PAN AMERICAN JOURNAL OF PUBLIC HEALTH

Original research

Harmonization and integration of data from prospective cohort studies across the Region of the Americas

Janeil Williams,[1] Olga Tchuvatkina,[2] Marshall K. Tulloch-Reid,[3] Joette McKenzie,[1] Novie Younger-Coleman,[1] Ian Hambleton,[1] Kimlin Ashing,[3] and Camille Ragin[3]

Suggested citation Williams J, Tchuvatkina O, Tulloch-Reid MK, McKenzie J, Younger-Coleman N, Hambleton I, et al. Harmonization and integration of data from prospective cohort studies across the Region of the Americas. Rev Panam Salud Publica. 2025;49:e54. https://doi.org/10.26633/RPSP.2025.54

ABSTRACT

Objectives. To develop a generalizable extraction, transform, and load (ETL) process and workflow for prospective harmonization of data from active cohort studies being conducted in different geographic locations across the Region of the Americas.

Methods. This study harmonized and merged data from two active prospective cohort studies, the Living in Full Health (LIFE) project in Jamaica and the Cancer Prevention Project of Philadelphia (CAP3) in the United States. The RedCAP data collection platform was leveraged in harmonizing and pooling baseline prospective cohort data that was collected from June 2019 to December 2024.

Results. The merged data from this harmonization methodology displayed good coverage on the mapped variables. Seventeen of 23 (74%) of the questionnaire forms harmonized greater than 50% of the variables. Statistical tests on the age-adjusted prevalence of health conditions demonstrated regional differences that could be used to investigate disease hypotheses in the Black Diaspora.

Conclusion. This study developed a successful data harmonization process that can guide similar projects. Active data harmonization is a useful strategy that can reduce costs and leverage resources required to conduct multi-site cohort studies, while fostering data sharing and collaborative research across the Region of the Americas.

keywords Data harmonization; data sharing; data pooling; extraction, transform, and load; cohort studies; Americas.

Population-based prospective cohort studies have been a staple in the epidemiologic research toolbox for understanding the causes, progression, and risk factors for diseases, especially slow-progressing noncommunicable diseases (NCDs) (1). Some of these NCDs, such as cancers and cardiovascular diseases, have a wide range of risk factors (genetic or biologic, lifestyle, environmental) with complex interactions that are difficult to disentangle. Cohort study designs, with comprehensive biomedical and bio-behavioral data from diverse geographic populations, can help researchers understand the underlying risk factors (2). However, due to high implementation costs, few cohort studies are able to collect data in multiple large and diverse populations (3).

By integrating smaller cohort databases, researchers can garner unique insights into chronic disease etiology (4). Pooled data from the individual cohorts increase the statistical power of studies by increasing the sample size of the target population (5). Integrated databases also provide the capability to address questions that otherwise may not have been answered, especially when there is little heterogeneity in an exposure of

[1] Caribbean Institute for Health Research, The University of the West Indies, Kingston, Jamaica.

[2] Cancer Prevention and Control Program, Fox Chase Cancer Center–Temple Health, Philadelphia, United States of America. ✉ Camille.Ragin@fccc.edu

[3] Department of Population Sciences, Beckman Research Institute, City of Hope, United States

interest in a single population, when the exposure of interest is rare or when the interactions between an exposure and environmental factors are being explored (6).

Despite the benefits of pooled data, there are challenges that arise when attempting to integrate databases from multiple cohort studies that collected data in various settings. There is an inherent variability among the questions because they were likely adapted to be population- and/or context-specific or may focus on distinct outcomes. Therefore, attempts to merge data from various study sites can be time consuming and require substantial effort (7–9). To improve efficiency and address some of these challenges, we undertook the development of a data harmonization platform that could provide a structured approach (7).

Data harmonization is the integration of data from two or more separate data sources into a single dataset that can be analyzed (7,9,10). Most attempts at data harmonization use custom tools and widely available software packages (11,12). Generally, these use an extraction, transform, and load (ETL) process, with a prerequisite being the mapping of the different variables. Data harmonization approaches can either be prospective or retrospective based on the time point in which they are conducted. Prospective harmonization occurs before or during data collection, whereas retrospective harmonization occurs after the data is collected (7). Although both approaches can be undertaken, there are few guidelines available, and their cost-effectiveness has not been thoroughly evaluated (13); most have been conducted with cohorts that are completed and not actively recruiting participants.

Given these challenges, we sought to create and test a generalizable ETL process for prospective harmonization of multi-cohort studies. This article describes the development of a harmonization platform comprising two active cohorts in different geographic locations, and through a preliminary analysis, demonstrates the scientific benefits of data harmonization by comparing the prevalence of risk factors and common chronic health conditions between the two data sources.

