PAN AMERICAN JOURNAL OF PUBLIC HEALTH

Special report

Data for population-based health analytics: the Cohorts Consortium of Latin America and the Caribbean

Rodrigo M. Carrillo-Larco[1] and Ian R. Hambleton[2]

Suggested citation Carrillo-Larco RM, Hambleton IR. Data for population-based health analytics: the Cohorts Consortium of Latin America and the Caribbean. Rev Panam Salud Publica. 2024;48:e59. https://doi.org/10.26633/RPSP.2024.59

ABSTRACT

Objective. We describe the daily operations of the Cohorts Consortium of Latin America and the Caribbean (CC-LAC), detailing the resources required and offering tips to Caribbean researchers so this guide can be used to start a data pooling project.

Methods. The CC-LAC began by developing a steering committee – that is, a team of regional experts who guided the project's set up and operations. The Consortium invites investigators who agree to share individual-level data about topics of interest to become members and they then have input into the project's goals and operations; they are also invited to coauthor papers. We used a systematic review methodology to identify investigators with data resources aligned with the project and developed a protocol (i.e. a manual of procedures) to document all aspects of the project's operations.

Results. If a study recruited people from more than one country, then the sample from each country was counted as a separate cohort, thus in 2024 our combined data resources include >30 separate units from 13 countries, with a combined sample size of >174 000 participants. Using this unique resource, we have produced region-specific risk estimates for cardiometabolic risk factors (e.g. anthropometrics) and cardiovascular disease, and we have developed a region-specific cardiovascular risk score for use in clinical settings.

Conclusions. Data pooling projects are less expensive than collecting new data, and they increase the longer-term value and impact of the data that are contributed. Data pooling efforts require systematic and transparent methodology, and expertise in data handling and analytics are prerequisites. Researchers embarking on a data pooling endeavor should understand and be able to meet the various data protection standards stipulated by national data legislation as these standards will likely vary among jurisdictions.

Keywords Dataset; data pooling; big data; data science; epidemiology.

WHAT IS DATA POOLING AND WHY IS IT IMPORTANT?

Data pooling refers to a type of epidemiological study in which data from primary research studies are collated (or pooled), homogenized and analyzed together. This is also referred to as individual-level meta-analysis. By pooling several studies, often dozens or hundreds, data pooling endeavors accrue large statistical power and lower associated statistical uncertainty. This allows research questions to be tackled that one or several studies may not be powered to resolve. Data pooling projects regularly bring together data from different populations, thus increasing variability in the profile of the underlying sample and, therefore, allowing informative results for multiple populations or countries. Perhaps most importantly, but often overlooked, data pooling projects convene many investigators who bring wide-ranging expertise, and this collective knowledge can lead to meaningful insights when applying results to local contexts.

Community-based data collection efforts, especially when longitudinal in design, represent a major investment of time

[1] Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA, United States of America ✉ Rodrigo M. Carrillo-Larco, rmcarri@emory.edu

[2] The University of the West Indies at Cave Hill, Bridgetown, Saint Michael, Barbados

and resources, so the ability to reuse data to address additional targeted research goals provides an opportunity to maximize resources and increase the impact of primary research studies. Data pooling projects can lead to the development of new networks and collaborations and generate new avenues for training students and early-career investigators by providing them with rich data to explore new ideas and to acquire new skill sets, particularly in data science.

Data pooling efforts in Latin America and the Caribbean

In Latin America and the Caribbean, much epidemiological research about cardiometabolic disease has involved collecting primary data from a specific community or population, with a sample size usually of hundreds or a few thousand. There are notable exceptions, such as the CARMELA study in seven countries (1-6) and the CESCAS study in three countries (7), all in Latin America. The dearth of epidemiological evidence from large, population-based longitudinal studies in Latin America and the Caribbean, and particularly in the Caribbean, has been partly driven by limitations in funding and collaborative networks and – to some extent – limited expertise to conduct these studies. To date, population-based community studies have mostly been cross-sectional, which limits their ability to explore causality and long-term effects; in contrast, however, the Cohorts Consortium of Latin America and the Caribbean (CC-LAC) concentrates on developing longitudinal data to fill this critical gap.

Before the CC-LAC was established, during 2018–2019, the only data pooling project in Latin America and the Caribbean collated cross-sectional epidemiological surveys and studies. This was the Latin American Consortium of Studies in Obesity (known as LASO) (8, 9), which was a pioneer in this field in Latin America and the Caribbean and provided important evidence, although the project has now ceased operations. Globally, the NCD (noncommunicable disease) Risk Factor Collaboration (NCD-RisC) has spearheaded the largest data pooling project for cardiometabolic risk factors, also collating data from cross-sectional epidemiological surveys and studies in Latin America and the Caribbean. NCD-RisC produces global (10-15) and region-specific outputs (16). The Global Burden of Disease study, organized by the Institute for Health Metrics and Evaluation, has provided metrics for all diseases and countries, including those in Latin America and the Caribbean (17). NCD-RisC and the Global Burden of Disease studies provide open access health metrics, such as national averages, and prevalence and mortality rates, with repeated cross-sectional evidence running sometimes from as early as the 1970s, and they continue to regularly update their estimates as new studies become available. Depending on the availability of data, it may be that the most recent estimates for some countries are based on statistical modeling (e.g. if there are not new data for a given country but there are new data for other countries) to allow researchers to run inferences for all countries.

Because these global efforts focus on cross-sectional evidence, they provide snapshots in time, but they cannot deliver estimates of disease incidence and risk nor are there prediction tools that can guide risk-based prevention of cardiometabolic diseases in Latin America and the Caribbean. Thus, there was an opportunity to put together the infrastructure to allow data pooling from regional longitudinal studies to advance the regional evidence about cardiometabolic diseases.

