Practice of Epidemiology

Toward Rigorous Data Harmonization in Cancer Epidemiology Research: One Approach

Betsy Rolland*, Suzanna Reid, Deanna Stelling, Greg Warnick, Mark Thornquist, Ziding Feng, and John D. Potter

* Correspondence to Dr. Betsy Rolland, Cancer Prevention Fellowship Program, National Cancer Institute, 9609 Medical Center Drive, Room 3E460, Rockville, MD 20892 (e-mail: brolland@uw.edu).

Initially submitted October 17, 2014; accepted for publication May 12, 2015.

Cancer epidemiologists have a long history of combining data sets in pooled analyses, often harmonizing heterogeneous data from multiple studies into 1 large data set. Although there are useful websites on data harmonization with recommendations and support, there is little research on best practices in data harmonization; each project conducts harmonization according to its own internal standards. The field would be greatly served by charting the process of data harmonization to enhance the quality of the harmonized data. Here, we describe the data harmonization process utilized at the Fred Hutchinson Cancer Research Center (Seattle, Washington) by the coordinating centers of several research projects. We describe a 6-step harmonization process, including: 1) identification of questions the harmonized data set is required to answer; 2) identification of high-level data concepts to answer those questions; 3) assessment of data availability for data concepts; 4) development of common data elements for each data concept; 5) mapping and transformation of individual data points to common data elements; and 6) quality-control procedures. Our aim here is not to claim a “correct” way of doing data harmonization but to encourage others to describe their processes in order that we can begin to create rigorous approaches. We also propose a research agenda around this issue.

cancer epidemiology; data harmonization; data pooling

WHAT IS DATA HARMONIZATION AND WHY IS IT IMPORTANT?

Retrospective data harmonization involves pooling heterogeneous data from disparate data sets and transforming them into 1 common data set to use for new analyses. This can be done either by using variables that are common (e.g., age or sex) or similar (e.g., smoking status) across data sets or by deriving new variables from disparate questions with a common theme (e.g., total alcohol intake). We call these harmonized variables common data elements (CDEs). CDEs represent the ideal version of the variable, one that will allow analyses to be conducted across the combined data set.

Pooling of existing data sets has a number of benefits, including 1) the ability to ask questions for which individual studies are underpowered and 2) the related ability to focus on rare outcomes (1). However, data harmonization is a complex operation that is expensive and time-consuming. The greatest challenge with harmonizing epidemiologic data is ensuring that the variables used are thoroughly understood before combining them. Two key questions to address are 1) the heterogeneity of the questions asked during data collection and 2) the meaning of specific variables. Data collection instruments can ask questions in similar yet subtly different ways, ensuring that careless harmonization leads to inappropriate conclusions. For example, one study may ask the...
question: How many hours a week do you participate in vigorous sport activities? Another study may ask: How many days per week do you participate in moderate or vigorous physical activity? Whether these 2 variables can be harmonized depends strongly on the scientific questions being asked.

Variable names and descriptions can mean different things to different people. The description “months of medication use” seems straightforward, but the possible values may not be as expected. In research on how postdoctoral researchers in cancer epidemiology understand data sets, Rolland and Lee (2) found that how the data were actually coded to produce this variable differed from what was included in the data dictionary and that this was not uncommon.

Science is composed of a myriad of actions, assumptions, and decisions, both explicit and tacit, making thorough documentation of a research project difficult. These actions, assumptions, and decisions are reflected not only in the questions asked but also in what probes are used, how ambiguous answers are coded, how unknown responses are treated, and how variables are constructed or derived after data have been collected. Even with substantial documentation, it can be challenging to understand data that someone else has collected, without engaging in time-consuming conversations with the original data collectors (2). Trying to harmonize multiple data sets magnifies this challenge. (See Web Appendix 1, available at http://aje.oxfordjournals.org/, for some pitfalls of harmonization.)

Few investigators write extensively about their data-harmonization procedures, despite the widespread use of harmonized data. Even papers that reference the methodological issues of data pooling tend to gloss over the actual process of data harmonization itself (3–5). In the paper by Fortier et al., for example, there are details in the Methods section on how the data are selected; then the harmonization process itself is summed up thus:

In order to classify the assessment items and to ensure the validity and reproducibility of the pairing results, sets of comprehensive ‘pairing rules’ specific to each variable are defined. Development of pairing rules is context specific and involves a systematic process of iteration between scientific experts and trained research officers. Using these pairing rules, trained research officers determine whether or not a variable can be recreated using the assessment items collected by each participating study (4, p. 1317).