METHODS

This study was conducted from June 2019 to December 2024, and followed the US National Institutes of Health reporting guidelines. One of the primary aims of the study was to develop a generalizable ETL process that could be used to merge two active cohort studies in different geographic areas: the Living in Full Health (LIFE) project (Jamaica) and the Cancer Prevention Project of Philadelphia (CAP3; United States) (14).

The LIFE project is a partnership between the University of the West Indies (Kingston, Jamaica), the Fox Chase Cancer Center (Philadelphia, US), and the City of Hope Cancer Center (Duarte, US), which form part of the African Caribbean Cancer Consortium team science project. The main objective of the LIFE project is to repurpose and extend the 2016-2017 Jamaica Health and Lifestyle Survey (JHLS-III), which enrolls a nationally representative sample of approximately 3 000 individuals to determine prevalence and risk factors for cancer, coronary vascular disease, and other chronic diseases in Jamaica. The LIFE project enhanced the JHLS-III survey by collecting additional epidemiologic, social, environmental, and medical data, as well as additional biospecimens and body measurements from participants, and by enrolling new volunteers living in the recruitment communities (15). The CAP3 project was a cross-sectional cohort study in the US that enrolled African American/Black participants from Philadelphia (Pennsylvania) and other cities in the states of New Jersey and New York .

The CAP3 consists of ethnically diverse cancer-free participants (US-born Black individuals, and immigrants of African or Caribbean origin) who completed an interviewer-administered questionnaire and donated mouthwash or buccal and urine samples for storage and analysis (14). Using REDCap software (Research Electronic Data Capture, Vanderbilt University; https://projectredcap.org/) and its available Application Programming Interfaces (APIs), we established a secure data integration and sharing platform that enabled linkages between the two cohorts, with an extensible platform and process that can be easily integrated with other regional Caribbean and US cohorts.

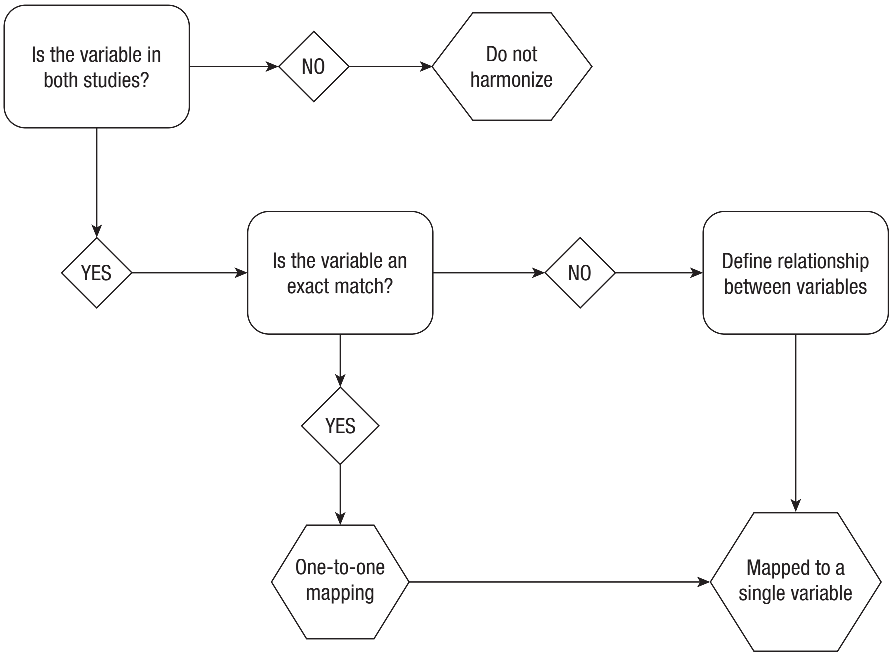
Data mapping and variable harmonization

The first step in the harmonization process involved mapping the variables across both projects onto a single variable. The identification of shared data elements was done prospectively to ensure that the correct variables were included in the data collection instruments at the individual sites. A series of working group sessions were held that focused on selecting the questions that would form the basis of the LIFE study questionnaire. These group sessions included epidemiologists, psychologists, laboratory scientists, and research assistants who were familiar with the instruments from both studies. Given that the objective of the LIFE study was to convert the JHLS-III into a cohort study, attempts were made to retain the questions from the survey instrument that measured critical exposures and outcomes. Survey questions from the CAP3 study, which was already in its data collection phase, were used to collect data on additional exposures that were not already measured in the JHLS-III survey and were to be captured in the LIFE cohort (16). Items from the LIFE questionnaire were grouped into different domains of interest, including demographic factors, medical history, socioeconomic variables, and standard NCD risk factors, such as physical activity, nutrition, developed environment, and occupational exposures. Then, an algorithm was developed to map the variables from the LIFE data source to the CAP3 dataset. The algorithm for mapping all variables is described in Figure 1. First, only variables that were in both studies, and intrinsically represented that same construct, were harmonized. If both variables were collecting the same data and were of the same type, then they were mapped directly onto a single output variable. For variables collecting the same data but the data type differed, a user-defined mapping table was used. This was more common for variables that had different coding for different options. In these instances, the coding used within LIFE data source was recoded to be consistent with that of the CAP3 variable. After the harmonization was completed, the extraction, transform, and loading process was used to pool the data into a single database.