The rationale for and origins of the Consortium

The CC-LAC began as a solution to the lack of longitudinal big data on cardiometabolic conditions in Latin America and the Caribbean (18). Although this approach was new to this area, data pooling projects for longitudinal studies already existed in other parts of the world, and these efforts have advanced knowledge about cardiometabolic conditions elsewhere (19-24). In recognition of the geographical diversity of cardiometabolic disease epidemiology, the CC-LAC was conceived to deliver specific evidence about cardiovascular diseases and cardiometabolic risk factors in Latin America and the Caribbean.

It is widely recognized that cardiometabolic conditions have overtaken communicable diseases as the leading causes of morbidity, mortality and disability worldwide and in Latin America and the Caribbean (16, 25-27). Nevertheless, given the heterogeneous profiles observed within Latin America and the Caribbean in terms of demographics, epidemiology and access to health care (28-30), dissecting the distribution, risk factors and consequences of cardiometabolic conditions both in the area alone and in comparison with other world regions, has provided additional insights for regional policy-makers.

In this article we report on the practicalities of developing a regional data pooling project in the field of cardiometabolic diseases. These practical steps can be readily extended to other fields in health and beyond, and data pooling across the area remains an underutilized tool for maximizing the usefulness of individual-level research outputs in clinical medicine and public health.

ESTABLISHING THE CONSORTIUM

Figure 1 summarizes the procedures followed by the CC-LAC. In short, we established a steering committee that helped implement operations. Potential data sources were identified following systematic methods, and those who had collected the data were invited to join the Consortium. Data were cleaned and harmonized for pooled analyses.

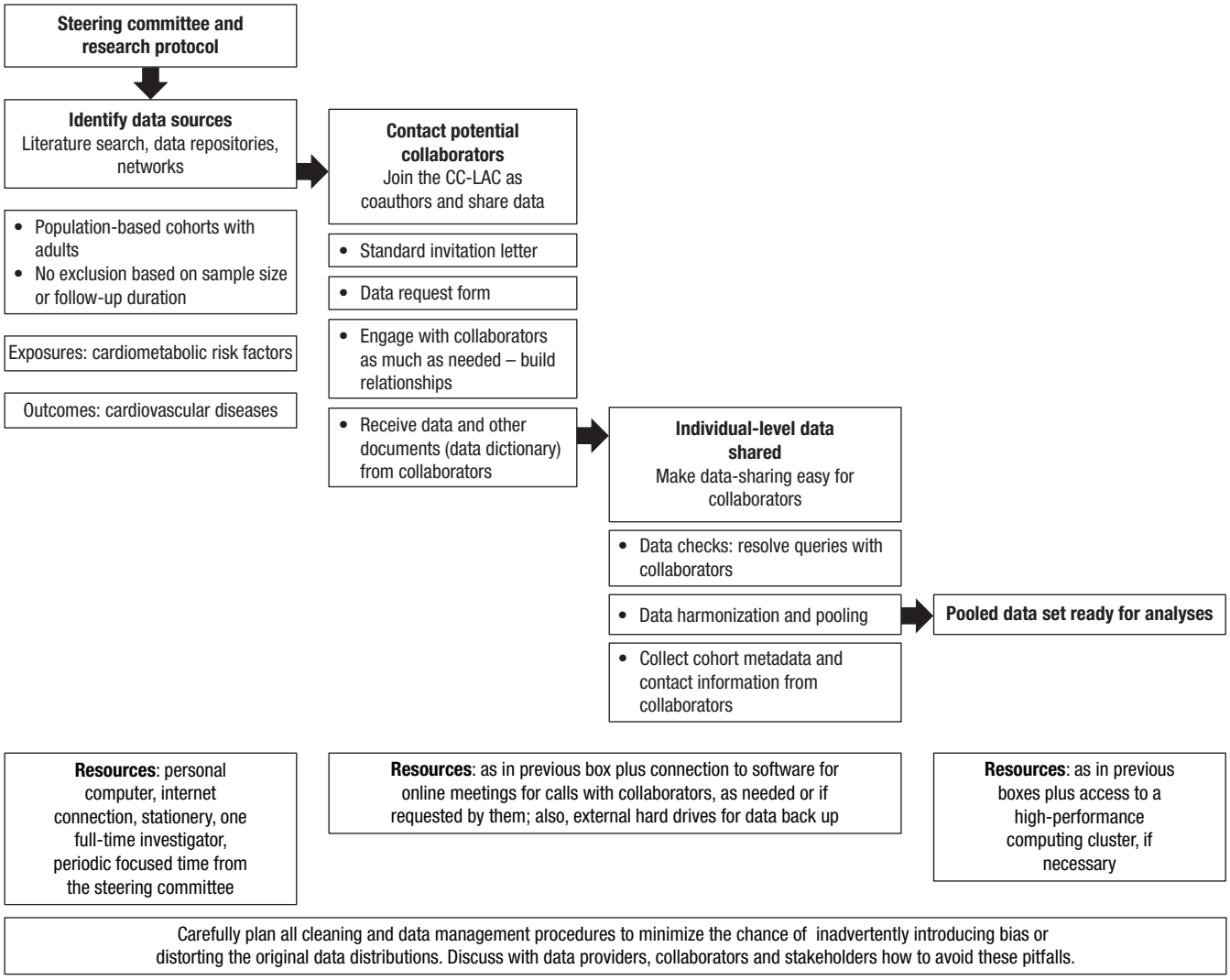
Co-creation of a regional data pooling initiative

The CC-LAC first convened a project steering committee of experts with experience in at least one of three domains: (i) subject-matter expertise in cardiometabolic health; (ii) geographical expertise covering South America, Central America and the Caribbean; or (iii) experience in academia, government or an international health organization, such as the Pan American Health Organization. The steering committee oversees CC-LAC's operations, including guiding the predefined program of analyses and the interpretation of findings. Similar to any epidemiological study, CC-LAC operations are based on a research protocol, which was developed by the members of the steering committee. The key features of the protocol are summarized below.