There are no details on how the scientific experts and trained research officers derived and refined their pairing rules or how long it took. Though such an explanation might be complex, without it, researchers are unable to apply the methods themselves or even fully evaluate the methods proposed.

This lack of discussion on how to pool data for a new analysis means that investigators in each study craft their own methods of harmonization, with little empirical evidence to support any one method. Pooled studies have markedly increased power because of their larger sample sizes; it behooves us to be sure that the conclusions being drawn are as accurate as possible. To that end, we describe here the harmonization processes needed for 4 different studies, the analyses of which were directed by 2 of us (M.T., Z.F.), both senior biostatisticians.

THE DATA HARMONIZATION PROCESS

In this section, we describe how data harmonization was conducted at the Fred Hutchinson Cancer Research Center (Seattle, Washington) on data sets from several projects, including the Asia Cohort Consortium (ACC) (6–9), the Early Detection Research Network (EDRN) (http://edrn.nci.nih.gov/), the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) (http://www.fredhutch.org/en/labs/phs/projects/cancer-prevention/projects/gecco.html), and Population-based Research Optimizing Screening through Personalized Regimens (PROSPR) (http://appli research.cancer.gov/prospr/). These data-harmonization efforts took place under the aegis of the Comprehensive Center for the Advancement of Scientific Strategies (COMPASS), a group in the Public Health Sciences Division of the Fred Hutchinson Cancer Research Center. COMPASS has provided expertise in study and research consortia coordination since the 1980s, serving as the data management and coordination center for the Carotene and Retinol Efficiency Trial (CARET), as well as for the operational components of the EDRN, the Transdisciplinary Research on Energetics and Cancer (TREC) initiative, and PROSPR. COMPASS also provides data harmonization support for GECCO and ACC analyses. These opportunities have allowed COMPASS to accumulate experience from multiple data-harmonization projects across widely different types of data, collaborators, and scientific questions.

The harmonization process has varied for each project, yet commonalities exist. Here we present an idealized overview of this process, including examples of specific harmonization efforts. Not every project followed these steps exactly, and some iterated through them in different orders, but most engaged in these activities at some point during harmonization.

In general, harmonization follows these steps:

1. Identification of the questions the harmonized data set is required to answer
2. Identification of the high-level data concepts required to answer those questions
3. Assessment of data availability for data concepts
4. Development of CDEs for each data concept
5. Mapping and transformation of individual data points to CDEs
6. Quality-control procedures

Step 1: Identification of the questions the harmonized data set is required to answer

A data set is harmonized for a specific project, to answer specific questions. It is obvious, then, that everyone involved must be in agreement on the aims of that project and the questions, since these drive the steps that follow. The primary research questions drive the statistical analysis plan by identifying the dependent and key independent variables and their required properties (e.g., timing of collection). Nonetheless, it almost always makes sense to think about other, related questions that could also be answered with such a harmonized data set. The ACC took this approach in its first harmonization project, which focused on body mass index (weight (kg)/height (m)²)
in Asian populations. Because this was the first collaborative project of the ACC, project leaders took a less-is-more approach and focused on data points that were critical to answering the question of association between mortality and body mass index in Asian populations (6). Early discussions also explored what questions might be built on this foundation, such as additional projects on diabetes and rare cancers, which were pursued subsequently (7, 10, 11).

**Step 2: Identification of the high-level data concepts required to answer those questions**

Once all participants have agreed on the overarching goals of the project with the harmonized data set, conversations turn to what kinds of data must be harmonized in order to answer the questions of interest. In GECCO, investigators needed a wide range of epidemiologic data to use with genome-wide association study data in order to investigate interactions. These included data on use of nonsteroidal antiinflammatory drugs, hormone use, dietary exposures, and tobacco and alcohol use—categories of data that were already established risk factors for colorectal cancer. Likewise, PROSPR started with conceptual models of the screening process for breast, cervical, and colon cancer, identifying elements of the process that were of interest in future analyses, such as time since last screening, patient demographic characteristics, and disease risk factors.

Identification of these high-level data concepts is not an easy process, as it requires researchers to take a step back from the detailed data, think conceptually about their research questions, and then negotiate around those concepts with their colleagues until they reach agreement on what is important. For investigators accustomed to moving directly to individual data points, it may feel like a waste to start at such a high level, but it has been our experience that time spent on this step makes the work that follows substantially smoother and quicker. Such conversations need to involve a variety of people, including the Principal Investigators (PIs) of the original studies and their data managers, who collectively have an understanding of the subtle nuances within the data they collect and manage, including questions of data reliability and availability. Answers to these questions can influence the final form of the project’s scientific questions.

