, on the other hand, provides essential contextual information about data, including their origin, structure, format, and provenance, facilitating their interpretation and integration across different studies.[7] While this seminar will focus on the practical aspects of data harmonization, it is also essential to consider that harmonization does not mandate data integration or pooling. Pooling data from different studies involves not only aligning data structures but also determining whether combining and analyzing the data – as if it were similar – is scientifically valid.[8] These scientific considerations include comparability of underlying populations and study design as well as characteristics of individual variables, for example, ensuring comparable measurement approaches. While integrating data requires harmonization, we argue there is value in harmonizing data to common standards, even without immediate plans for integration, to promote interoperability and future research.

There are two main approaches to harmonization, prospective and retrospective.[9] Prospective harmonization involves planning and implementing harmonization strategies before data collection begins. This approach can include using standardized data collection tools, developing common protocols, and ensuring consistency in methods and measurements across individual studies. Retrospective harmonization, on the other hand, involves harmonizing data after collection. This approach may involve re-analyzing data or mapping data elements to a common framework. Moreover, data harmonization may be either stringent or flexible. Stringent harmonization requires all aspects of data collection and analysis to be standardized, ensuring consistency but lacking flexibility. Flexible harmonization, on the other hand, allows different collection and analytic methods as long as the data remains comparably meaningful, offering

more adaptability but often a more time-consuming harmonization process. Prospective approaches enable harmonization with minimal loss of information as plans for interoperability are specified at the onset. Prospective harmonization can be either stringent or flexible. However, retrospective harmonization is flexible by design as it must accommodate differences in data collection and variable construction across independent studies, without prior planning.[10] This may often result in loss of information as harmonization is dictated by the study with the least granular measure. Despite this limitation, the large amounts of extant data mandate the use of retrospective approaches. Thus, both prospective and retrospective approaches play complementary roles in promoting data interoperability and facilitating cross-study analyses within the EHS community.

The current scientific literature for EHS contains many examples of harmonization, both from large, planned consortia and from individual efforts at data pooling.[11-19] These studies have generated new knowledge by increasing sample size and exposure variability. Most have been retrospective in nature, including the ESCAPE (European Study of Cohorts for Air Pollution Effects) program,[20] the Pooled Study of Phthalate Exposure and Preterm Birth,[13] and the MINDMAP project.[14] The methodology used for retrospective harmonization within these studies is often reported as either a completely manual process or a combination of automated and manual processes. For the purposes of this paper, a manual effort or process is one in which key steps and decision-making is performed by a researcher, while an automated process or computational approach is one in which researchers use software to facilitate harmonization. For example, within the Environmental influences on Child Health Outcomes (ECHO) program, retrospective data harmonization efforts for cohorts' existing data include the creation of a common data model and tools for data transformation[21] but also include the role of multi-person teams reviewing, deriving analytic variables, and ensuring accuracy and applicability of the resulting data.[12] There are some examples of prospective harmonization, typically in large, planned consortia or collective research programs. For example, EPIC (European Prospective Investigation into Cancer and Nutrition) and the Canadian Partnership for Tomorrow Project are two ongoing large multicenter prospective cohort studies.[22,23]

Data harmonization can be done through either manual or computational approaches, each characterized by strengths and limitations. While the manual approach allows for nuanced decision-making, computational approaches allow for automatic processing and can make processing more efficient under optimal conditions. Computational approaches can also support larger volumes of data because of their potential for scalability. Open-source tools, schemas, and approaches for harmonization within EHS are also described in the literature, including some designed for both prospective and retrospective harmonization.[12,24-27] For instance, the International Society of Exposure Science (Europe chapter) has provided a platform for preregistration of EHS studies to harmonize the data life cycle and implement Findability, Accessibility, Interoperable, and Reuse (FAIR) guiding principles via its FAIR Environment and Health Registry (FAIREHR) to facilitate the generation of structured, reusable high-quality metadata for effective data integration, interoperability and (re)use.[28] Some additional examples of these include DataHarmonizer, Biolink, and DataSHaPER.[29-31] These tools report varying levels of success in harmonization, and the level of current maintenance and availability is not always clear. These tools often require considerable manual review and effort, even when

software is used for parts of the harmonization process. The time and resources needed to implement these approaches constitute a limiting factor in their broader use by the EHS community.

Given the challenges and complexity of harmonization in EHS studies, the Data Harmonization Use Case Working Group within the Environmental Health Language Collaborative was established with the aim of addressing the challenges faced in conducting harmonization of related studies by generating strategies and resources to create awareness among the EHS community. The goal of this Seminar is to present current approaches to data harmonization through an illustrative use case, identify the gaps in existing harmonization approaches, and present recommendations and future directions to advance harmonization efforts within the EHS community. We focus on environmental epidemiology studies, although many of the themes presented apply to other domains within EHS.[32-36]

2. Use Case Methods

2.1 Data Sources and Methods

The use case sought to present a practical demonstration of the initial stages of the data harmonization process needed to conduct pooled analyses. For this exercise, we obtained publicly available research data from epidemiological cohort studies housed within the Human Health Exposure Analysis Resource (HHEAR) Data Repository. We selected two different research questions to examine how harmonization may vary based on the data types needed for a specific analysis. The first exercise focused on two childhood asthma studies (Study A: 2016-1407[37,38] and Study B: 2016-1450[39,40]). The second exercise focused on three studies examining biomarkers of exposure to metals and birth outcomes within pregnancy cohorts (Study C:2017-1740,[41,42] Study D: 2016-34,[43,44] Study E: 2017-1945[45,46]). Identification numbers for each study were obtained directly from the HHEAR Data Repository. For each study, we obtained a subset of the original data relevant to the research question and the corresponding data dictionary and codebook. These studies vary in their geographic scope, exposure distributions, and time periods. The example studies were selected to demonstrate the practical aspects of harmonization and do not account for the scientific considerations of whether pooling across these studies will result in valid estimates. Data dictionaries for all studies and programming code are posted on our GitHub repository (https://github.com/NIEHS/EHLC_Data_Harmonization/).

2.1.a Mapping Measures

To assess our ability to harmonize the datasets within a case study, we first conducted a manual mapping exercise between pairs of datasets to determine which variables could be identified as equivalent or similar (See Figure 1). In total, four mappings were generated: Asthma: Study A x Study B, Metals and Birth Outcomes: Study C x Study D, Study C x Study E, and Study D x Study E. For each mapping, a matrix of variables from the first dataset with variables from the second dataset was generated. For each pair of data elements, we assigned a category of:

- . = inequivalent concepts (e.g., sex and spirometry reading).

- e = equivalent concept and equivalent coding of the variables. For example, “age” and “age at visit” are marked as equivalent concept and coding (ECC) if the data dictionary indicates the ages are coded the same way and both variables represent the age at visit.
- d = equivalent concept and different coding, recoding without loss of fidelity possible. For example, sex coded as 0/1 or M/F can be harmonized without information loss by recoding.
- el = equivalent concept and coding; however, there may be differences due to measurement approaches between studies (e.g., potential for different protocols for taking blood pressure).
- r = similar concept but a loss of information will occur from mapping the variables due to potential differences in equipment, protocol, and/or measurement. Determination of the category often required manually consulting the study data dictionary and coding documents, as well as inspecting the data values. For example, studies with different income brackets for annual household income can possibly be combined by recoding to less-granular income brackets. As another example, a variable that encodes the frequency of using medicine to control asthma can be combined with a variable that encodes only if medicine is used to control asthma by harmonizing definitions but will result in a loss of the frequency information.