Defining the regional Consortium

The CC-LAC operates as a virtual consortium. To join the collaboration, teams need to share individual-level data. Teams sharing eligible data with the CC-LAC can include up

FIGURE 1. Schematic representation of the operations of the Cohorts Consortium of Latin America and the Caribbean



Source: Figure prepared by the authors based on their research.

to three investigators who join the Consortium and potentially become coauthors of any scientific output. In exceptional circumstances, teams can include up to five investigators. These circumstances include cases in which (i) the original study closed years ago and substantial work is needed to retrieve and process the data sets; or (ii) the original investigators would be able to refresh their study by linking their records with clinical or vital registration systems, but this process requires substantial work. Teams who become members must be available to answer study-specific questions relating to the data they submit and have the responsibility of commenting on working manuscripts by providing study-specific, subject-specific and location-specific insights as needed. Through these processes, papers authored by the CC-LAC include a broad collection of arguments, ideas and perspectives from across the region.

Types of studies

The CC-LAC focuses on cohort studies evaluating cardiometabolic risk factors, defined as a study that collects data about predefined cardiometabolic risk factors at baseline, with follow up at one or more time points to record a range of established cardiovascular outcomes from the same individual. Studies of any sample size or any follow-up duration are eligible for inclusion.

Types of participants

Cohort studies that included adults (i.e. aged 18 years and older) who were selected using probability sampling from a general underlying population (i.e. population-based cohorts) are eligible for inclusion. This includes cohorts from screening or prevention programs targeting the general population (31) and cohorts recruiting from particular demographic groups, such as the Mexican Teachers' Cohort (32). Studies made up exclusively of people with pre-existing cardiometabolic conditions (e.g. a cohort of people living with diabetes) or cardiovascular diseases (e.g. stroke survivors) or people with a known high-risk profile (e.g. smokers) are not eligible for inclusion in the CC-LAC.

Types of exposure

Data about cardiometabolic risk factors include anthropometric measurements, blood pressure, lipid profiles and biomarkers of diabetes. Data from cohorts were eligible for inclusion if investigators could provide data for one or many cardiometabolic risk factors. The CC-LAC has no exclusion criteria focusing on the method of measuring the risk factors.

Types of outcome measures

The CC-LAC includes cohorts in which fatal and nonfatal cardiovascular outcomes were ascertained at follow up using at least one of three accepted methods: (i) outcomes were confirmed using vital registration systems (e.g. death certificates), (ii) outcomes were adjudicated by qualified physicians or (iii) outcomes were confirmed using health records. If the cardiovascular outcomes were reported by the participant or a third-party informant (i.e. outcomes were not verified by an accepted reliable source), that cohort could be included in the CC-LAC, but the data were not pooled. Seven cohorts were included in this group (i.e. researchers were invited to join but data were not pooled). Although the CC-LAC was initially conceived to focus on cardiometabolic risk factors and cardiovascular outcomes, these cohorts have been included with a view to incorporating their data into future analyses not involving cardiovascular disease.

Methods for identifying studies

Our initial search to identify cohorts for inclusion was guided by two principles: (i) the search must be systematic and (ii) it must include as many studies, countries and populations as possible. The CC-LAC followed a process that closely resembled a systematic literature review.

Electronic searches. First, the CC-LAC defined a literature search strategy and searched MEDLINE, Embase, and SciELO to identify population-based cohorts in studies looking at cardiometabolic conditions in Latin America and the Caribbean. Articles published in any language and at any time were considered. Search terms are available in the published cohort profile (18). The steering committee reviewed the results of the literature search, suggested additional potentially eligible cohorts not identified in the search and helped with communications with cohort investigators.

Other resources. To complement the electronic search, the CC-LAC sought data from publicly available data repositories and used informal "snowballing" enquiries to contact researchers across the area to help identify additional data sources.

Studies and data extraction

In this section we describe in detail how we filtered the search results and contacted potential collaborators. We also describe the process of data-cleaning and give examples of some of the challenges we faced.

Selection of studies. The searches as well as the in-depth evaluation of the search results were conducted by a single full-time investigator in consultation with the steering committee. The search process, including contacting potential collaborators, took around 12 months for the single investigator working full-time on CC-LAC development.

Data extraction. Having identified published articles about eligible cohorts, cohort metadata were extracted, including where the study was conducted, the years of data collection and the sample size. Contact information was instrumental for inviting the investigators to join the CC-LAC and share data, and contact information for the corresponding author was recorded. Contact information for other authors was recorded when possible (e.g. using academic profiles on Google Scholar, university websites and other publications).

Investigator contact and data collation

Data collation followed a standardized and consistent six-step process. This process was repeated for each cohort.

Step 1 – invitation. A standard invitation letter was emailed to the investigators associated with each eligible cohort (Supplementary Figure 1). The invitation letter included information about (i) the rationale and motivations of CC-LAC, (ii) why their cohort was relevant to the Consortium, (iii) how the CC-LAC planned to operate and (iv) the researchers involved in the CC-LAC. If the initial email invitation was not answered within 2 weeks, a first reminder was sent. After two reminders, invitations were sent to other investigators associated with the eligible cohort. The CC-LAC also tried to get in touch with investigators using the steering committee's network of contacts.