**Step 3: Assessment of data availability for data concepts**

Of course, what does not exist cannot be harmonized; so, before moving forward, data concepts must be compared with the data available from the original studies. One way of doing this is to send a list of data concepts to study PIs and ask them to return a list of available data points that may be useful in exploring each concept. This has the advantage that the people who know the data best review their data and choose those that best match. However, it does require that study PIs devote sufficient time to the task and that data concepts be as precise and bounded as possible. A second way of investigating data availability is to have study PIs send a list of all available data, after which the harmonization team categorizes the data and matches them to data concepts. Although, on the surface, the latter appears to be easier on the study PIs, in reality the questions generated by such an approach are likely to require as much work to answer as the former. A third approach is to collect the original study instruments, such as questionnaires or interview protocols, as well as data dictionaries. This has the advantage of laying out clearly what questions were asked of participants. However, participant answers may be transformed between the data collection and the existing database in ways that may not be well documented, which can be confusing for the harmonization team. In reality, a harmonization project may draw upon all 3 of these approaches.

We have found that this process of assessing data availability is iterative and can take several rounds of discussion before a final determination can be made as to whether data are sufficiently available across participating studies. If not all studies are able to provide data for a given data concept, there are several potential paths forward. First, the group can decide to collect data for a particular concept from a subset of studies. This diminishes the power of the analyses, but it may be preferable to abandoning a data concept entirely. Second, the data concept can be adjusted to a higher level of abstraction, if data on that level are, indeed, available in all studies. At this step, it may be determined that some data concepts do not have sufficient comparable representation in the studies being harmonized for that concept to be a good candidate for harmonization. That, in turn, may make some already agreed-upon scientific questions effectively unanswerable with the data available, returning the data harmonization process to step 1.

**Step 4: Development of CDEs for each data concept**

This is the step that requires the deepest knowledge of both the existing data sets to be harmonized and the scientific purpose of the harmonized data. Like the data-availability analysis, the development of CDEs is iterative and can take a substantial investment of time. The object is to produce individual data points (with data type and possible values) that represent the data concepts covered in step 3. The ACC investigators were interested in participants’ history of disease, so they developed CDEs that included the following, all of which had possible values of yes, no, and unknown/refused:

- Baseline diagnosis of stroke
- Baseline diagnosis of diabetes
- Baseline diagnosis of cancer, not including basal/squamous cell skin cancer

These CDEs were relatively straightforward for participating cohort studies to produce. A more challenging area of harmonization can be seen in GECCO’s work on CDEs on red meat consumption. The goal was to harmonize data on red meat intake over 13 studies, all of which had collected the data differently. To achieve this goal, COMPASS staff examined each data-collection instrument and data repository to determine comparability of the desired data elements. They found that frequency, duration, and quantity were asked about in a variety of ways, some as categorical variables with differing cutpoints, others as continuous variables. Consider harmonizing the following questions:

**Am J Epidemiol. 2015;182(12):1033–1038**
• About 2 years ago, on average, how often did you eat a serving of red meat (not chicken or fish)? Please specify the number of servings per day, per week, or per month. A serving of red meat is 2–3 ounces [oz] or a piece of meat about the size of a deck of cards.
  - __number of servings
  - Interval of frequency for meat consumption
    - 1) servings per day
    - 2) servings per week
    - 3) servings per month
• Average use, during the past year: beef, pork, or lamb as a main dish, e.g., steak, roast, ham, etc. (4–6 oz):
  - Never or less than once per month
  - 1–3 per month
  - 1 per week
  - 2–4 per week
  - 5–6 per week
  - 1 per day
  - 4–6 per day
  - 6+ per day
• How often do you eat red meat (beef, lamb . . .)?
  - Once/week or less
  - 2–4 times/week
  - More than 5 times/week