In addition, we sought to understand whether computer techniques, such as natural language processing (NLP) and heuristics (e.g., matching age-related variables), could replicate manual comparisons and ultimately be used to assist humans in mapping exercises. Informed by related work, the authors have found such approaches can significantly improve the speed of curators in extracting information from documents without a drop in precision (i.e., positive predictive value) or recall (i.e., sensitivity).[47] To assess this, we computed mappings between study variables and compared those to the manual mappings in terms of precision and recall. Computed mappings were binary (0 = no match, 1 = match), and the manual mappings were recoded to binary (0 = no match, 1 = any kind of match, codes e, d, el, r).

To compare the two approaches, we computed the number of true positive (TP = number of times the computer and manual mappings both assigned a match), number of false positive (FP = number of times the computer mapping assigned a match and the manual mapping assigned a non-match), and number of false negatives (FN = number of times the computing mapping assigned a non-match and the manual mapping assigned a match). From these results, the precision, recall, and F1 scores were computed (see Table 1 for formulas). A high precision score indicates that the computer and manual approaches agreed when the computer approach indicated a match. A high recall score indicates the computer approach indicated a match when the manual approach indicated a match. The F1 score is a way of averaging the precision and recall, a high score indicates both precision and recall are high whereas a lower score indicates the precision or recall, or both, are lower.

Computed mappings were generated by first converting variable names and descriptions from the data dictionaries to embedding vectors using the OpenAI application programming interface and using the OpenAI text-embedding-3-large model (the top open-access embedding model from OpenAI at the time of writing). Embedding vectors are numerical vectors, often of high dimension (embedding-3-large uses a vector of length 3072 digits) that have the property that semantically related word phrases are nearby in the embedding space. The similarity between

each pair of variables was then computed as the cosine similarity of their embedding vectors. Under this approach, the similarity score between two variables increases as the semantic similarity of their name and description increases. Finally, a threshold was applied to the similarity score to create a binary value indicating a match. The threshold was varied to improve the overall performance of the model by finding the threshold that yielded the highest F1 score. The overall methodology is commonly used to identify similar words, sentences, and documents (see https://openai.com/index/introducing-text-and-code-embeddings/).[48]

The approach to create computer-based mappings is primarily dependent on the method of computing the similarity between two variables. To assess the impact of different similarity methods, we reran the same analysis described here using three additional approaches to computing similarity. The details and results of this exercise are included in supplemental materials.

For these exercises, a set of Python scripts were developed, making use of several commonly used analysis packages including the OpenAI library, Pandas, and Scikit-learn. All code is available on the GitHub repository.

3. Discussion

3.1.a Interpretation of Results

Results from the mapping exercise are provided in Table 2. The total number of paired data elements for each mapping exercise is a direct function of the number of variables in each study. As expected, the vast majority of paired data elements were inequivalent concepts. Among paired data elements that could have been potential matches, the plurality mapping in all four exercises was “r,” suggesting similar concepts but also a potential loss in information that could occur upon pooling due to different operational definitions. Note that a variable from a dataset could be mapped to more than one variable from the other dataset. This was infrequent but occurred primarily due to variables related to treatments, symptoms, and race, wherein multiple variables described some aspect of the same underlying concept. For example, one survey asks for the current asthma control treatment regimen, whereas another survey asks first whether the participant has taken medicine to control asthma and, if yes, asks for information on the regimen. The single question from the first survey maps with loss of information to both questions from the second survey.

The childhood asthma mapping also had a larger proportion of paired elements labeled “el,” indicating an equivalent concept but with potential differences in measurement, equipment, or protocol that could affect the ability to pool the data. For example, the measures “fvc_best” and “spiro_fvc_visit1” appear to be similar and refer to forced expiratory volume, a lung function measurement from spirometry. However, the inclusion of “best” in the first measure name could suggest multiple measures within one study visit, as is common with spirometry. As the second measure name does not include “best” or any other indicator of multiple measurements, alignment of the two variables isn’t fully clear.. Additionally, spirometry protocol and equipment differences may impact the ability to integrate the two variables for harmonization. In such cases, the level of detail in the data dictionaries was insufficient to address the concerns.

Results from our investigation of the use of computational approaches for matching variable names are shown in Table 1. These results show modest and study-dependent agreement between computational approaches to mapping and human mapping decisions. Generally, we see greater agreement with human-generated mappings for the variables within the studies of biomarkers of metal exposure and birth outcomes than we do for the childhood asthma studies. Precision is a positive predictive value, so the Study A x Study B combinatorial approach resulted in only 26% actual matches between variables. The Study C x Study D combinatorial approach resulted in a much higher precision of 80%.

Examination of the mismatches between manual and computer-generated results suggests that our computational approach was too limited in the information available to compute mappings accurately. This assumption is supported by the comparison of four similarity approaches provided in the supplementary materials, which demonstrate that larger and more sophisticated large-language models provide better performance over small and simpler models.

Potentially, both retrospective and prospective harmonization efforts could benefit further from triangulation approaches. Triangulation is the process by which results from different approaches are integrated with the hope of achieving a consensus and avoiding the biases associated with any singular approach or method.[49,50] An example of triangulation approaches as applied to data harmonization in the environmental health sciences would be aligning data from in vitro, animal-, and human-based studies in order to compare the effects of fine particulate matter on lung function across these different scientific investigations. Triangulation can benefit data harmonization both as a means of increasing confidence in harmonization efforts as well as by aligning data generated from different scientific approaches or measurements. Triangulation can be applied to theories, methods, data, or observers, and is often AI-assisted in practice although manual approaches do exist.[51] In the case of harmonization for environmental health sciences, more than one of these types of triangulation efforts could apply, but data and methodological triangulation have clear relevance.[52] An example of helpful triangulation for data harmonization efforts could include leveraging concept matches within variable descriptions, comparing data distributions, and conducting variable name text-matching to all contribute toward the harmonization potential of data sets.[53] Each additional mapping criterion adds confidence to a comparison of data elements. Similarly, different models for harmonization, as described in our supplement, could be applied and harmonized rather than selecting a single best model.

Full results of the mapping exercise can be found on GitHub (https://github.com/NIEHS/EHLC_Data_Harmonization).

3.1.b Insights and Challenges

This practical exercise of implementing existing data harmonization approaches revealed several insights and challenges. Given the number of variables within each study, there were many potential matches and the need to consult data dictionaries for all mappings made the process time-consuming and error prone. Variations in data dictionary entries for similar concepts also made the mapping exercise difficult and would hamper automation attempts. Data dictionaries are often customized for the studies for which they have been developed, and there is generally a lack of common structure across data dictionaries used in EHS. Review of

the literature did not reveal existing tools that structure data dictionaries in such a way as to benefit the generalization and automation of harmonization approaches. While the project-specific scope of harmonization may have been ambitious due to the use of totally disparate studies, working toward a generalizable customization of data dictionaries will help to move the field forward.

Mappings labeled “e” (equivalent concept and equivalent coding) or “d” (equivalent concept and different coding) often represented common measures (e.g., participant age, weight, race), but differences in variable names hindered mapping. This is a common issue affecting harmonization applications. For instance, Adhikari et al. report on a harmonization effort of two Canadian pregnancy cohorts and highlight how differences in variable names and uninformative variable names required investigators to delve deeper into variable descriptions to align variables as common as age and body mass index.[54] Common data elements (CDEs),[55] with standard variable names and constructions, are one potential solution that could facilitate harmonization of these types of commonly collected data. Even in retrospective harmonization efforts, aligning each dataset or source to CDEs—rather than harmonizing them directly with one another—can enhance efficiency and facilitate data integration while minimizing information loss, especially when it is not known in advance which datasets will be integrated.