Step 2 – data request form. After contact was made, the CC-LAC shared a request form (Supplementary Figure 2). This was a spreadsheet with three tabs: sheet 1 had fields for the investigators to input information such as their names, institutional affiliation and contact details; sheet 2 requested metadata about their cohort and data collection procedures, such as the years of data collection and whether participants fasted before blood tests; and sheet 3 consisted of a list of variables required for data pooling for which investigators recorded additional information. To complete sheet 3, investigators recorded whether their cohort included each requested variable (i.e. yes or no), the name of each variable (i.e. as a safeguard in the event of inadequate labeling of the data set), and the units of measurement, as applicable, recognizing that the units would vary among studies. In addition to the formal request form, the CC-LAC asked investigators to share other documents that could help better understand their data. These included protocols, data dictionaries and other relevant technical documentation and published articles.

Step 3 – additional contact as required. Some investigators requested additional information usually via a teleconference to learn more about the CC-LAC. The CC-LAC team engaged as much as possible with the investigators to build meaningful, trusting and long-lasting relationships.

Step 4 – data sharing. Investigators returned the completed data request form along with their data resources and documents they deemed useful. The investigators for each cohort had the autonomy to decide how to share their data, and data could be transmitted in any format. Most investigators chose to share deidentified data by email using simple open-access formats (such as ASCII files) or common proprietary formats (such as Stata or SPSS files). Collaborators were also given flexibility about how their transmitted data were structured. The goal throughout this step was to minimize the burden on the cohort investigators.

Step 5 – data cleaning. The CC-LAC team inspected in detail and thoroughly cleaned the data submitted from the investigators for each eligible cohort. This process is described in the Data management section and included exhaustive cross-checks with previous publications and shared documents, and with the information provided in the data request form. (Supplementary Table 1 provides a list of data checks worth considering.) If discrepancies were found, the CC-LAC team contacted the cohort investigators to resolve these queries. Investigators received detailed examples, with screenshots of the discrepancies, to ease the process of resolving each inconsistency. The CC-LAC did not consider a data set ready for pooling until all queries had been resolved.

Step 6 – create the database of contacts. The CC-LAC team built and securely maintains a data set containing the contact information of all CC-LAC members (Supplementary Figure 3). This database is used to automatically extract names and affiliations for research papers.

Data management

In preparation for Step 5, the CC-LAC team converted all data sets to a standardized format with standardized names for the variables. All similar measurements were transformed to the same units, and the number and proportion of missing values for each variable were recorded after accounting for questionnaire skip patterns. Finally, a new data set was saved for each eligible cohort that contained only the variables of interest converted to consistent names and units.

A supplementary data set of metadata was also constructed to accompany each cohort data set. Metadata included the study name, the years of data collection, the location of the cohort and a unique identifier for each cohort. These unique identifiers had the following pattern: XXX_YYYY_ZZZZ, where XXX refers to the International Organization for Standardization (ISO)-3 code for the country where the cohort was based, YYYY refers to the baseline year, and ZZZZ is an alphanumeric code for each cohort (e.g. an acronym for the cohort or other name to allow easy identification).

Because the CC-LAC was focusing on cardiometabolic conditions, it included participants' demographics, anthropometric measurements, blood pressure, biomarkers of diabetes, lipid biomarkers, other self-reported health diagnoses and cardiovascular outcomes. Because most of these attributes have a limited range of accepted measurements, harmonization was not particularly challenging for our chosen subject matter. It is likely that other variables (e.g. dietary recall, physical activity measures) would be more time-consuming and challenging to harmonize.

Data analysis

The CC-LAC had two initial goals: (i) to describe associations between cardiometabolic risk factors and cardiovascular outcomes in Latin America and the Caribbean and (ii) to develop a risk score for cardiovascular diseases in the area. The analysis plan for these projects was developed by the steering committee.

For the analysis of risk associations, early decisions were to conduct multivariate imputation by chained equations (i.e. the MICE package in R statistical software, R Core Team, Vienna, Austria) to maximize the available data; to conduct a complete-case analysis for robustness; to compute relative risks, rather than hazard ratios, for comparison purposes because equivalent global metrics were reported as relative risks; and for the same comparison purposes, regression models were age-specific and adjusted by age and sex.

For the analysis of the cardiovascular risk score, we decided to follow the method established by Globorisk (33, 34) because it allows for country-specific recalibration. We conducted a complete-case analysis for the cardiovascular risk score. To maintain consistency with the Globorisk methodology and also because we acknowledge the limited availability of data about blood biomarkers in rural and resource-limited settings, we developed two risk models: office-based and laboratory-based, whereby the former did not include laboratory variables.

For data harmonization and analysis, we used R statistical software. When collaborators shared data in Stata or SPSS files, we first used the originating software to open the data sets and double-check labels and variable names. Notably, because we pooled cohort data from studies conducted in Spanish- and Portuguese-speaking countries, some variable names and labels were at risk of being corrupted because of the different characters in these languages.

Ad hoc R scripts were developed for data cleaning, management and analysis. For example, the R script for data cleaning would read the original data sets, check for inconsistencies and harmonize variable names; the R script for data management would read all extracted data sets after cleaning and merge them into one big data set; and the R scripts for the analyses were specific for each paper. Each process was led and verified by a member of the team. We did not build software or programs to automate any of the processes described here. It is important to carefully plan all cleaning and data management procedures to minimize the chance of inadvertently introducing bias or distorting the original distributions of data. It is also important to discuss with data providers, collaborators and stakeholders how to avoid these potential issues.

Participant confidentiality

The CC-LAC pooled only deidentified data. Data provided by the investigators did not include personal identifiers such as names, identification numbers, addresses or phone numbers. Data were stored on a secure server at Imperial College London, United Kingdom (where the full-time investigator was based), meaning that only the full-time CC-LAC investigator had access to the data and used personal log in credentials to access it within the university network. We allowed each cohort investigator team to determine whether they needed or wanted to sign a formal data-sharing agreement.

