International Journal of Population Data Science

Journal Website: www.ijpds.org

Data harmonization and data pooling from cohort studies: a practical approach for data management

Kamala Adhikari[1,*], Scott B Patten[1], Alka B Patel[1,2], Shahirose Premji[3], Suzanne Tough[1,4], Nicole Letourneau[1,4,5,6], Gerald Giesbrecht[1,4], and Amy Metcalfe[1,7,8]

Submission History

Submitted:	10/05/2021
Accepted:	17/08/2021
Published:	30/11/2021

[1]Department of Community Health Sciences, University of Calgary, Calgary, Canada
[2]Applied Research and Evaluation- Primary Health Care, Alberta Health Services, Calgary, Canada
[3]School of Nursing, Faculty of Health, York University, Calgary, Canada
[4]Department of Pediatrics, University of Calgary, Calgary, Canada
[5]Faculty of Nursing University of Calgary, Calgary, Canada
[6]Deprtment of Psychiatry, University of Calgary, Calgary, Canada
[7]Department of Obstetrics and Gynecology, University of Calgary, Calgary, Canada
[8]Department of Medicine, University of Calgary, Calgary, Canada

Abstract

Abstract

Data pooling from pre-existing datasets can be useful to increase study sample size and statistical power in order to answer a research question. However, individual datasets may contain variables that measure the same construct differently, posing challenges for data pooling. Variable harmonization, an approach that can generate comparable datasets from heterogeneous sources, can address this issue in some circumstances. As an illustrative example, this paper describes the data harmonization strategies that helped generate comparable datasets across two Canadian pregnancy cohort studies: All Our Families; and the Alberta Pregnancy Outcomes and Nutrition.

Variables were harmonized considering multiple features across the datasets: the construct measured; question asked/response options; the measurement scale used; the frequency of measurement; timing of measurement, and the data structure. Completely matching, partially matching, and completely un-matching variables across the datasets were determined based on these features. Variables that were an exact match were pooled as is. Partially matching variables were harmonized or processed under a common format across the datasets considering the frequency of measurement, the timing of measurement, the measurement scale used, and response options. Variables that were completely unmatching could not be harmonized into a single variable.

The variable harmonization strategies that were used to generate comparable cohort datasets for data pooling are applicable to other data sources. Future studies may employ or evaluate these strategies, which permit researchers to answer novel research questions in a statistically efficient, timely, and cost-efficient manner that could not be achieved using a single data source.

Keywords

data harmonization; data pooling or combination; comparable dataset; cohort studies; harmonization strategies

*Corresponding Author:
Email Address: kamala.adhikaridahal@ucalgary.ca (Kamala Adhikari)

https://doi.org/10.23889/ijpds.v6i1.1680

Introduction

Data pooling from multiple studies into a single dataset provides opportunities to increase the statistical power of a study and to answer novel research questions that could not be addressed using data from a single study [1, 2]. Data pooling from existing data sources allows investigators to conduct research more rapidly and at a lower cost than primary data collection would allow, providing opportunities for timely translation of knowledge into practice.

Individual datasets from different studies or data sources often measure the same construct differently, which poses challenges for data pooling. These challenges are addressed by data harmonization. Data harmonization refers to efforts that provide comparability of datasets from heterogeneous sources and allows for combining, pooling, or integrating them in a coherent way [3].

Data harmonization can take a prospective or retrospective approach. Prospective data harmonization occurs at the initial stage of study design, or at least before data collection. For this, investigators agree on a common core set of variables or measures, compatible data collection tools, and standard operating procedures, often leading to a high degree of homogeneity [3, 4]. Retrospective harmonization is a flexible approach, which targets the synthesis of already-collected information. For this, researchers define a core set of variables, and then assess the compatibility of information collected and the potential for creating single harmonized variables. If harmonization is possible, strategies for data processing are developed [3–6].