Most paired data elements in the mapping exercise (Table 1) are labeled “el” or “r.” Determining whether and how these variables could be harmonized was difficult from the limited descriptions in the data dictionaries and would require more comprehensive study metadata. In our case, two studies used the same survey instrument, and inclusion of the survey instrument name and use of the survey instrument variable names in the metadata would have allowed rapid mapping. The solution to these problems requires investigative teams with both data science skills and domain-specific knowledge to identify the relevant metadata that would aid in future harmonization efforts. For example, to achieve high levels of harmonization, Fortier et al. describe a very flexible approach that requires extensive and in-depth knowledge of study variables using predefined keywords and systematic pairing rules.[31] Having structured metadata to accompany datasets would enable greater harmonization efficiency and could facilitate automated and semi-automated computational approaches. Thus, environmental health training must expand to include issues of data representation and management, including the FAIR data principles, which stress the importance of comprehensive metadata. Within the US, NIH has already recognized this need, funding a series of supplements to training programs designed to increase knowledge of and experience with FAIR.[56]Our computational approach to mapping showed modest and study-dependent agreement with human mapping. Generally, agreement was higher for the studies focusing on biomarkers of metals exposure and birth outcomes (e.g., Study C x Study D) than for the studies of air pollution and asthma (e.g., Study A x Study B). This could have occurred for several reasons. Measures of childhood asthma include those derived from spirometry and symptom-based questions that can vary from study to study. Birth outcomes, on the other hand, are generally standard due to common reporting on birth records, such as birth weight and gestational age. Improvements in computational approaches are likely as the current approach did not make use of data distributions, data types, or context from the study publications.

Table 3 shows a detailed discussion of manual and computational approaches, but it is important to recognize that, in practice, hybrid approaches that combine aspects of both will

typically be the most fit for purpose. Both manual and computational approaches have benefits and trade-offs, including variable need for skill and resources. In the absence of mature, community-accepted computational tools, each harmonization effort requires the construction of new tools or substantial modification of existing tools. These efforts consume time and resources for code development and testing that would otherwise have been spent on manual harmonization. It is therefore difficult to determine which approach is more resource-effective, and it may be context-dependent. However, as tools are developed, the upfront investment will become less significant and the computational approach more sustainable.

Investigators often seek guidance to identify potential candidate studies ideal for harmonization efforts. As demonstrated by the results of the above exercise, retrospective harmonization requires considerable alignment between studies to achieve success. Thus, most harmonization efforts select studies based on the practical considerations of having commonalities in topic, measurement and populations as well as buy-in from investigators to ensure access to information that may not be clearly recorded in a data dictionary. Pre-existing consortia and multi-investigator initiatives with thoughts of harmonization prior to study completion present potential sources of data that could serve as candidate studies for initial efforts at harmonization and integration. Similar to meta-analyses, a considerable prior literature can often be of benefit for harmonization efforts as it often translates into investigators aligning on covariates, ensuring similar data across studies. Even with combined experience and two very similar datasets, retrospective harmonization is challenging. When fewer variables and/or studies are being retrospectively harmonized, a greater percentage of harmonization with potentially minimal information loss is plausible.[54,57] We and others have shown that retrospective harmonization could generally be successful in 20%–60% of the variables, with availability of metadata playing a large role in increasing the percent of variables that can be mapped across studies/sources. Even with complete metadata, however, the degree of possible harmonization is a function of the granularity of the data originally captured in each study. There are inevitable trade-offs between information content and harmonization level. Using a common data element, such as birth date, captures meaningful detail and minimizes information loss during harmonization (see Figure 2). When collecting data with less-granular derivations, such as life stage, the degree of harmonization potential is limited. The decision to integrate data that can only be harmonized at a less-granular level will depend on the scientific considerations specific to the research question or application. In certain contexts, information loss or potential for misclassification due to differences in data granularity or measurement approaches may be acceptable. But, in general, we advocate for prospective harmonization and the use of CDEs that possess high degrees of granularity in data collection whenever possible to facilitate interoperability and data reuse for future research.

Prospective harmonization with CDEs offers the advantage of collecting the data across cohorts in a unified fashion from the beginning. This does not necessarily decrease the amount of work involved, as effort is transferred from the post-data collection period for retrospective harmonization to the period prior to and during data collection for prospective harmonization. This also transfers the effort to expert members of the study team. What prospective harmonization guarantees is the amount of information that we can effectively use, given its homogeneity and quality, and a greater level of confidence that relationships assessed are more likely to be true. On the other hand, although the results are desirable, prospective

harmonization is not without difficulties, and considerations of privacy and resource availability may prevent full implementation. The cost of collection and maintenance for multiple highly detailed measures can be prohibitive. In addition, a lack of community agreement on standards used for prospective harmonization is an obstacle to adherence.

Beyond the data management aspects of harmonization, it is important to recognize that pooling data across studies can lead to bias in certain contexts. While outside the scope of this seminar, evaluation of statistical approaches for analysis of pooled data, including the determination of whether pooling is scientifically appropriate, is an area that requires additional research and guidelines for practice. In addition, we used deidentified data as the foundation of our use case. The identifiability of data is an important consideration for data integration efforts, however, irrespective of the types of data harmonization efforts used. Therefore, strict adherence to data privacy and security measures like masking, encryption, and access control within the ethico-legal and regulatory frameworks is paramount in preventing reidentification of individuals.[36] Moreover, several new solutions, organizations, and tools are being developed or proposed to address these growing privacy concerns. Some examples of these efforts include the Federated Data Consortium model, DataSHIELD technology, and proposals for coherent data access agreements and applications of machine learning algorithms for anonymity.[32,33]

3.2 Core Requirements for Data Harmonization

Some level of automation is necessary for data harmonization, particularly for extremely large datasets that are challenging or impossible to manually harmonize because of the large number of variables present. Automated solutions for data harmonization are dependent on a shared understanding of guiding principles for harmonization (Table 4) and community-endorsed standards. Our observations during the data harmonization exercise have also allowed us to identify a core set of community and technical requirements that any solution for data standards and harmonization must meet to promote convenient and cost-effective interoperability in EHS.

3.2.a Community Engagement

At the community level, any effort toward standards for data harmonization must have well-defined goals and objectives to be successful. Clearly defined and transparent goals and objectives ensure that all involved parties and stakeholders are working toward a common purpose and have a common understanding of what needs to be achieved in any harmonization effort. Buy-in from all relevant stakeholders, including investigators, data providers, data users, subject matter experts, equipment vendors, and funders, is critical to establish at the very beginning of any data harmonization effort, whether retrospective or prospective. Appropriately designed standards can be used for either retrospective or prospective initiatives. Standards should be coupled with best practices in open-source software development, such as shared public GitHub repositories, use of software collaboration platforms (e.g., Jupyter Colab), code review and approval protocols, thorough documentation, changelogs, versioning, etc.

Harmonization efforts are improved by collective effort/contribution. This will not only allow an individual project to benefit from the collective expertise of a broader group, it also will motivate and propagate broad interest in contributing to data harmonization efforts. The benefits of

engaging the broad EHS community include sharing best practices, providing feedback and suggestions, and collaborating on open-source tool and resource development. We propose that the EHLC can serve as a hub for fostering community engagement in data harmonization efforts within the EHS community.