Harmonization and ETL implementation process

Both the LIFE and CAP3 projects used the REDCap application for data management. This platform is a free full-featured and secure web application that is used globally to build data

FIGURE 1. Algorithm used to map variables from LIFE to CAP3 dataset


Is the variable in both studies?
NO
Do not harmonize
YES
Is the variable an exact match?
NO
Define relationship between variables
YES
One-to-one mapping
Mapped to a single variable

Abbreviations: CAP3, the Cancer Prevention Project of Philadelphia (United States); LIFE, the Living in Full Health (LIFE) project (Jamaica).
Source: Prepared by the authors.

collection survey tools, manage survey databases, and provide data quality checks (17). The platform also supports APIs that can be leveraged to interact with the platform. These APIs provided useful functions that were used for custom-built software solutions that processes data externally. The REDCap platform is compliant with HIPAA (the US Health Insurance Portability and Accountability Act), GDPR (the EU General Data Protection Regulation), and FISMA (the US Federal Information Security Management Act). The built-in role-based security features were used to maintain the privacy and confidentiality of the data. Data access was limited to key personnel who were involved in the harmonization process.

The algorithm for the harmonization was implemented using a data collection form within a REDCap project (Figure 2). The source variables from the LIFE project were individually mapped to a destination variable within the CAP3 project; other metadata to guide the recoding of variables were included. The mapping was done by researchers (JM, JW) who had extensive knowledge of the variables within both the LIFE and CAP3 projects. The codebooks for both projects were consulted to guide the process, ensuring that the naming convention and data types were consistent. Additionally, the codebooks were cross-checked to ascertain the coding for single or multiple selection variables. The mapping table output from this process was used to direct the data integration process. This consisted of the source variable, the destination variable, the option values that would be recoded where necessary, and a flag to indicate if the variable would be included in the pooled dataset.

Data pooling and quality assurance

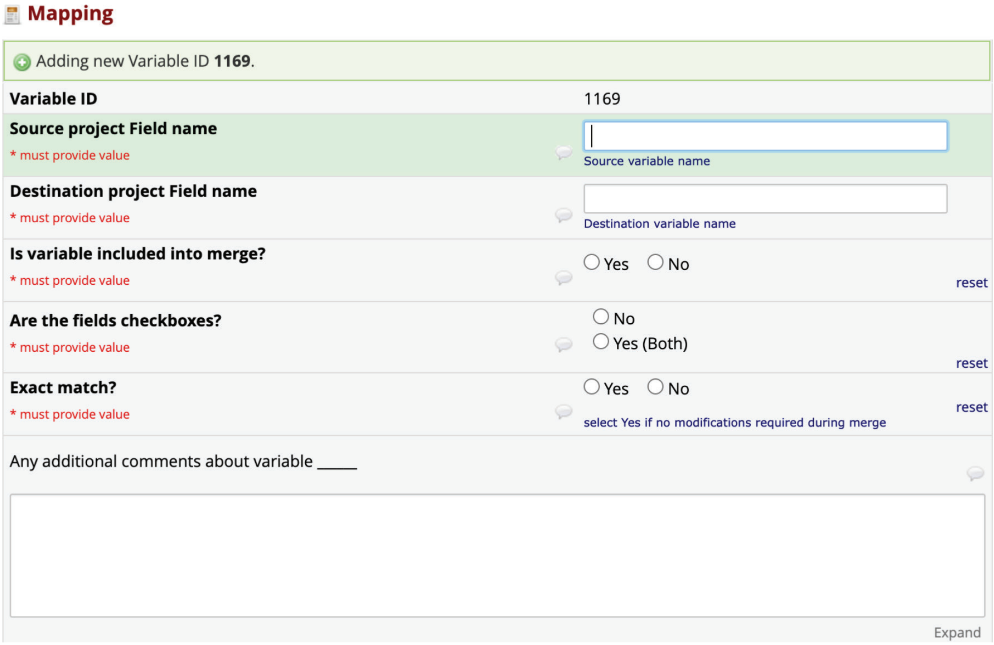
After the mapping was completed, the data from the mapping table was used to create a REDCap project that would serve as the integrated database. Then we developed a custom application (using Java programming language) to routinely download the data for both studies using the requisite API calls. The data were subsequently uploaded to the integrated project. This process was executed weekly on an automated server-side job schedule. Logging information was generated post-execution and indicated the job status, which included the number of records uploaded to the integrated database. Then, the researchers were notified via email for review.