Data harmonization is particularly valuable when the outcome and/or risk factor is rare, since examining interactions among risk factors and investigating population subgroups requires a large sample size to ensure adequate study power. It is not always feasible to accomplish this with primary data collection from a single study given the resources required. Additionally, measurement of the same construct using multiple measurement scales is generally unreasonable or unfeasible for a single study unless the primary aim of the study is to compare the results from the multiple scales.

The use of multiple existing datasets from the studies that were conducted in similar target populations using comparable methodologies but different measurement scales can address these issues. However, data harmonization (in this case, retrospective) involves extensive data processing or data cleaning and management and variable transformation processes. While these processes are critical [6], the literature or guidelines on how to do this remain limited [6].

Our research project aimed to improve the understanding of risk factors for preterm birth using data from two pregnancy cohort studies conducted in Alberta, Canada–All Our Families (AOF: n = 3,351) and Alberta Pregnancy Outcomes and Nutrition (APrON: n = 2,187) [7–10]. Specifically, our research intended to develop and validate a prediction model for preterm birth, to evaluate the suitability of and comparability of multiple anxiety scales to measure anxiety during pregnancy, and to examine if neighborhood socioeconomic status modified the association between anxiety and/or depression status during pregnancy and preterm birth [11–13]. Achieving these goals required data harmonization.

This paper describes the data harmonization strategies that helped generate comparable datasets across these two studies to address our research objectives. It presents examples of data harmonization strategies that were used to generate comparable datasets. These strategies may be employed or evaluated in subsequent studies, and may serve as useful starting points for other projects.

Methods

Data sources

We obtained two de-identified datasets from the two prospective pregnancy cohort studies (AOF: n = 3,351 and APrON: n = 2,187). Both datasets are available for secondary analysis and are housed in SAGE (Secondary Analysis to Generate Evidence), a secure data repository developed by PolicyWise for Children & Families, which houses these datasets (https://policywise.com).

The AOF and APrON studies are ongoing cohort studies of mother and child dyads. Both cohort studies use quality control procedures to maintain the quality of study-specific data. To illustrate, both studies use data management standards for data storage, data entry, data dictionary and data cleaning. The data are double-entered by trained research assistants with discrepancies resolved by a master coder. All implausible or unusual values are re-entered to verify the data. In some cases, participants are contacted for clarification, and in other cases, the studies collected additional information that allowed them to correct implausible values. Where such corrections are not possible, the data are set to missing.

Each dataset was linked (by SAGE) with neighbourhood socioeconomic status measured by both the average household income and the Pampalon material deprivation index. Both measures were derived from 2011 Statistics Canada census data [14–16].

The AOF and APrON studies are comparable in many ways including target population, recruitment time periods, inclusion criteria, sampling design, data collection methods, cohort characteristics (such as age, income, and parity), and participant follow-ups and retention during the perinatal period (Supplementary Table 1) [7–10]. Both studies collect data about mothers, children, and partners, using methods including questionnaires, health records, and lab samples.

Given the similarity between the study populations and methodologies, pooling data from these studies was justifiable [1]. However, each study measured/recorded the same construct/variables differently and therefore, data harmonization strategies were used to generate a comparable dataset across the studies.

Data harmonization focused only on the maternal data obtained from questionnaires. Both studies collected data using questionnaires on perinatal health, including maternal demographics, socioeconomic status, lifestyle, social support, depression, anxiety, and preterm delivery [7–10]. Details on the description and comparability of these cohort studies is available elsewhere [7–10], and are summarized in Supplementary Table 1.