3.2.b Technical Requirements and Tool Development

On a technical level, perhaps the most critical need in EHS data harmonization is human- and machine-readable metadata specifications tailored to different domains within EHS, such as toxicology, field epidemiology, and exposomics. Metadata specifications, like those found in a data dictionary, would provide a guide on what metadata needs to be available for collecting and reporting data within a particular domain, along with linkages to allowable choices of semantic and syntactic standards. For example, sensor metadata specifications would consist of instrument, deployment, and output domains and provide a guide for standardizing, harmonizing, storing, and integrating sensor data with other types of EHS data.[58,59] The metadata specifications can be stored in open graph-based metadata repositories , accessible via application programming interfaces, for study team consumption when creating, harmonizing, evaluating, and versioning various data collections. Within a data dictionary, relevant metadata should include variable names, units, allowable values, concept numbers, and instrumentation. There are several tools that attempt to provide standalone metadata management services that can assist in building data dictionaries, such as CEDAR, CDE repository, PhenX Toolkit, NCI Thesaurus, and caDSR II.[55,60-63] While CEDAR focuses mostly on experimental metadata, the rest of the tools concentrate on the clinical domain. The lack of tools to build data dictionaries specific to the needs of the EHS community is a gap that needs to be filled. Moreover, metadata specifications, when made openly available, will help to ensure that EHS datasets abide by the FAIR principles.

As another example of tool development for automating harmonization steps, Feric et al. illustrate an open-source tool for online retrospective data harmonization.[21] The tool offers a range of rules to harmonize typical variable types and coding definitions. Particularly useful is the tool's support for harmonization of biomarkers by providing a set of libraries that allow adjustments via modeling. Another nice feature of the tool is the summary reports that display the continuous and categorical features of the harmonized datasets. Although they caution that the current tool is configured for their study, the user has the possibility of fine tuning the tool to their own needs. To do this, the user will need to edit the data model, the data adapters (i.e., the mapping functionality), and possibly the visualization functions. Open-source tools shared through GitHub and other repositories in turn support the community efforts described above and are strongly encouraged over siloed software.

The adoption of CDEs[55] provides a socio-technical solution to address differences in the representation of common measures (e.g., age) across EHS studies. For electronic health record data, the adoption of a common clinical data model (CDM) such as HL7®'s Fast Healthcare Interoperability Resources (FHIR®) CDM[64,65] or OHDSI's Observational Medical Outcomes Partnership (OMOP) CDM[66] provides another solution to standardize data representation across study elements. While the choice of CDM is determined by healthcare

systems rather than researchers, tools are available to map between CDMs,[66] thus facilitating data harmonization across EHS studies. Finally, quality control measures and continuous evaluation should be put in place to ensure that any data harmonization effort is accurate and reliable. Quality control measures may include a combination of manual data quality checks by a person with relevant expertise and automated checks, although we encourage automation for greater efficiency. Tools for automated quality checks and validation testing will need to be created and publicly stored and maintained, along with benchmarks for their performance.

Storage and maintenance of quality control benchmarks through a central repository, perhaps a shared GitHub repository, will facilitate validation of data harmonization efforts by stakeholders. A shared process and informatics platform for continuous evaluation is critical for success in data harmonization efforts involving diverse teams with different goals and available resources. This may involve informally gathering feedback from stakeholders or more formal GitHub pull requests to ensure that all relevant stakeholders and EHS harmonization efforts are aligned with respect to data standardization.

3.3 Example Initiatives for Promoting the Adoption of Community Data Standards and Tools for Data Integration and Harmonization

Several efforts are well positioned as pillars to advance data integration and harmonization. For instance, the National Institutes of Health (NIH) CDE Repository[55] represents an effort to broadly identify and endorse CDE across fields within biomedical science, although the primary focus is on clinical data and population survey instruments.[55] Multiple NIH Institutes and Centers now contribute to the repository, and community (e.g., professional societies) and regulatory bodies (e.g., Clinical Data Interchange Standards Consortium) continue to lead the establishment of data-type–specific standards to support flow cytometry, microarrays, clinical trials, and other survey instruments and study types.

Other broad community initiatives and programmatic efforts, such as the Environmental Health Language Collaborative,[67] the Human Health Exposure Analysis Resource,[68] and the NIH Climate Change and Health Initiative,[69] are working to identify gaps that exist within the CDE initiative and others; support the flow of information within the diverse subdomains of environmental health; and inform funding needs for the development of tools, standards, and protocols.

Within the biomedical “knowledge representation” community, large-scale consortia and programs such as the Biomedical Data Translator Consortium,[70-72] the Monarch Initiative,[73,74] and the Illuminating the Druggable Genome program[75,76] have advanced a community-contributed, open-source, universal data model and schema for representing biomedical knowledge: the Biolink Model.[30] This model can serve as a general framework and approach for linking study-level data elements to high-level ontologies (e.g., Gene Ontology, Human Phenotype Ontology). The Biolink Model community provides a toolkit for human- and machine-readable documentation, metadata, quality control, and tools for normalization across identifier systems. The linkage between study-level data elements and knowledge frameworks also provides a means to conduct data integration, harmonization, and analysis at different levels of

abstraction (e.g., analysis of data based on chemical class instead of specific chemical structure), which is an important consideration given the diversity inherent in environmental health data.

Recent funding opportunities, such as those announced by the NIH Office of Data Science Strategy and the National Institute of Environmental Health Sciences, aim to support the development of tools and standards for environmental health research.[77] As shown in this paper and others,[21,78,79] tools that can semi-automate the creation of machine- and human-readable data dictionaries, as well as data harmonization, are critical. Ideally, such tools will build upon existing data elements and knowledge concepts represented in the CDE Portal and Biolink Model, thus helping to span the gap from data to knowledge.[80]

In addition to funding support, other incentives that promote prospective and retrospective data harmonization are important and obtainable. For example, data citations will increase as the number of datasets containing Digital Object Identifiers (DOIs) increases. Moreover, search services such as the National Library of Medicine's Data Catalog will make it easier to find and reuse datasets and standards.[81] Importantly, tools that allow researchers to find datasets that can be integrated into proposed research will increase the value of both the data assets and the research that uses those assets, as well as stimulate broader collaboration. Recent efforts to support preregistration of studies, such as the FAIR Environmental and Health Registry, will incentivize collaboration.[28] Likewise, standards-based foundational exposomics datasets will incentivize prospective data harmonization.

3.4 Conclusion

The benefits of data harmonization are illustrated by a recent report that investigated phthalate exposure and preterm birth using data pooled from across 16 cohorts.[13] New knowledge is generated when data across different geographies, time periods, and populations are integrated and analyzed.[13] As illustrated in the use case presented here and in related referenced work, the ability to retrospectively harmonize data is sufficiently time-consuming and difficult that the research community should be encouraged and supported in prospectively planning for data harmonization as part of existing and future studies. This seminar highlights several of the considerations in prospective planning, such as the adoption of common data elements, the use of well-described data dictionaries, the implementation of tools that promote reuse of common surveys and data elements, the establishment of linkages to ontologies, and engagement with communities to ensure that data collection efforts lead to products that can be reused and integrated with other study data.

This seminar also highlights the considerable challenges the community needs to address to fully realize the benefits of data harmonization. There is a need for tools that minimize the burden on researchers for both prospective and retrospective harmonization efforts, as well as a need for continued evaluation of the impact, capabilities, and limitations of such tools. Standards for machine-readable data dictionaries, which would support automation of harmonization tasks, do not yet exist. Development of guidance and training materials that support researchers in prospective planning would be beneficial. Finally, given the challenges of harmonization and the diversity of environmental health, the field would benefit from increasing

the number of experts who can serve to advise and support prospective planning efforts, in both the mechanics of harmonization and in identifying scientific opportunities that would stem from and justify the investment.

Acknowledgments: The authors would like to thank the Environmental Health Science Language Collaborative Data Harmonization Working Group members for their time, dedication, and enthusiasm in elaborating community needs, discussing potential solutions, and contributing to the development of efforts to further the goals and abilities of the environmental health sciences community. The authors would also like to thank Bren Ames, Albert Donnay, Jennifer Fostel, Stephanie Holmgren, Kristan Markey, and Paul Whaley for their contributions to the Data Harmonization Working Group's trajectory.

Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Numbers P42ES017198 (DK) and R00ES027022 (JAS); by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Numbers OT2TR003430 (KF) and UM1TR004409 (RG); by the National Institute of Environmental Health Sciences of the National Institutes of Health under Project Number 2P42ES010356-20 (OL); by the National Institutes of Health under Project Number 1OT2OD031940-01 (OL); and by the National Heart Lung and Blood Institute of the National Institutes of Health under Project Numbers 1OT3HL147154-01, 1OT2HL167310, and 1OT2HL156812 (OL). Additional support was provided by the Intramural Research Program of the National Institutes of Health under Project Number ZIA-ES103364 (CS). Data for this project were obtained from the publicly available data in the Human Health Exposure Analysis Resource (HHEAR) Data Repository, supported by the National Institutes of Health under Award Number U2CES026555. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

This work was also supported in part by the National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health under contract GS00Q14OADU417 (Order No. HHSN273201600015U) to ICF. The views expressed are those of the authors and do not necessarily represent the views or policies of NIEHS. Any mention of trade names, products, or services does not imply an endorsement by the U.S. government or NIEHS. NIEHS does not endorse any commercial products, services, or enterprises.

6. References

1. Brook JR, Doiron D, Setton E, Lakerveld J. Centralizing environmental datasets to support (inter)national chronic disease research: Values, challenges, and recommendations. *Environmental Epidemiology*. 2021;5(1)
2. Stieb DM, Boot CR, Turner MC. Promise and pitfalls in the application of big data to occupational and environmental health. *BMC Public Health*. May 9 2017;17(1):372. doi:https://doi.org/10.1186/s12889-017-4286-8
3. Ives C, Pan H, Edwards SW, et al. Linking complex disease and exposure data—insights from an environmental and occupational health study. *J Expo Sci Environ Epidemiol*. 2023/01/01 2023;33(1):12-16. doi:https://doi.org/10.1038/s41370-022-00428-7
4. Cheng C, Messerschmidt L, Bravo I, et al. A general primer for data harmonization. *Scientific Data*. 2024/01/31 2024;11(1):152. doi:https://doi.org/10.1038/s41597-024-02956-3

5. Chung MK, House JS, Akhtari FS, et al. Decoding the exposome: Data science methodologies and implications in exposome-wide association studies (ExWASs). *Exposome*. 2024;4(1):osae001. doi:https://doi.org/10.1093/exposome/osae001
6. Chow S-M, Nahum-Shani I, Baker JT, et al. The ILHBN: Challenges, opportunities, and solutions from harmonizing data under heterogeneous study designs, target populations, and measurement protocols. *Transl Behav Med*. 2023;13(1):7-16. doi:https://doi.org/10.1093/tbm/ibac069
7. Teradata. *What is data harmonization?* 2024. Accessed November 12, 2024. https://www.teradata.com/insights/data-platform/what-is-data-harmonization
8. Bravata DM, Olkin I. Simple pooling versus combining in meta-analysis. *Eval Health Prof*. 2001/06/01 2001;24(2):218-230. doi:https://doi.org/10.1177/01632780122034885
9. Fortier I, Burton PR, Robson PJ, et al. Quality, quantity and harmony: The DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol*. Oct 2010;39(5):1383-93. doi:https://doi.org/10.1093/ije/dyq139
10. Fortier I, Doiron D, Burton P, Raina P. Invited commentary: Consolidating data harmonization—How to obtain quality and applicability? *Am J Epidemiol*. 2011;174(3):261-264. doi:https://doi.org/10.1093/aje/kwr194
11. Andersen ZJ, Pedersen M, Weinmayr G, et al. Long-term exposure to ambient air pollution and incidence of brain tumor: The European Study of Cohorts for Air Pollution Effects (ESCAPE). *Neuro Oncol*. Feb 19 2018;20(3):420-432. doi:https://doi.org/10.1093/neuonc/nox163
12. Jacobson LP, Parker CB, Cella D, Mroczek DK, Lester BM. Approaches to protocol standardization and data harmonization in the ECHO-wide cohort study. *Pediatr Res*. Feb 16 2024;doi:https://doi.org/10.1038/s41390-024-03039-0
13. Welch BM, Keil AP, Buckley JP, et al. Racial and ethnic disparities in phthalate exposure and preterm birth: A pooled study of sixteen U.S. cohorts. *Environ Health Perspect*. Dec 2023;131(12):127015. doi:https://doi.org/10.1289/ehp12831
14. Wey TW, Doiron D, Wissa R, et al. Overview of retrospective data harmonisation in the MINDMAP project: Process and results. *J Epidemiol Community Health*. May 2021;75(5):433-441. doi:https://doi.org/10.1136/jech-2020-214259
15. Doiron D, Burton P, Marcon Y, et al. Data harmonization and federated analysis of population-based studies: The BioSHaRE project. *Emerg Themes Epidemiol*. Nov 21 2013;10(1):12. doi:https://doi.org/10.1186/1742-7622-10-12
16. Fortier I, Dragieva N, Saliba M, Craig C, Robson PJ. Harmonization of the Health and Risk Factor Questionnaire data of the Canadian Partnership for Tomorrow Project: A descriptive analysis. *CMAJ Open*. Apr-Jun 2019;7(2):E272-e282. doi:https://doi.org/10.9778/cmajo.20180062
17. Miller RL, Rivera J, Lichtiger L, et al. Associations between mitochondrial biomarkers, urban residential exposures and childhood asthma outcomes over 6 months. *Environ Res*. Dec 15 2023;239(Pt 1):117342. doi:https://doi.org/10.1016/j.envres.2023.117342
18. Buckley JP, Engel SM, Braun JM, et al. Prenatal phthalate exposures and body mass index among 4 to 7 year old children: A pooled analysis. *Epidemiology*. May 2016;27(3):449-58. doi:https://doi.org/10.1097/ede.0000000000000436
19. Chen J, Braun D, Christidis T, et al. Long-term exposure to low-level PM2.5 and mortality: Investigation of heterogeneity by harmonizing analyses in large cohort studies in Canada, United States, and Europe. *Environ Health Perspect*. 2023;131(12):127003. doi:https://doi.org/10.1289/EHP12141
20. Beelen R, Hoek G, Vienneau D, et al. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – The ESCAPE project. *Atmos Environ*. 2013/06/01/ 2013;72:10-23. doi:https://doi.org/10.1016/j.atmosenv.2013.02.037