A web application was developed to query the integrated database on the completeness of the data. This web application provided a quantification of the key chronic health exposure variables and their completeness for future research needs. To ensure that the data were consistent, quality checks were routinely conducted on the integrated dataset. A random sample from the integrated database was pulled weekly and cross-checked against the LIFE and CAP3 source data. When data entry errors were noted, these errors were corrected and updated within the source cohort database on REDCap—the corrected values would propagate the merged database on the subsequent run of the automated process; the integrity of the merged database was maintained by preventing direct data entry.

Evaluation of the data harmonization process

To evaluate the data harmonization process and highlight the benefit of data pooling, the coverage of the mapped variables was determined and the prevalences of common chronic health conditions were estimated; sex differences were identified using Pearson χ^2 test of association. Additionally, to highlight the key benefit of the integrated database, the age-adjusted prevalences were estimated and further stratified by the participants' country of origin. Wald test was conducted to identify region-specific differences in prevalences. Stata, version 13 (StataCorp), was used to conduct the analyses.

FIGURE 2. The data collection form used to map variables from the source project (LIFE) to the destination project (CAP3)



Mapping

Adding new Variable ID **1169**.

Variable ID	1169
Source project Field name * must provide value	Source variable name
Destination project Field name * must provide value	Destination variable name
Is variable included into merge? * must provide value	○ Yes ○ No reset
Are the fields checkboxes? * must provide value	○ No ○ Yes (Both) reset
Exact match? * must provide value	○ Yes ○ No select Yes if no modifications required during merge reset

Any additional comments about variable _____

Expand

Abbreviations: CAP3, the Cancer Prevention Project of Philadelphia (United States); LIFE, the Living in Full Health (LIFE) project (Jamaica).

Ethical approval

Both the CAP3 study and the LIFE study were reviewed and approved by the Fox Chase Cancer Center Institutional Review Board. The LIFE study was also reviewed and approved by the University of the West Indies, Mona, Research Ethics Committee. In both studies informed consent was obtained from the participants. Additionally, data sharing agreements were signed between the collaborating institutions.

RESULTS

Harmonization coverage

We successfully implemented a process to harmonize and pool data from the LIFE and the CAP3 cohort studies using custom-built ETL software tools that leveraged the built-in REDCap APIs. The coverage of the mapping, grouped by survey instruments, is described in Table 1. These survey forms grouped similar questions (eg, demographic information). Of 23 forms within the LIFE study, a total of 7 had complete mapping with variables in the CAP3 cohort study. Seventeen of 23 (74%) questionnaire forms had 50% or more of variables mapped; six of 23 (26%) had coverage of less than 50%. A total of 3 188 variables were available within the LIFE study, of which 1 056 (33%) were completely mapped and included in the integrated database. All of the different survey modules within the study were represented in the integrated database.

TABLE 1. Variables within each survey instrument of the LIFE study and the percentage mapped onto corresponding variable within the CAP3 study

Questionnaire sections	Variables, No.	Mapped coverage (%)
Demographic information	166	93 (56)
Tobacco exposure	32	32 (100)
Alcohol use	12	7 (58)
Marijuana use	16	1 (6)
Chronic health conditions	310	108 (35)
Family history of chronic diseases	1860	484 (26)
Female health	37	26 (70)
Male health	70	59 (84)
Colon cancer screening	6	6 (100)
Lung cancer screening	4	4 (100)
HPV screening	4	4 (100)
Nutrition	321	93 (29)
Supplement use	47	25 (53)
Oral health	15	15 (100)
Physical activity	20	17 (85)
Environmental exposures	10	7 (70)
Social environment	18	10 (56)
Health care access	29	29 (100)
Quality of life	33	16 (48)
Religion	15	10 (67)
Medication use	50	46 (92)
Sun exposure	7	7 (100)
Measurement	106	3 (<1)

Abbreviations: CAP3, the Cancer Prevention Project of Philadelphia (United States); HPV, human papillomavirus; LIFE, the Living in Full Health (LIFE) project (Jamaica).
Source: Prepared by the authors from the study results.