OUTCOMES

Scientific articles with country-specific guidance

To date, the CC-LAC has authored three academic papers. The first was a cohort profile that summarizes the history of the CC-LAC and provides a description of the cohorts included in it (18). In the second, the CC-LAC described the association between cardiometabolic risk factors and cardiovascular

outcomes, comparing the magnitude of such associations with those from global cohorts and quantifying the number of cardiovascular events attributable to the risk factors in Latin America and the Caribbean (35). In the third, the Consortium developed the first cardiovascular risk score for Latin America and the Caribbean, and the risk score was calibrated for 31 countries across the area (36). The published articles include instruments for decision-makers and clinical practitioners: the second paper provides country-specific infographics that describe the associations of risk factors with cardiovascular outcomes and the third paper provides country-specific cardiovascular risk charts.

CONSIDERATIONS FOR ESTABLISHING A DATA POOLING PROJECT

Researchers should consider several points when planning and operating a data pooling project. We offer specific recommendations based on our experience with the CC-LAC.

Resources

Although we argue that data pooling projects are less expensive than collecting new data, pooling projects have costs, and these should be considered at the project planning stage. For a project the size of CC-LAC (with >30 separate cohort-country units for pooling, a total sample size of >174 000 participants and 25 key variables that were standardized), we employed one full-time researcher for 3 years to deal with all day-to-day data-finding and data-handling activities. If the use of advanced statistical models is planned, then a computer with a large amount of memory is required as well as adequate data storage capacity, including the facility to back up data, depending on the amount and type of data to be pooled. If the analyses involve computationally intensive procedures, consider accessing high-performance cloud computing (such as Amazon Web Services or Google Cloud) as a cost-effective alternative to local computing. Additionally, there should be resources for publication costs and for online and, if possible, in-person meetings with the key investigators (e.g. the steering committee). Finally, the location of the computer hub should be strategic as well.

Consider establishing operations, in terms of the country and specific location, where access to power and internet connectivity are secure and stable. If the organization plans to host meetings or visiting researchers, consider a location with easy and cost-effective access. Particularly for the Caribbean, seek a location with a low risk of natural disasters or similar events that may disrupt the work. Although it is important to adhere to national data protection rules, maximizing the use of virtual operations can offer important security against the uncertainties and disruptions of natural disasters. If the plan is to grow operations by collecting more data or new types of data (e.g. images) or by expanding into other health fields, carefully and purposely plan and budget for the resources that may be needed immediately and into the future. As always, sustainability should be a key consideration. For long-term endeavors, also plan how to monitor data and mitigate time bias (i.e. when data become too outdated to be representative of the current epidemiological landscape); additionally, consider specific analytical strategies, such as multilevel regressions incorporating a random intercept by study, year or decade, or sensitivity analyses whereby the robustness of the findings is verified by year or decade.

Long-term maintenance and updating

The costs of long-term maintenance for a pooled data resource are relatively low compared with the initial start-up costs. Nonetheless, costs continue, and longer-term goals should be incorporated into start-up planning, with approvals secured from data providers for these future uses of their data. If the endeavor is operating within an academic environment, employing interns or students on a doctoral program might be a cost-effective way to maintain the data. The two primary areas of work related to longer-term maintenance are likely to be finding, preparing and pooling new eligible studies, and statistical analyses linked to subsequent publications.

Data protection governance

For CC-LAC we request only deidentified data resources, and the host institutions did not require the Consortium to sign data-sharing agreements. Data protection legislation across the Caribbean that closely follows the European Union's General Data Protection Regulation is now common, and it is more likely that data-sharing agreements will be needed prior to collaborating institutions releasing their data. Agreements will almost certainly be required if special categories of personal data are included, such as genetic information, biometric data, and geolocation and health data, as well as personal data revealing racial and ethnic origin, political opinions, religious or ideological convictions, or trade union membership. In some cases, data-sharing agreements will be made between investigators, and in other cases these agreements will be enacted at the institutional level. For the latter case, it is important to ensure that the institution hosting the data pooling has the capacity to handle this process, which may involve legal support, contract management and, in some cases, expertise in intellectual property. Regardless of any data agreement, the ownership of data in any data pooling endeavor remains with those who originally produced the data. In other words, because someone is sharing data with the Consortium, for example, it does not make the Consortium the owner of that data. It is important to be respectful of data ownership and always use the pooled data only within the limits of any agreement. It is important to communicate regularly with data providers to build and maintain trust.

Thus, the data pooling team must understand and meet the data protection requirements for each contributing country. Ensuring compliance with multiple data protection laws and securing the pooled data resources against data breaches are crucial and require careful technical management. In the case of CC-LAC, before starting the data collation process, we enacted data security protocols as advised by a university data specialist. Key among these protocols was controlling access to the pooled data resource. Similarly, it is essential to consider and implement physical and electronic security measures to protect sensitive data, as needed. While the specific measures may vary depending on the data, physical measures regularly include restricting access to the physical area where data are stored; electronic measures should include encryption of data, both in transit and at rest.

Ethical considerations

Data pooling endeavors are generally classed as secondary analyses of data, but in the case of health information they are still associated with data collected from individuals. We strongly recommend the development of a comprehensive protocol and manual of procedures, similar to any epidemiological research, and that potential research protocols are submitted to the local ethics committee for formal review. The pooled data will be deidentified rather than formally anonymized, and because re-identification is theoretically possible, international ethical standards must be met: for example, the data must be used only for the purposes agreed to by those who provided the data; there must never be an attempt to identify individuals; and information must always be reported in an aggregated form. Always acknowledge the original participants in any manuscripts developed from the data; this can be done in a brief paragraph in publications. The CC-LAC protocols were approved by the institutional review board at *Universidad Peruana Cayetano Heredia* (Lima, Peru) and Imperial College London (London, United Kingdom).