21. Feric Z, Bohm Agostini N, Beene D, et al. A secure and reusable software architecture for supporting online data harmonization. *Proc IEEE Int Conf Big Data*. Dec 2021;2021:2801-2812. doi:https://doi.org/10.1109/bigdata52589.2021.9671538
22. Borugian MJ, Robson P, Fortier I, et al. The Canadian Partnership for Tomorrow Project: Building a pan-Canadian research platform for disease prevention. *CMAJ*. Aug 10 2010;182(11):1197-201. doi:https://doi.org/10.1503/cmaj.091540
23. Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): Study populations and data collection. *Public Health Nutr*. Dec 2002;5(6b):1113-24. doi:https://doi.org/10.1079/phn2002394
24. Ramírez-Andreotta MD, Walls R, Youens-Clark K, et al. Alleviating environmental health disparities through community science and data integration. *Front Sustain Food Syst*. Jun 2021;5doi:https://doi.org/10.3389/fsufs.2021.620470
25. Rashid SM, McCusker JP, Pinheiro P, et al. The Semantic Data Dictionary - An approach for describing and annotating data. *Data Intell*. Fall 2020;2(4):443-486. doi:https://doi.org/10.1162/dint_a_00058
26. Mrdakovic Popic J, Haanes H, Di Carlo C, et al. Tools for harmonized data collection at exposure situations with naturally occurring radioactive materials (NORM). *Environ Int*. May 2023;175:107954. doi:https://doi.org/10.1016/j.envint.2023.107954
27. Fortier I, Raina P, Van den Heuvel ER, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. *Int J Epidemiol*. Feb 1 2017;46(1):103-105. doi:https://doi.org/10.1093/ije/dyw075
28. Zare Jeddi M, Galea KS, Viegas S, et al. FAIR environmental and health registry (FAIREHR)- supporting the science to policy interface and life science research, development and innovation. *Front Toxicol*. 2023;5:1116707. doi:https://doi.org/10.3389/ftox.2023.1116707
29. Gill IS, Griffiths EJ, Dooley D, et al. The DataHarmonizer: A tool for faster data harmonization, validation, aggregation and analysis of pathogen genomics contextual information. *Microbial Genomics*. 2023;9(1)doi:https://doi.org/10.1099/mgen.0.000908
30. Unni DR, Moxon SAT, Bada M, et al. Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clin Transl Sci*. Aug 2022;15(8):1848-1855. doi:https://doi.org/10.1111/cts.13302
31. Fortier I, Doiron D, Little J, et al. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *Int J Epidemiol*. Oct 2011;40(5):1314-28. doi:https://doi.org/10.1093/ije/dyr106
32. Budin-Ljøsne I, Burton P, Isaeva J, et al. DataSHIELD: An ethically robust solution to multiple-site individual-level data analysis. *Public Health Genomics*. 2015;18(2):87-96. doi:https://doi.org/10.1159/000368959
33. World Economic Forum. *Sharing sensitive health data in a federated data consortium model: An eight-step guide*. 2020. https://www3.weforum.org/docs/WEF_Sharing_Sensitive_Health_Data_2020.pdf
34. Saulnier KM, Bujold D, Dyke SOM, et al. Benefits and barriers in the design of harmonized access agreements for international data sharing. *Scientific Data*. 2019/12/02 2019;6(1):297. doi:https://doi.org/10.1038/s41597-019-0310-4
35. Avraam D, Jones E, Burton P. A deterministic approach for protecting privacy in sensitive personal data. *BMC Medical Informatics and Decision Making*. 2022/01/28 2022;22(1):24. doi:https://doi.org/10.1186/s12911-022-01754-4
36. Rujano MA, Boiten J-W, Ohmann C, et al. Sharing sensitive data in life sciences: An overview of centralized and federated approaches. *Brief Bioinform*. 2024;25(4):bbae262. doi:https://doi.org/10.1093/bib/bbae262
37. Phipatanakul W. *Pediatric inner-city environmental exposures at school and home and asthma study. Targeted Lab Analytic Data*. 2016. *HHEAR Data Repository*. Accessed April 24, 2024. https://doi.org/10.36043/1407_118

38. Phipatanakul W. *Pediatric inner-city environmental exposures at school and home and asthma study. Epidemiological Data*. 2016. *HHEAR Data Repository*. Accessed April 24, 2024. https://doi.org/10.36043/2016-1407_EPI_68
39. Liu A. *Denver asthma panel study. Targeted Lab Analytic Data*. 2016. *HHEAR Data Repository*. Accessed April 24, 2024. https://doi.org/10.36043/1450_356
40. Liu A. *Denver asthma panel study. Epidemiological Data*. 2016. *HHEAR Data Repository*. Accessed April 24, 2024. https://doi.org/10.36043/1450_355
41. Cardenas A. *Mitochondrial DNA biomarkers of prenatal metal mixture exposure: Intergenerational inheritance and infant growth. Targeted Lab Analytic Data*. 2017. *HHEAR Data Repository*. Accessed April 24, 2024. https://doi.org/10.36043/1740_225
42. Cardenas A. *Mitochondrial DNA biomarkers of prenatal metal mixture exposure: Intergenerational inheritance and infant growth. Epidemiological Data*. 2017. *HHEAR Data Repository*. Accessed April 24, 2024. https://doi.org/10.36043/2017-1740_EPI_58
43. Christiani D. *Relating metals exposure to birth and early childhood outcomes via the metabotype of cord blood. Targeted Lab Analytic Data*. 2016. *HHEAR Data Repository*. Accessed May 1, 2024. https://doi.org/10.36043/34_732
44. Christiani D. *Relating metals exposure to birth and early childhood outcomes via the metabotype of cord blood. Epidemiological Data*. 2016. *HHEAR Data Repository*. Accessed May 1, 2024. https://doi.org/10.36043/34_94
45. Breton C. *Maternal and developmental risks from environmental and social stressors (MADRES). Targeted Lab Analytic Data*. 2017. *HHEAR Data Repository*. Accessed April 24, 2024. https://doi.org/10.36043/1945_159
46. Breton C. *Maternal and developmental risks from environmental and social stressors (MADRES). Epidemiological Data*. 2017. *HHEAR Data Repository*. Accessed April 24, 2024. https://doi.org/10.36043/1945_177
47. Wieder WR, Pierson D, Earl S, et al. SoDaH: The SOils DAta Harmonization database, an open-source synthesis of soil data from research networks, version 1.0. *Earth Syst Sci Data*. 2021;13(5):1843-1854. doi:https://doi.org/10.5194/essd-13-1843-2021
48. Neelakantan A, Weng L, Power B, Jang J. Introducing text and code embeddings. OpenAI. Accessed November 22, 2024, https://openai.com/index/introducing-text-and-code-embeddings/
49. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *Int J Epidemiol*. Dec 1 2016;45(6):1866-1886. doi:https://doi.org/10.1093/ije/dyw314
50. Arias Valencia MM. Principles, scope, and limitations of the methodological triangulation. *Invest Educ Enferm*. Jun 2022;40(2)doi:https://doi.org/10.17533/udea.iee.v40n2e03
51. Liu Y, Gaunt TR. Triangulating evidence in health sciences with Annotated Semantic Queries. *Bioinformatics*. 2024;40(9)doi:https://doi.org/10.1093/bioinformatics/btae519
52. Noble H, Heale R. Triangulation in research, with examples. *Evid Based Nurs*. Jul 2019;22(3):67-68. doi:https://doi.org/10.1136/ebnurs-2019-103145
53. Long RA, Ballard S, Shah S, et al. A new AI-assisted data standard accelerates interoperability in biomedical research. *medRxiv*. 2024;doi:10.1101/2024.10.17.24315618
54. Adhikari K, Patten SB, Patel AB, et al. Data harmonization and data pooling from cohort studies: A practical approach for data management. *Int J Popul Data Sci*. 2021;6(1):1680. doi:https://doi.org/10.23889/ijpds.v6i1.1680
55. National Institute of Health (NIH). NIH CDE Repository. Accessed April 18, 2024. https://cde.nlm.nih.gov/home
56. National Institute of Health (NIH). *About the administrative supplements for workforce development at the interface of information sciences, artificial intelligence and machine learning (AI/ML), and biomedical sciences*. 2021. https://datascience.nih.gov/artificial-intelligence/initiatives/Workforce-Gap-Data-Governance-AI