Evaluation of the data harmonization—prevalences of chronic health conditions

Table 2 shows the comparison of crude prevalences for seven common chronic diseases in the LIFE and CAP3 cohorts. Diabetes and hypertension were the most common NCDs. In the combined cohort, sex-specific prevalences for diabetes were

8% in males and 19% in females, and for hypertension, 22% in males and 40% in females. The prevalences of the other five conditions ranged from 1% to 3% for both sexes. For the LIFE study, there were statistically significant sex differences in the prevalences of diabetes, dyslipidemia, CVDs, and hypertension. Sex differences were noted in dyslipidemia in the CAP3 study. As shown in Table 2, the combined cohort demonstrated sex differences for all the health conditions, except for CVDs.

The age-adjusted prevalences of each of the health conditions, stratified by migration status, are presented in Table 3. The highest prevalence of diabetes was found among emigrants of Asian and/or European (25.3%), followed by Caribbean emigrants to the United States (23.4%). The lowest prevalence of diabetes was observed among African emigrants (8.9%). For hypertension, the prevalence was highest among Jamaican emigrants (60.1%), followed by Eurasian emigrants (57.2%). The lowest prevalence was observed among African emigrants (39.5%). Dyslipidemia had the highest prevalence among African emigrants, which was estimated to be around 38%. This was followed by other Caribbean emigrants (37.3%). The prevalence of six of the seven health conditions was estimated to be statistically significant to the geographic birth region of the participants.

Evaluation of the coverage of data harmonization

The coverage of the mapping is described in Table 1. When considering the different survey instruments, the majority had mapping of 50% or more. However, the shortfall in coverage was due to two main factors: first, cultural and/or geographic context influenced the addition or deletion of variables. For example, the marijuana use instrument is more extensive in the LIFE study due to the cultural importance of marijuana in Jamaica. Second, differences in the study protocol introduced differences in the data collection instruments. For example, the LIFE cohort study seeks to more comprehensively investigate the risk factors associated with NCDs, and thus, includes a deeper enquiry into the family medical history for several health conditions. Consequently, the differences in the data collection protocols diminished overall coverage in the mapped variables.

DISCUSSION

Understanding the benefits and challenges of data harmonization

This study describes an approach to data harmonization and data pooling using datasets from two comparable, geographically separate cohort studies. This process used expert knowledge, coupled with automated software to execute an ETL process that leveraged the very popular REDCap data collection platform. The advantages of using harmonized data in research are well documented (4,7,8,10). We demonstrated that data pooling allows researchers to conduct comparative analyses between different data sources. As shown, the harmonized

TABLE 2. Self-reported crude prevalence of chronic health conditions by cohort and in the combined dataset, No. (%)

Chronic health condition	LIFE			CAP3			Combined		
	Male (n = 1864)	Female (n = 3225)	*P* value	Male (n = 673)	Female (n = 683)	*P* value	Male (n = 2537)	Female (n = 3908)	*P* value[a]
Cancer	18 (1)	56 (2)	0.168	6 (1)	22 (3)	0.250	24 (1)	78 (2)	0.035
CVD	19 (<1)	59 (<1)	0.023	17 (3)	22 (3)	0.444	36 (1)	81 (1)	0.055
Diabetes	160 (9)	523 (16)	<0.001	51 (8)	53 (8)	0.074	211 (8)	576 (14)	<0.001
Dyslipidemia	132 (7)	585 (18)	<0.001	134 (20)	215 (32)	0.019	266 (11)	800 (20)	<0.001
Hypertension	438 (23)	1412 (44)	<0.001	126 (19)	137 (20)	0.091	558 (22)	1532 (40)	<0.001
Stroke, mini-stroke, transient ischemic attack	48 (3)	94 (3)	0.473	10 (1)	5 (<1)	0.441	58 (2)	99 (3)	0.689

Abbreviations: CAP3, the Cancer Prevention Project of Philadelphia (United States); CVD, cardiovascular disease (including angina, coronary heart disease, myocardial infarction); LIFE, the Living in Full Health (LIFE) project (Jamaica).
[a] *P* values for Pearson χ^2 test of association for country-specific sex differences.
Source: Prepared by the authors from the study results.

TABLE 3. Age-adjusted prevalences of self-reported health conditions among study participants, by country of origin, %

Health condition	Jamaican residents	US immigrants' geographic origin					*P* value[a]
		Jamaica	Caribbean	Africa	Eurasia	Not new immigrants-	
Cancer[b]	1.6	12.6	6.2	—	4.0	5.3	<0.001
CVD	1.5	—	2.5	0.7	3.4	3.6	<0.001
Diabetes	18.0	13.0	23.4	8.9	25.3	16.6	0.312
Dyslipidemia	14.6	33.3	37.3	38.1	26.7	33.7	<0.001
Hypertension	39.9	60.1	43.2	39.5	57.2	47.3	<0.001
Stroke, mini-stroke, transient ischemic attack	2.9	—	0.5	5.7	—	2.9	<0.001