Data quality for pooling

Ensuring data quality encompasses a set of requirements that the project ought to thoroughly check, such as the completeness, consistency, uniqueness and accuracy of the data. To assess data quality, we checked for duplicate observations; identified outliers or unusual values, alone and in combination (e.g. a combination of weight and height that would lead to an unlikely body mass index); explored inconsistencies between variables (e.g. females with prostate cancer); and summarized the number and proportion of missing values for each variable.

Data accuracy for pooling

Data accuracy (i.e. the closeness of the data to real-world observations) is one aspect of data quality that is difficult to assess. Domain knowledge is critical, for instance to explore whether the weight or height distributions by age and sex are broadly as expected. We explored accuracy by creating visualizations of the distributions of clinical measurements. Expectations depend on local knowledge, perhaps drawing on information from relevant nationally representative data. For basic health measures, for example, cross-sectional health surveys can be used as objective measurements to compare with expectations of the distribution. As always, if questions arise, these should be discussed with those who provided the data.

Data bias and representativeness

It is important to explore the sampling frame and sampling process for each included study to understand the purported and actual representativeness of each sample. Comparing the key demographics sampled, such as age and sex, with, for example, census data can facilitate this exploration.

While some eligible studies may be nationally representative, others will reflect a geographical subregion or population subgroup, whose demographic profile may not mimic that of a national population. This is not a reason to exclude a study, but it is important that these differences are made transparent and that sampling details are recorded as part of the pooling study's metadata. Thus, despite a pooled data resource having a large sample size, its representativeness may be limited. Using CC-LAC as an example, from its >30 pooled cohorts, only 1 was from the English-speaking Caribbean, potentially limiting its usefulness for this subregion. This does not invalidate the work by the CC-LAC, but careful interpretation and extrapolation of findings are essential.

Skill requirements

Effective data pooling requires two key skill sets: data handling and statistical analysis, and domain-specific knowledge.

A basic requirement is for staff with training in data handling and statistical analysis, including the practical skill of developing reproducible statistical algorithms using a computing language, ideally from an open-source environment, such as R or Python. Data quality checks can be time-consuming. In the case of CC-LAC, enacting a comprehensive set of data quality routines took between 1 and 2 full days for each cohort, with extra time needed to resolve queries with data providers.

Domain-specific knowledge allows the data pooling team to better understand the expectations for individual variables, and this context can improve the data preparation process. Most data pooling endeavors will involve both junior and more experienced researchers, and because data preparation will involve an iterative series of questions, it is important to foster a sense of open dialogue among all members of the data pooling team.

Data pooling requires patience and respect for the workloads and competing demands of the many investigators providing data. The data pooling project will be contacting multiple colleagues, asking for data, engaging with them to resolve queries and develop confidence in the data set, and coauthoring papers. Problem-solving and progress will sometimes take perseverance.

Conclusions

Data pooling projects using health data can answer high-impact questions with increased statistical power to inform decisions in public health and clinical medicine. They generally require fewer resources than collecting new data and help to increase the longer-term impact of projects collecting primary data. Data pooling projects encourage data efficiency by reusing already available data, and they allow researchers to develop skills in areas that are highly valuable, such as data science and big data. Like any other type of health research, data pooling projects must follow systematic protocols and be transparent about their methods; they must also meet the requirements of data governance legislation in each country from which data are contributed. Each data pooling effort should aim to fill critical gaps in science, leveraging the collective expertise of those sharing and pooling data.

Authors' contributions. RMC-L and IRH conceived the idea for manuscript, wrote the first draft and edited the final submitted version. RMC-L and IRH produced the supplementary materials. RMC-L produced the figures with support from IRH. Both authors reviewed and approved the final version of the manuscript.

Conflicts of interest. None declared.

Funding. No specific funding was received for this manuscript. The CC-LAC is funded by the Wellcome Trust (grant no. 214185/Z/18/Z).

Disclaimer. Authors hold sole responsibility for the views expressed in the manuscript, which may not necessarily reflect the opinion or policy of the *Revista Panamericana de Salud Pública/Pan American Journal of Public Health* or the Pan American Health Organization.