There are at least 2 ways to approach this harmonization, both of which require specific assumptions and decisions to be made by the harmonization team. First, a conversion must be made between “times per day” and “times per week” or “per month,” which is relatively straightforward. One path is to convert categorical variables to continuous variables by mapping each category to a value in that category’s range (such as the minimum, maximum, or midpoint). A second path is to convert the continuous variables to categorical variables by, for example, placing the answer “5 servings” in the “4–7 times/week” category. The decision on how to proceed must be based on the scientific questions being asked; for example, if we are exploring how low quantities of red meat consumption affect a participant’s risk of colorectal cancer, our harmonization may lean toward maintaining specificity and create a continuous CDE variable. On the other hand, if categories of consumption are sufficient for addressing our question, then harmonizing to a categorical CDE variable may be simpler, although harmonizing a categorical variable to a categorical CDE may be challenging if a category in the variable overlaps with multiple categories in the CDE. The bottom line is that there are multiple acceptable answers and ways to produce a CDE that can answer the project’s question, but all involve assumptions and decisions that must be thoroughly discussed to understand the ramifications for the analyses. Sometimes, it is possible to take more than 1 approach to a CDE and to explore, in sensitivity analyses, the impact of choosing one version of a CDE over another.

Once CDEs have been established, it is critical to the success of the harmonization project that they be documented as explicitly and meticulously as possible. Data type, possible values, and a definition of what the variable represents must be documented, followed by discussion with the data suppliers to ensure that everyone agrees on what the resulting CDE means.

**Step 5: Mapping and transformation of individual data points to CDEs**

Once a list of CDEs has been developed, original study data must be mapped to produce the required CDEs. There are 2 main routes to mapping: 1) the original study data can be sent to the harmonization team, then mapped by them to the CDEs; or 2) the original study team can map the data to the CDEs, then send the harmonized data to the harmonization team. What is important is that the process takes advantage of each original study team’s deep understanding of its own data, its nuances and quirks, while also leveraging the harmonization team’s expertise on the CDEs. Balancing these two ideas can be tricky, since it requires detailed conversations about individual data points, and budgets for these activities are usually limited.

Mapping difficulty can range from the simple, such as directly importing the data delivered by the data provider or making minor adjustments to the content (e.g., changing 2 = no to 0 = no), to more complex transformations. One example of a complex transformation is the conversion of data on supplemental calcium intake collected by Newcomb et al. (12). The study questionnaire asked several questions about calcium supplementation, including questions on frequency and duration. The CDE was “calcium from supplements (mg/day),” so mapping needed to account for the quantity of calcium currently being consumed by each participant via calcium supplements, multivitamins, or both. Ideally, this would mean assigning a standard calcium concentration, in mg/tablet, to calcium supplements and to multivitamins, calculating the intake in mg/day, and adding the numbers together as indicated. In reality, data often conflicted, with some participants answering “no” to the first question about use but actually providing values for the amount used and frequency (12). Decisions and rules that addressed inconsistencies in the data needed to be included with rules about transforming data to mg/day.

In the GECCO project, the COMPASS staff took the first mapping route, requesting data from the original studies and then mapping and transforming them as needed at the Fred Hutchinson Cancer Research Center. In cases where 1 or more variables were to be mapped at the GECCO Coordinating Center, data-collection instruments and data repositories were examined and GECCO staff and the staffs and PIs of participating studies were consulted to develop an understanding of the data sufficient to allow documentation of the mapping. An internal tracking database was created to manage mapping documentation; this described how contributed data from each study should be transformed and harmonized to the GECCO CDEs. This information was then used by COMPASS programmers to create scripts to transform study data upon receipt at the GECCO Coordinating Center. Note that this activity was performed in advance of any data’s being received at the coordination center; mapping was done at the data element level before being carried out at the data level. Data sets were extracted from each contributing study’s database(s) based upon the Data Request Packet (see Web...
Appendix 2) and securely transmitted to COMPASS, where the data were ingested and transformed into the harmonized GECCO CDEs.

PROSPR, on the other hand, took the second route to mapping and harmonization. Each year, PROSPR research centers received a Data Request Packet containing data dictionaries of CDEs and CDE definitions. It was the responsibility of the PROSPR research centers to map and transform data from their repositories into the defined CDEs. One reason for doing this was that many of the participating studies were pulling data from electronic medical records, in which data can be distributed across multiple data systems, not simply stored in a single study database. Here, the data are so complex that it made the most sense for those who were deeply involved with the data to make the mapping decisions.

**Step 6: Quality-control procedures**

Quality control is the next step in data harmonization, as with any data set. The harmonization team needs to assess the harmonized data to ensure it is in the correct format and falls within expected value ranges. Potential issues include: data points that are outliers, unexpected values that are not in the CDE definition, missing data, incorrect data types, unexpected duplicate or orphan records, and illogical data distributions. If eligibility criteria exist for inclusion in the harmonized data set, eligibility logic checks must also be done. If data fail these checks, study investigators might be contacted to discuss data issues and/or clarify mapping logic; this could then lead to additional iterations of data delivery, transformation, and quality control, until the harmonization team is satisfied that they have high-quality data. As will be discussed below, this is an area that is ripe for further methodological research. What exactly do high-quality harmonized data look like? Are there specific approaches that can be used?