57. Bauermeister S, Phatak M, Sparks K, et al. Evaluating the harmonisation potential of diverse cohort datasets. *Eur J Epidemiol*. Jun 2023;38(6):605-615. doi:https://doi.org/10.1007/s10654-023-00997-3
58. Burnett N, Gouripeddi R, Facelli J, et al. Development of sensor metadata library for exposomic studies. *ISEE Conference Abstracts*. 2018;2018(1)doi:https://doi.org/10.1289/isesisee.2018.P01.0320
59. Gouripeddi R, Lundrigan P, Kasera S, et al. Exposure health informatics ecosystem. *Total Exposure Health*. Taylor & Francis; 2020:47:chap 16.
60. Musen MA, Bean CA, Cheung K-H, et al. The center for expanded data annotation and retrieval. *Journal of the American Medical Informatics Association*. 2015;22(6):1148-1152. doi:https://doi.org/10.1093/jamia/ocv048
61. Hamilton CM, Strader LC, Pratt JG, et al. The PhenX Toolkit: Get the most from your measures. *Am J Epidemiol*. Aug 1 2011;174(3):253-60. doi:https://doi.org/10.1093/aje/kwr193
62. National Cancer Institute (NCI). caDSR II. https://cadsr.cancer.gov/onedata/Home.jsp
63. National Cancer Institute (NCI). NCI thesaurus. Accessed November 22, 2024, https://ncit.nci.nih.gov/ncitbrowser/
64. FHIR Foundation. *HL7 FHIR Foundation: Enabling health interoperability through FHIR*. 2024. Accessed November 12, 2024. https://fhir.org/
65. The Office of the National Coordinator for Health Information Technology. *What Is HL7® FHIR®?* . https://www.healthit.gov/sites/default/files/page/2021-04/What%20Is%20FHIR%20Fact%20Sheet.pdf
66. Observational Health Data Sciences and Informatics (OHDSI). *Standardized data: The OMOP common data model*. 2024. Accessed November 12, 2024. https://www.ohdsi.org/data-standardization/
67. National Institute of Health (NIH). Environmental health language collaborative harmonizing data. Connecting knowledge. Improving health. Updated December 11, 2023. Accessed April 25, 2024, https://www.niehs.nih.gov/research/programs/ehlc
68. The Human Health Exposure Analysis Resource (HHEAR). Human Health Exposure Analysis Resource. Accessed April 25, 2024, https://hhearprogram.org/
69. National Institute of Health (NIH). NIH Climate Change and Health Initiative. Accessed April 25, 2025, https://www.nih.gov/climateandhealth
70. National Institute of Health (NIH). About Biomedical Data Translator. Updated April 22, 2024. Accessed April 25, 2024, https://ncats.nih.gov/research/research-activities/translator/about
71. Toward a universal biomedical data translator. *Clin Transl Sci*. Mar 2019;12(2):86-90. doi:https://doi.org/10.1111/cts.12591
72. Fecho K, Thessen AE, Baranzini SE, et al. Progress toward a universal biomedical data translator. *Clin Transl Sci*. May 25 2022;15(8):1838-47. doi:https://doi.org/10.1111/cts.13301
73. Monarch Initiative. What is Monarch? Accessed April 25, 2024, https://monarchinitiative.org/
74. Putman TE, Schaper K, Matentzoglu N, et al. The Monarch Initiative in 2024: An analytic platform integrating phenotypes, genes and diseases across species. *Nucleic Acids Res*. Jan 5 2024;52(D1):D938-d949. doi:https://doi.org/10.1093/nar/gkad1082
75. National Institute of Health (NIH). Illuminating the druggable genome (IDG). Accessed April 25, 2024, https://ncats.nih.gov/research/research-activities/idg
76. Oprea TI, Bologa CG, Brunak S, et al. Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov*. May 2018;17(5):377. doi:https://doi.org/10.1038/nrd.2018.52
77. National Institute of Health (NIH). *Department of Health and Human Services. Grants.* https://grants.nih.gov/grants/guide/rfa-files/RFA-ES-23-002.html

78. Kapsner LA, Mang JM, Mate S, et al. Linking a consortium-wide data quality assessment tool with the MIRACUM Metadata Repository. *Appl Clin Inform*. Aug 2021;12(4):826-835. doi:https://doi.org/10.1055/s-0041-1733847
79. Doiron D, Marcon Y, Fortier I, Burton P, Ferretti V. Software Application Profile: Opal and Mica: Open-source software solutions for epidemiological data management, harmonization and dissemination. *Int J Epidemiol*. Oct 1 2017;46(5):1372-1378. doi:https://doi.org/10.1093/ije/dyx180
80. Schmitt CP, Cox S, Fecho K, et al. *Scientific discovery in the era of big data: More than the scientific method*. 2015. https://renci.org/wp-content/uploads/2015/11/SCi-Discovery-BigData-FINAL-11.23.15.pdf
81. National Institute of Health (NIH). Dataset catalog beta. National Library of Medicine. Accessed May 2, 2024, https://www.datasetcatalog.nlm.nih.gov/index.html

Tables

Table 1. Comparison of Human and Machine-generated Mappings (Binary Classification).

Study 1[a]	Study 2[a]	Threshold	Precision	Recall	F1	TP	FP	FN
C	D	0.65	0.80	0.73	0.76	8	2	3
E	C	0.55	0.56	0.82	0.67	14	11	3
E	D	0.65	0.67	0.67	0.67	6	3	3
B	A	0.45	0.26	0.60	0.36	12	34	8

Note: Threshold = the threshold yielding the highest F1 score; precision = TP/(TP + FP); recall = TP/TP + FN); F1 = 2(precision * recall)/(precision + recall); TP = true positives; FP = false positives; FN = false negatives.*

[a]Values presented in these columns represent the Human Health Exposure Analysis Resource (HHEAR) repository identification number.

Table 2. Number of Variables Matched to Coding Definitions by Manual Mapping of Variables between Datasets.

	Metals and Birth Outcomes			Childhood Asthma	
Coding Definition	Study D (2016-34) x Study E (2017-1945)	Study C (2017-1740) x Study E (2017-1945)	Study C (2017-1740) x Study D (2016-34)	Study A (2016-1407) x Study B (2016-1450)	Total
Equivalent concept, coding (e)	3	4	2	1	10
Equivalent concept and different coding (d)	1	2	3	2	8
Similar variable concept, different operational concepts (r)	5	9	5	10	29
Equivalent concept of variables, differential methodology (el)	0	2	1	7	10
Total number of paired data elements	752 (47 × 16)	893 (47 × 19)	304 (19 × 16)	7,938 (63 × 126)	

Note: The following coding definitions were applied for the mapping exercise: e = equivalent concept and equivalent coding of variables; d = equivalent concept and different coding of variables; r = similar concept of variables but with the potential to lose information by pooling due to different operational concepts; el = equivalent concept of variables but with potential differences in measurement, equipment, or protocol that could affect data pooling. Study numbers listed are sourced directly from the HHEAR Repository. Because inequivalent concepts were not mappable, they were not included in this table.

Environ Health Perspect, DOI: 10.1289/ehp15410, Accepted 4/15/2025

Table 3. A Comparative Analysis of Manual Curation Versus Computational Approaches in the Context of Data Matching.

Consideration	Manual Approach	Computational Approach
Method	Human-driven process requiring manual effort	Automatic process leveraging computational algorithms
Time efficiency	Time-consuming due to manual review and decision-making	Potentially faster due to automation of matching process
Accuracy	Dependent on human judgment and expertise	Relies on computational algorithms and models
Complexity	Moderate complexity due to manual review and decision-making	High complexity involving data processing and algorithm implementation
Flexibility	Allows for nuanced decision-making based on domain knowledge	Limited flexibility in handling complex matching scenarios
Scalability	Limited scalability for large datasets	Potentially scalable for processing large volumes of data
Resource requirement	Requires human resources with domain expertise; Costs are primarily expert labor and will vary depending on the study size and complexity	Requires computational resources and technical expertise; Costs will also include expert labor and will vary depending on the study size, complexity and availability of existing tools
Suitability for complex data	May struggle with complex matching scenarios	Can handle complex data structures and matching criteria
Potential for improvement	Dependent on experience and training	Can be enhanced through algorithm refinement and optimization
Integration with data standards	Dependent on manual adherence to standards	Can be integrated with existing data standards and ontologies

Table 4. Summary of Guiding Principles for Data Harmonization.