Abbreviations: CAP3, the Cancer Prevention Project of Philadelphia (United States); CVD, cardiovascular disease (including angina, coronary heart disease, myocardial infarction); LIFE, the Living in Full Health (LIFE) project (Jamaica).
[a] *P* values for Wald test of equality of coefficients.
[b] CAP3 study includes participants with a diagnosis more than 5 years before recruitment and not currently receiving treatment.
Source: Prepared by the authors from the study results

data allowed us to explore and compare sex differences in the prevalences within the Jamaican and US cohorts. These differences in crude estimates of disease burden may be associated with differences in risk factors, geographic exposures, or disease awareness affecting men and women differently in these two settings. By further investigating the contribution of the risk factors for different chronic health conditions between the Jamaican-based cohort and the US-based cohort, investigators may also gain insights into the nature and the attributable risks of these exposures related to ethnic and cultural differences. Because the US serves as a major migratory destination for Caribbean nationals, pooling the data may also provide further insights on the effects of regional migratory patterns on the risk of developing chronic diseases, particularly CVDs (18).

Considering the global context, this data harmonizing strategy provides an opportunity for low-resourced countries with limited scientific and technical expertise to undertake robust high-quality research by quickly pooling research data. Furthermore, this harmonization strategy is advantageous given the cost effectiveness, accessibility, and the user-friendliness of the platform. This strategy may be particularly useful to low-income Caribbean countries that have limited capacity to undertake large cohort studies, particularly studies exploring the causes of cancer. Additionally, this approach may allow researchers to adequately investigate health conditions with emergent risk factors that may be unique to Latin America and the Caribbean. Not only will this approach contribute to global health research, but it can inform regional policymakers on potential strategies that can bolster comprehensive health system strengthening.

While harmonization is beneficial to scientific research, several challenges persist that will eventually arise when attempting to harmonize research data (10). The first challenge is to ensure that the data are homogeneous, i.e., each element is capturing the same concepts, irrespective of the variable naming conventions. To overcome this challenge in our study, subject matter experts undertook the entirety of the mapping exercise. This was critical because it ensured data equivalence and maintained the conceptual intent of the variables.

Importantly, the stage at which the harmonization is conducted will affect the quality of the pooled data. In retrospective data harmonization, the harmonization process is conducted with data that have already been collected. The disadvantages in this approach may be that the original intent of the questions has been lost over time, the original dataset (and codebooks) may not be available, and researchers familiar with the original study may not be involved. The alternative approach is prospective data harmonization, which occurs when the process is undertaken before data collection. In our prospective approach, substantial time and effort were required to harmonize data at the beginning of the study. Despite this, the prospective approach guaranteed a higher level of data equivalence because the potential discrepancies were remedied iteratively.

The technical implementation of a data harmonization strategy can itself become a major challenge. However, the primary objectives of the harmonization should guide the implementation process. Given the nature and intent of these cohort studies, special focus was given to variables that are known and/or potential risk factors for chronic health conditions; the mapping of these variables provided researchers with the ability to explore both cohort-specific and comparative analyses on the different risk factors for common chronic diseases. Other factors that may influence the implementation phase of the process may be the available technologies. For example, the availability of the data infrastructure will determine the type of software that is used in harmonization process. Similarly, the location and access to servers for hosting all the cohort databases will impact the ease of accessing the data. This issue may be addressed with a new initiative on open data, which is being initiated by the University of the West Indies in collaboration with the Inter-American Development Bank (19).

In addition to the technical considerations, there is an ethical aspect of this strategy that must be addressed, particularly the privacy and confidentiality of the data. First, institutions that engage in data harmonization and pooling must ensure that all standard ethical approval processes are followed, with a special focus on obtaining informed consent. Institutions must also ensure that the required data infrastructures are compliant with international standards. Lastly, all data sharing agreements must be framed so that all the legal rights of the participants are maintained, irrespective of the geographic jurisdiction.

Limitations

This approach to data harmonization can serve as a model for other Caribbean datasets. There are long-standing concerns regarding a lack of robust study designs for evaluation of many NCDs, including cancer and CVDs, in the Region of the Americas (20). Limited funding and a shortage of human and other resources (statistical expertise, epidemiologists, basic scientists, laboratory facilities) often limit the Region’s ability to conduct large cohort studies. As a result, only a handful of population-based studies have been conducted in the past decade. The LIFE project is one of the two active US National Institutes of Health−sponsored population-based cohort studies investigating the causes of NCDs in the English-speaking Caribbean. Data sharing and harmonization in these and other projects conducted in varied settings in the Caribbean can strengthen our understanding of the epidemiologic mechanisms of these conditions by taking advantage of the genetic, environmental, and cultural diversity that exist, especially those of underrepresented populations. Finally, data harmonization presents an opportunity to drive health research in the Caribbean and wider Region because it fosters data sharing, and ultimately, regional scientific collaboration.