**A DATA-HARMONIZATION RESEARCH AGENDA**

A central question of data harmonization, as with all science, is that of reproducibility. As pooled analyses continue to grow in both size and scope, it is imperative that epidemiologists be confident that their approach to harmonization is based on rigorous, empirical evidence. To that end, we suggest a research agenda on the subject of data harmonization and pooled analyses. We start with 4 questions:

1. *Can we standardize and optimize the harmonization process, both scientifically and organizationally, to increase rigor while minimizing costs?*

Currently, each project harmonizes data from pooled studies in its own way. As such, we have no guarantee that another team working with the same data would harmonize and derive the same results. We have here documented one group’s harmonization process, but it is essential that other groups provide details on their own harmonization processes and that we then conduct research into how comparable those processes are. Furthermore, we need to balance this focus on rigor with organizational and financial constraints. Harmonization is expensive (see Web Appendix 3) and time-consuming. Are there ways we can reduce these costs both to the studies providing data and to those doing the harmonization?

2. *What measures can be used to judge how successful a harmonization has been?*

In addition to improving the process of harmonization, we propose that there is a need for research into the development of objective measures of the results. Are there ways to compare the results of our harmonization with results from individual studies or with a simulation of the process using idealized data? By doing so, the comparison becomes a metric of the harmonization process itself, ensuring that results reflect true differences in the data rather than differences between forms of data collection.

3. *What is the minimal (and optimal) level of harmonization that is needed to ensure correct inferences?*

Whereas the process described in this paper uses CDEs developed for each project, often reusing those CDEs developed for previous, similar projects, other methods are possible. For example, the approach of the Public Population Project in Genomics and Society (P²G), an organization that develops tools to promote collaboration and data harmonization, is to develop standardized CDEs for the research community to use in harmonization (13). Furthermore, the level at which the data are harmonized generally depends on 1) what data are available and 2) how much time and money are available for the pooled analyses. It would be a great boon to researchers to have empirical evidence that points to 1 type or level of harmonization as optimal or tells us that harmonization can be successful at any of these levels, given the right approach. In fact, if looser harmonization leads to the same results as a more rigorous harmonization process, substantial time and effort could be saved in pooled analyses. For any 1 variable, there will be degrees of acceptable harmonization, not just a dichotomy of adequate/inadequate, though even that would be helpful if it could be applied rigorously. Just as crucially for any specific analysis is the question of how many variables need to be adequately harmonized in order to allow correct inferences. A well-defined measure of harmonization will allow us to approach an optimal level of harmonization.

4. *What level and types of documentation are needed from collaborating studies to ensure maximal efficiency of the harmonization process?*

Data harmonization can be a slow, painful, and difficult process, as it often requires the elucidation of tacit knowledge or digging into the study archives to understand how the study was done. As more pooled analyses are conducted, we can seek ways to reduce that burden by asking for answers to many of the questions up front, reducing the back-and-forth required of busy investigators. It is simply impossible to anticipate all future uses of a data set and to document appropriately for all of them; however, if we can identify a series of questions that are most frequently asked when conducting pooled analyses, we can increase efficiency.

**CONCLUSION**

Data harmonization is a required part of all pooled analysis projects, both retrospective and prospective; yet the
harmonization process is currently done without the benefit of empirical evidence to support existing methods, and reproducibility of any one group’s processes is questionable. In this paper, we have described how one group has approached data harmonization within 4 large consortia and laid out the beginnings of an agenda for future research. By doing so, we hope to spark both examination of the data-harmonization process and interest in developing rigorous and consistent methods of harmonization.

ACKNOWLEDGMENTS

Author affiliations: Cancer Prevention Fellowship Program, National Cancer Institute, Bethesda, Maryland (Betsy Rolland); Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington (Betsy Rolland, Suzanna Reid, Deanna Stelling, Greg Warnick, Mark Thornquist, John D. Potter); Department of Biostatistics, MD Anderson Cancer Center, Houston, Texas (Ziding Feng); Centre for Public Health Research, Massey University, Wellington, New Zealand (John D. Potter); and Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington (John D. Potter).

This work was supported by the Fred Hutchinson Cancer Research Center and National Institutes of Health grant U24-086368 (Z.F.).

Conflict of interest: none declared.

REFERENCES