Principles	Benefits of Implementation	Obstacles to Implementation
FAIR Adherence: Alignment with FAIR (Findability, Accessibility, Interoperability, and Reuse) advances data harmonization efforts.	Enables researchers to improve discoverability and reuse of their data, while meeting funding and compliance requirements of NIH and other funding agencies. Strengthens the broader scientific ecosystem, fostering a culture of collaboration and innovation.	Time and resources needed to make data/metadata FAIR can be extensive, particularly in the absence of automated tools.
Standardized Tools and Formats: Use of common data elements (CDEs) and community-recognized tools facilitates both retrospective and prospective harmonization efforts.	Saves time and effort and prevents errors and uncertainty during data harmonization.	Different granularity or units for the same measures in different studies can limit capacity to retrospectively harmonize and integrate data. Variable transformation can lead to large data loss at integration, so that pooled data cannot address questions of interest.
Standardized Metadata Collection: Clear and machine-readable metadata are the foundation of strong harmonization efforts.	Promotes findability and interoperability, even without immediate access to data itself. Enables data privacy by facilitating findability and transparency without initial data release.	Community agreement is required on minimum standards for metadata within and across disparate domains in the environmental health sciences.
Quality Control: Continuous evaluation and adaptation are critical elements of a harmonization pipeline.	Ensures rigorous and reproducible science will be achieved by data harmonization and integration efforts.	Manual data quality checks are inefficient, but a lack of existing software tools limit alternatives.

Figure Captions and Notes

Figure 1:

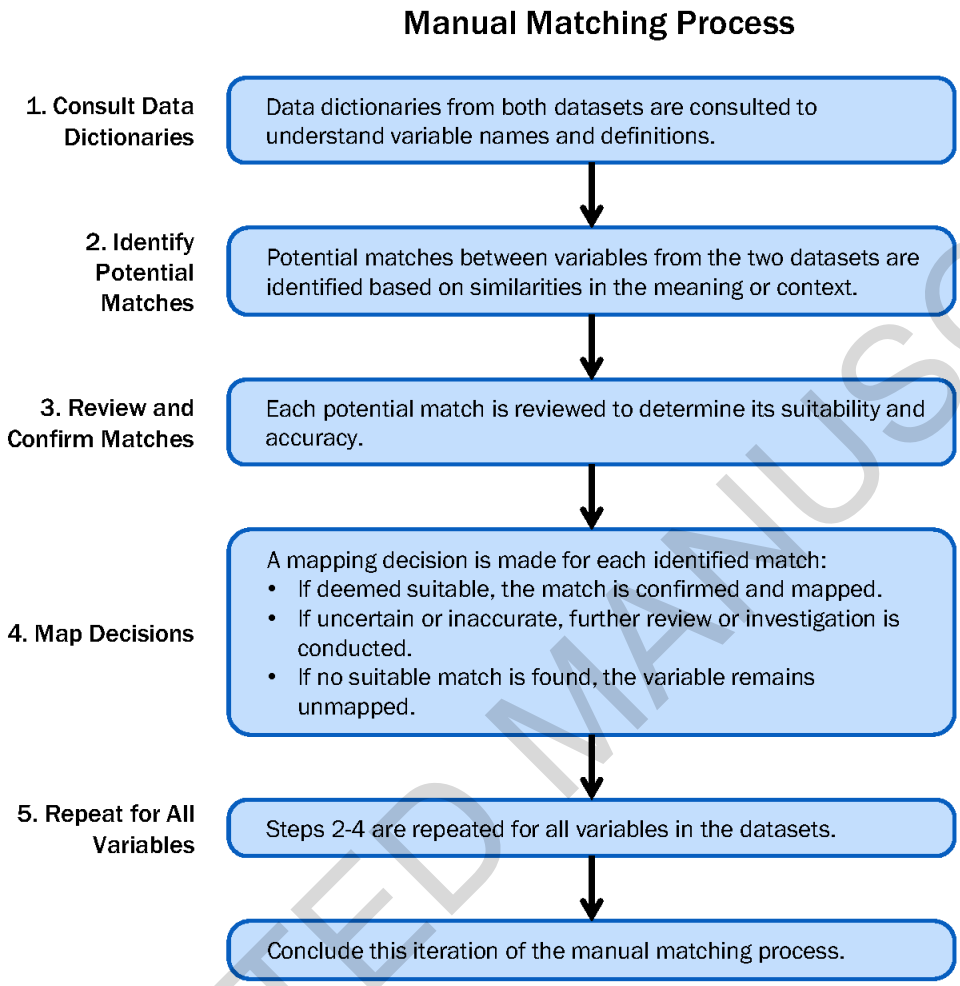

Figure 1 Caption: Flowchart of the Manual Matching Process for Variables between Datasets.

Figure 2:

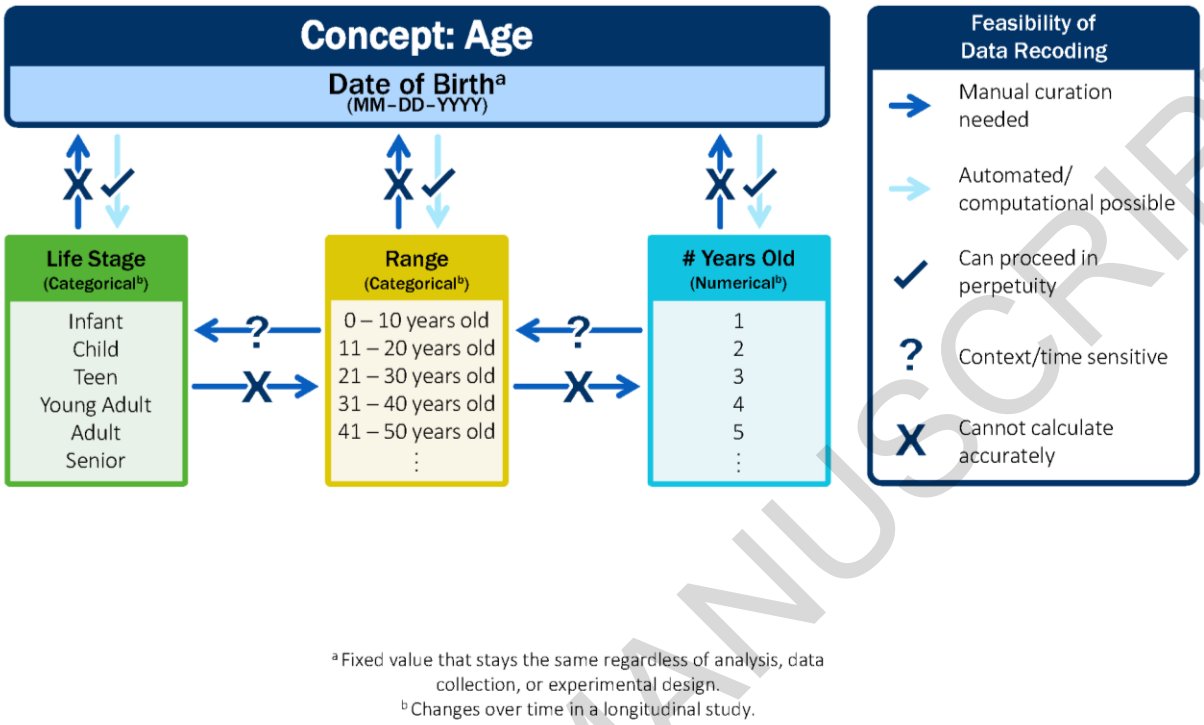

Figure 2 Caption: Data Recoding Feasibility Example Using Date of Birth as a Variable.

Figure 2 Notes:

[a]Fixed value that stays the same regardless of analysis, data collection, or experimental design.

[b]Changes over time in a longitudinal study.