REFERENCES

1. Escobedo J, Buitrón LV, Velasco MF, Ramírez JC, Hernández R, Macchia A, et al. High prevalence of diabetes and impaired fasting glucose in urban Latin America: the CARMELA Study. Diabet Med. 2009;26(9):864-71.
2. Hernández-Hernández R, Silva H, Velasco M, Pellegrini F, Macchia A, Escobedo J, et al. Hypertension in seven Latin American cities: the Cardiovascular Risk Factor Multiple Evaluation in Latin America (CARMELA) study. J Hypertens. 2010;28(1):24-34.
3. Schargrodsky H, Hernández-Hernández R, Champagne BM, Silva H, Vinueza R, Silva Ayçaguer LC, et al. CARMELA: assessment of cardiovascular risk in seven Latin American cities. Am J Med. 2008;121(1):58-65.
4. Silva H, Hernandez-Hernandez R, Vinueza R, Velasco M, Boissonnet CP, Escobedo J, et al. Cardiovascular risk awareness, treatment, and control in urban Latin America. Am J Ther. 2010;17(2):159-66.
5. Touboul PJ, Vicaut E, Labreuche J, Acevedo M, Torres V, Ramirez-Martinez J, et al. Common carotid artery intima–media thickness: the Cardiovascular Risk Factor Multiple Evaluation in Latin America (CARMELA) study results. Cerebrovasc Dis. 2011;31(1):43-50.
6. Vinueza R, Boissonnet CP, Acevedo M, Uriza F, Benitez FJ, Silva H, et al. Dyslipidemia in seven Latin American cities: CARMELA study. Prev Med. 2010;50(3):106-11.
7. Rubinstein AL, Irazola VE, Poggio R, Bazzano L, Calandrelli M, Lanas Zanetti FT, et al. Detection and follow-up of cardiovascular disease and risk factors in the Southern Cone of Latin America: the CESCAS I study. BMJ Open. 2011;1(1):e000126.
8. Miranda JJ, Herrera VM, Chirinos JA, Gómez LF, Perel P, Pichardo R, et al. Major cardiovascular risk factors in Latin America: a comparison with the United States. The Latin American Consortium of Studies in Obesity (LASO). PLoS One. 2013;8(1):e54056.
9. Bautista LE, Casas JP, Herrera VM, Miranda JJ, Perel P, Pichardo R, et al. The Latin American Consortium of Studies in Obesity (LASO). Obes Rev. 2009;10(3):364-70.
10. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. Lancet. 2021;398:957-80.
11. NCD Risk Factor Collaboration (NCD-RisC). Repositioning of the global epicentre of non-optimal cholesterol. Nature. 2020;582:73-7.
12. NCD Risk Factor Collaboration (NCD-RisC). Rising rural body-mass index is the main driver of the global obesity epidemic in adults. Nature. 2019;569:260-4.
13. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128.9 million children, adolescents, and adults. Lancet. 2017;390:2627-42.
14. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19.1 million participants. Lancet. 2017;389:37-55.
15. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. Lancet. 2016;387:1513-30.
16. NCD Risk Factor Collaboration (NCD-RisC)—Americas Working Group. Trends in cardiometabolic risk factors in the Americas between 1980 and 2014: a pooled analysis of population-based surveys. Lancet Glob Health. 2020;8(1):e123-e33.
17. GBD 2019 Diabetes in the Americas Collaborators. Burden of diabetes and hyperglycaemia in adults in the Americas, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Diabetes Endocrinol. 2022;10(9):655-67.
18. Cohorts Consortium of Latin America and the Caribbean (CC-LAC), Carrillo-Larco RM, di Cesare M, Hambleton IR, Hennis A, Irazola V, et al. Cohort profile: the Cohorts Consortium of Latin America and the Caribbean (CC-LAC). Int J Epidemiol. 2020;49(5):1437-g.
19. Woodward M, Barzi F, Martiniuk A, Fang X, Gu DF, Imai Y, et al. Cohort profile: the Asia Pacific Cohort Studies Collaboration. Int J Epidemiol. 2006;35(6):1412-6.
20. Danesh J, Erqou S, Walker M, Thompson SG, Tipping R, Ford C, et al. The Emerging Risk Factors Collaboration: analysis of individual data on lipid, inflammatory and other markers in over 1.1 million participants in 104 prospective studies of cardiovascular diseases. Eur J Epidemiol. 2007;22(12):839-69.
21. Lewington S, Clarke R, Qizilbash N, Peto R, Collins R. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. Lancet. 2002;360:1903-13.
22. Singh GM, Danaei G, Farzadfar F, Stevens GA, Woodward M, Wormser D, et al. The age-specific quantitative effects of metabolic risk factors on cardiovascular diseases and diabetes: a pooled analysis. PLoS One. 2013;8(7):e65174.
23. Berrington de Gonzalez A, Hartge P, Cerhan JR, Flint AJ, Hannan L, MacInnis RJ, et al. Body-mass index and mortality among 1.46 million white adults. N Engl J Med. 2010;363(23):2211-9.
24. Magnussen C, Ojeda FM, Leong DP, Alegre-Diaz J, Amouyel P, Aviles-Santa L, et al. Global effect of modifiable risk factors on cardiovascular disease and mortality. N Engl J Med. 2023;389(14):1273-85.
25. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020;396:1204-22.
26. GBD 2019 Risk Factors Collaborators. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020;396:1223-49.
27. Miranda JJ, Barrientos-Gutiérrez T, Corvalan C, Hyder AA, Lazo-Porras M, Oni T, et al. Understanding the rise of cardiometabolic diseases in low- and middle-income countries. Nat Med. 2019;25(11):1667-79.
28. Atun R, de Andrade LOM, Almeida G, Cotlear D, Dmytraczenko T, Frenz P, et al. Health-system reform and universal health coverage in Latin America. Lancet. 2015;385:1230-47.
29. Cotlear D, Gómez-Dantés O, Knaul F, Atun R, Barreto ICHC, Cetrángolo O, et al. Overcoming social segregation in health care in Latin America. Lancet. 2015;385:1248-59.
30. de Andrade LOM, Filho AP, Solar O, Rígoli F, de Salazar LM, Serrate PC-F, et al. Social determinants of health, universal health coverage, and sustainable development: case studies from Latin American countries. Lancet. 2015;385:1343-51.
31. Muñoz OM, Rodríguez NI, Ruiz Á, Rondón M. Validación de los modelos de predicción de Framingham y PROCAM como estimadores del riesgo cardiovascular en una población colombiana [Validation of PROCAM and Framingham prediction models as estimators of cardiovascular risk in a Colombian population]. Rev Colomb Cardiolog. 2014;21(4):202-12.
32. Lajous M, Ortiz-Panozo E, Monge A, Santoyo-Vistrain R, García-Anaya A, Yunes-Díaz E, et al. Cohort profile: the Mexican Teachers' Cohort (MTC). Int J Epidemiol. 2017;46(2):e10.
33. Ueda P, Woodward M, Lu Y, Hajifathalian K, Al-Wotayan R, Aguilar-Salinas CA, et al. Laboratory-based and office-based risk scores and charts to predict 10-year risk of cardiovascular disease in 182 countries: a pooled analysis of prospective cohorts and health surveys. Lancet Diabetes Endocrinol. 2017;5(3):196-213.
34. Hajifathalian K, Ueda P, Lu Y, Woodward M, Ahmadvand A, Aguilar-Salinas CA, et al. A novel risk score to predict cardiovascular disease