Conclusions

This study demonstrated the ability to create a platform to harmonize and merge prospective cohort data from two different countries in the Region of the Americas, a successful undertaking that will allow researchers to disentangle the effects of location on outcomes of interest, as well as to explore unique exposures on complex diseases. The key advantages of leveraging this harmonization method is its ease of implementation, the relatively low infrastructure and technical requirements of the platform, and its adaptability in undertaking varied collaborative scientific enquiry.

Author contributions. MTR, CR, and KA conceived the original study; JW, JM, and CR collected the data; JM and OT

analyzed the data and performed quality checks; JW wrote and revised the manuscript and interpreted the results. All authors reviewed and approved the final version.

Conflicts of interest. None declared.

Funding. This study was funded by a grant from the US National Institutes of Health (No. R01 MD013347-01 awarded to Professors Tulloch-Reid, Ragin, and Ashing. This study was also supported in part by a grant from the US National Cancer Institute (No. P30 CA006927) and an appropriation from the Commonwealth of Pennsylvania (No. U54 CA221705 and R13 CA284826) to the African Caribbean Cancer Consortium, as well as institutional support from the Cancer Prevention Program at Fox Chase Cancer Center.

Disclaimer. The funders had no role in the study design, data collection or analysis, decision to publish, or preparation of the manuscript. Authors hold sole responsibility for the views expressed in the manuscript, which may not necessarily reflect the opinion or policy of the RPSP/PAJPH and/or the Pan American Health Organization (PAHO).

REFERENCES

1. Szklo M. Population-based cohort studies. Epidemiol Rev. 1998;20(1):81-90. doi:10.1093/oxfordjournals.epirev.a017974
2. Wang X, Kattan MW. Cohort studies: design, analysis, and reporting. Chest. 2020;158(1S):S72-S78. doi:10.1016/j.chest.2020.03.014
3. White E, Hunt JR, Casso D. Exposure measurement in cohort studies: the challenges of prospective data collection. Epidemiol Rev. 1998;20(1):43-56. doi:10.1093/oxfordjournals.epirev.a017971
4. Rolland B, Reid S, Stelling D, et al. Toward rigorous data harmonization in cancer epidemiology research: one approach. Am J Epidemiol. 2015;182(12):1033-1038. doi:10.1093/aje/kwv133
5. Roberts G, Binder D. Analyses based on combining similar information from multiple surveys. Int Stat Rev. 2009;77(1):1-23. doi:10.1111/j.1751-5823.2009.00076.x.
6. Thompson A. Thinking big: large-scale collaborative research in observational epidemiology. Eur J Epidemiol. 2009;24(12):727-731. doi:10.1007/s10654-009-9412-1
7. Cheng C, Messerschmidt L, Bravo I, et al. A general primer for data harmonization. Sci Data. 2024;11(1):152. doi:10.1038/s41597-024-02956-3
8. Fortier I, Wey TW, Bergeron J, et al. Life course of retrospective harmonization initiatives: key elements to consider. J Dev Orig Health Dis. 2023;14(2):190-198. doi:10.1017/S2040174422000460
9. Torres-Espín A, Ferguson AR. Harmonization-information tradeoffs for sharing individual participant data in biomedicine. Harv Data Sci Rev. 2022;4(3). doi:10.1162/99608f92.a9717b34
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):261-266. doi:10.1093/aje/kwr194
11. Mateus P, Moonen J, Beran M, et al. Data harmonization and federated learning for multi-cohort dementia research using the OMOP common data model: a Netherlands consortium of dementia cohorts case study. J Biomed Inform. 2024;155:104661. doi:10.1016/j.jbi.2024.104661
12. Adhikari K, Patten SB, Patel AB, Premji S, Tough S, Letourneau N, et al. Data harmonization and data pooling from cohort studies: a practical approach for data management. Int J Popul Data Sci. 2021;6(1). doi:10.23889/ijpds.v6i1.1443
13. Fortier I, Raina P, Van den Heuvel ER, Griffith LE, Craig C, Saliba M, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int J Epidemiol. 2017;46(1):103–115. doi:10.1093/ije/dyw075
14. Blackman E, Ashing K, Gibbs D, Kuo YM, Andrews A, Ramakodi M, et al. The Cancer Prevention Project of Philadelphia: preliminary findings examining diversity among the African diaspora. Ethn Health. 2021;26(5):659–675. doi:10.1080/13557858.2019.1672890
15. Younger-Coleman N, Webster-Kerr K, Ferguson T, McFarlane S. The Jamaica Health and Lifestyle Survey 2016-17 (JHLS III). Jamaica: Ian Randle Publishers; 2024.
16. Odedina FT, Ragin CC, Martin DN, Moser RP, Oliver JS, McDonald AC, et al. Standardized global behavioral and epidemiological measures for prostate cancer studies in Black men. In: Proceedings of the 2019 Conference on Prostate Cancer Research; 2019. doi:10.1101/203820376
17. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377–381. doi:10.1016/j.jbi.2008.08.010
18. Agyemang C, Van Den Born BJ. Non-communicable diseases in migrants: an expert review. J Travel Med. 2019;26(2). doi:10.1093/jtm/tay107
19. Edmunds A. Open data in the Caribbean and the IDB-financed CaribData project. Accessed 5 May 2025. Available from: https://journal.paho.org/en/file/1601/download?token=iu9pd9o2
20. Williamson GA, Rodrigo S, Blackman E, Ragin CC, Beck JR, Tulloch-Reid MK. An evaluation of regional cardiovascular disease and cancer research needs using conference abstracts. Ann Glob Health. 2021;87(1). doi:10.5334/aogh.3048

Manuscript submitted 20 November 2024. Revised version accepted for publication on 17 March 2025.

Armonización e integración de datos de estudios de cohorte prospectivos de toda la Región de las Américas

RESUMEN

Objetivos. Elaborar un proceso y un flujo de trabajo generalizables de extracción, transformación e inclusión para la armonización prospectiva de los datos de estudios de cohorte activos que se están realizando en diferentes lugares de la Región de las Américas.

Métodos. En el estudio se armonizaron y fusionaron los datos de dos estudios de cohorte prospectivos activos: el proyecto *Living in Full Health* (LIFE) de Jamaica y el proyecto *Cancer Prevention Project of Philadelphia* (CAP3) de Estados Unidos. Se utilizó la plataforma de recopilación de datos RedCAP para armonizar y agrupar los datos de cohorte prospectivos iniciales recopilados entre junio del 2019 y diciembre del 2024.

Resultados. Los datos fusionados obtenidos con esta metodología de armonización mostraron una buena cobertura de las variables mapeadas. En 17 de los 23 formularios del cuestionario (74%) se logró la armonización de más del 50% de las variables. Los análisis estadísticos de la prevalencia de los problemas de salud ajustada según la edad mostraron diferencias regionales que podrían emplearse para investigar hipótesis relacionadas con enfermedades asociadas a la diáspora africana.

Conclusiones. En el presente estudio se llevó a cabo con éxito un proceso de armonización de datos que puede servir de referencia para proyectos similares. La armonización activa de los datos es una estrategia útil que puede reducir los costos y aprovechar los recursos necesarios para realizar estudios de cohorte multicéntricos, además de fomentar la puesta en común de datos y la investigación colaborativa en toda la Región de las Américas.

Palabras clave Armonización de datos; compartición de datos; integración de datos; extracción, transformación y carga; estudios de cohortes; Américas.

Harmonização e integração de dados de estudos de coorte prospectiva na Região das Américas

RESUMO

Objetivos. Desenvolver um processo generalizável de extração, transformação e carregamento (ETL, na sigla em inglês) e um fluxo de trabalho para a harmonização prospectiva de dados de estudos de coorte ativos que estão sendo conduzidos em diferentes localizações geográficas na Região das Américas.

Métodos. Este estudo harmonizou e mesclou dados de dois estudos de coorte prospectiva ativos, o projeto *Living in Full Health* (LIFE), na Jamaica, e o *Cancer Prevention Project of Philadelphia* (CAP3), nos Estados Unidos. Utilizou-se a plataforma de coleta de dados RedCAP para harmonizar e agrupar dados de coorte prospectiva de linha de base que foram coletados de junho de 2019 a dezembro de 2024.

Resultados. Os dados mesclados dessa metodologia de harmonização apresentaram boa cobertura das variáveis mapeadas. Dezessete de 23 (74%) dos formulários do questionário harmonizaram mais de 50% das variáveis. Os testes estatísticos sobre a prevalência ajustada por idade das condições de saúde demonstraram diferenças regionais que poderiam ser usadas para investigar hipóteses de doenças na Diáspora Africana.

Conclusão. Este estudo desenvolveu um processo bem-sucedido de harmonização de dados que pode orientar projetos semelhantes. A harmonização de dados ativos é uma estratégia útil que pode reduzir os custos e angariar recursos necessários para realizar estudos de coorte multicêntricos, ao mesmo tempo em que promove o compartilhamento de dados e a pesquisa colaborativa na Região das Américas.

Palavras-chave Harmonização de dados; compartilhamento de dados; integração de dados; extração, transformação e carga; estudos de coortes; Américas.