risk in national populations (Globorisk): a pooled analysis of prospective cohorts and health examination surveys. Lancet Diabetes Endocrinol. 2015;3(5):339-55.
35. Cohorts Consortium of Latin America and the Caribbean (CC-LAC). Impact of common cardio-metabolic risk factors on fatal and non-fatal cardiovascular disease in Latin America and the Caribbean: an individual-level pooled analysis of 31 cohort studies. Lancet Reg Health Am. 2021;4:100068.
36. Cohorts Consortium of Latin America and the Caribbean (CC-LAC). Derivation, internal validation, and recalibration of a cardiovascular risk score for Latin America and the Caribbean (Globorisk-LAC): a pooled analysis of cohort studies. Lancet Reg Health Am. 2022;9:100258.

Manuscript submitted 18 February 2024. Revised version accepted for publication on 22 April 2024.

Datos para análisis poblacionales de la salud: el Consorcio de Cohortes de América Latina y el Caribe

RESUMEN

Objetivo. Se describe el funcionamiento cotidiano del Consorcio de Cohortes de América Latina y el Caribe, al tiempo que se brinda información detallada sobre los recursos necesarios y se proponen sugerencias para los investigadores del Caribe con el fin de que esta guía pueda ser utilizada para la puesta en marcha de un proyecto de agrupamiento de datos.

Metodología. El Consorcio de Cohortes de América Latina y el Caribe comenzó mediante la creación de un comité directivo, es decir, un equipo de expertos regionales que orientaron la puesta en marcha y el funcionamiento del proyecto. El Consorcio invita a afiliarse a investigadores que aceptan compartir datos a nivel individual sobre cuestiones de interés y que, a partir de entonces, pueden contribuir a la consecución de los objetivos y el funcionamiento del proyecto; también se les ofrece la posibilidad de ser coautores de artículos. Se utilizó una metodología de revisión sistemática para encontrar investigadores que dispusieran de datos de interés para el proyecto y se elaboró un protocolo, en forma de un manual de procedimientos, a fin de documentar todos los aspectos del funcionamiento del proyecto.

Resultados. Si en un estudio se reclutaba a personas de varios países, la muestra de cada país se contabilizaba como una cohorte diferente. En consecuencia, en el 2024 nuestras fuentes de datos combinadas incluyen más de 30 unidades diferentes de 13 países, con un tamaño muestral combinado de más de 174 000 participantes. A partir de este recurso que tiene unas características únicas, se han elaborado estimaciones del riesgo específicas de cada región para los factores de riesgo cardiometabólico (por ejemplo, antropometría) y las enfermedades cardiovasculares, y se ha creado una puntuación de riesgo cardiovascular específica de cada región para su utilización en el ámbito clínico.

Conclusiones. Los proyectos de agrupamiento de datos son menos costosos que la recopilación de datos nuevos, y aumentan el valor a más largo plazo y el impacto de los datos aportados. Requieren una metodología sistemática y transparente, así como conocimientos especializados en el manejo y análisis de datos. Los investigadores que participan en un proyecto de agrupamiento de datos deben conocer y ser capaces de cumplir las diversas normas de protección de datos previstas en la legislación de los distintos países, ya que es probable que difieran en las distintas jurisdicciones.

Palabras clave Conjunto de datos; metaanálisis como asunto; macrodatos; ciencia de los datos; epidemiología.

Dados para análises populacionais em saúde: o Consórcio de Coortes da América Latina e do Caribe

RESUMO

Objetivo. Descrevemos as operações diárias do Consórcio de Coortes da América Latina e do Caribe (CC-LAC), detalhando os recursos necessários e oferecendo dicas para os pesquisadores caribenhos a fim de que este guia possa ser usado para iniciar um projeto de dados colaborativos.

Métodos. O CC-LAC iniciou-se com a criação de um comitê diretor, ou seja, uma equipe de especialistas regionais que orientaram a configuração e as operações do projeto. O Consórcio convida pesquisadores que concordem em compartilhar dados de nível individual sobre temas de interesse a se tornarem membros e contribuir com os objetivos e as operações do projeto; além disso, esses pesquisadores são convidados a serem coautores de artigos. Usamos uma metodologia de revisão sistemática para identificar pesquisadores que tivessem recursos de dados compatíveis com o projeto e elaboramos um protocolo (ou seja, um manual de procedimentos) para documentar todos os aspectos das operações do projeto.

Resultados. Caso um estudo tivesse recrutado pessoas de mais de um país, a amostra de cada país foi considerada uma coorte distinta. Assim, em 2024, nossos recursos de dados combinados continham mais de 30 unidades diferentes de 13 países, com um tamanho amostral total de mais de 174 mil participantes. Graças a esse recurso exclusivo, produzimos estimativas de risco específicas para a região de fatores de risco cardiometabólicos (p. ex., medidas antropométricas) e doença cardiovascular, além de criarmos um escore de risco cardiovascular para uso clínico específico para a região.

Conclusões. Projetos de dados colaborativos são menos dispendiosos que a coleta de novos dados e aumentam o valor e o impacto dos dados compartilhados em longo prazo. Isso exige uma metodologia sistemática e transparente; conhecimento especializado em tratamento e análise de dados é um pré-requisito. Pesquisadores que decidam participar de um projeto de dados colaborativos devem conhecer e ser capazes de respeitar as diferentes normas de proteção de dados prescritas nas legislações nacionais sobre dados, uma vez que essas normas tendem a variar de um lugar para outro.

Palavras-chave Conjunto de dados; metanálise como assunto; big data; ciência de dados; epidemiologia.