Table 1: Variable harmonization

Variables	AOF cohort dataset	APrON cohort dataset	Harmonization process	Variables combined (and recoded if needed)
Maternal age	Variable name: Q1MMAGE2 Construct: Maternal age at recruitment Type of data: continuous Missing: . (period)	Variable name: MAQ Construct: Maternal age at recruitment Type of data: continuous Missing: 999	• Complete matching of construct • Complete matching of response or data type and coding, except missing value coding (partial matching) Action taken: Coded missing data on APrON as . (period) and both variables renamed with same name	Maternal age variables with continuous data combined and recoded as • <35 years • ≥35 years • . Missing
Marital status	Variable name: Q1MMSTAT1 Construct: Current marital status Data type: Categorical Response category and value level: • 1 Single • 2 Single with partner • 3 Married • 4 Common-law • 5 Divorced • 6 Separated • . Missing	Variable name: MAGB1 Construct: Current marital status Data type: Categorical Response category and value level: • 0 Single • 1 Married • 2 Divorced • 3 Common-law • 4 Widowed • 5 Separated • 999 Missing	• Complete matching construct • Partial matching of variable response and coding Action taken: Recoding AOF: Combined single and single with partner response into "single" and combined divorced, widowed and separated response into divorced/ separated/widowed, APrON: Combined divorced, widowed and separated response into divorced/ separated/widowed. Variable in both datasets were recoded as: • 0 Single • 1 Married/common-law • 2 Divorced/separated/ widowed • . Missing Variable renamed with same name	Variables with the following categories combined • 0 Single • 1 Married/common-law • 2 Divorced/separated/widowed • . Missing
Maternal ethnicity	Variable name: Q1METH1_2 Construct: Ethnic origin Data type: Categorical Response category and value level: • 0 Others • 1 White/Caucasian	Variable name: MAGB16 Construct: Ethnic origin Data type: Categorical Response category and value level: • 1 Caucasian • 2 Chinese • 3 Filipino • 4 Japanese • 5 Korean • 6 Latin American • 7 Aboriginal/Native • 8 South Asian • 9 South East Asian • 10 Arab • 11 West Asian • 12 Black • 13 Others	• Complete matching construct • Partial matching of variable response and coding Action taken: Recoding APrON: Combined coding 2–13 into "others" and recoded as • 0 Others • 1 White/Caucasian • Variable renamed with same name	Variables with the following categories combined • 0 Others • 1 White/Caucasian
Body mass index	Variable name: Q1MHW8 Construct: Pre-pregnancy weight in kg Variable name: Q1MHW5 Construct: Height in cm Data type: continuous Missing: .	Variable name: MAANTH2 and MBANTH2 Construct: Pre-pregnancy weight in kg Variable name: MAANTH3 and MBANTH Construct: Pre-pregnancy height in cm Data type: continuous Missing: 999	• Complete matching construct • Partial matching of data coding or management system Action taken: Variable managed and body mass index calculated AOF: Calculated body mass index APrON: • Combined 2 weight variables into one • Combined 2 height variables into one • Recoded missing (999) into (.) • Calculated body mass index	Combined continuous body mass index variable and recoded as 4 categories • 0 Underweight <18.5 • 1 Normal weight 18.5–24.9 • 2 Overweight 25–29.9 • 3 Obese 30+

(Continued)

Table 1: Continued

Variables	AOF cohort dataset	APrON cohort dataset	Harmonization process	Variables combined (and recoded if needed)
Parity	Variable name: Q1MPPI1_1 Construct: Parity (birth to a fetus >24 weeks) Data type: Categorical Response category and value level: • 0 No previous births • 1 Previous birth to a fetus (at least once) • . Missing If previous birth to a fetus, number of live births • 1 to 7 • Missing (.)	Variable name: MAPI3 Construct: Live born children have you had Data type: Categorical Response category and value level: • 0 to 4 • missing (999)	• Complete matching construct • Partial matching variable response and coding Action taken: Recoding In both datasets, responses were recoded as • 1 Primiparous • 2 Multiparous • 3 Grand multiparous (>2 live births) • . "missing"	Variables with the following categories combined • 1 Primiparous • 2 Multiparous • 3 Grand multiparous • . Missing
Depression during pregnancy	Variable name: Q1MEDPS Construct: EPDS score in first measurement (during recruitment: <24 weeks of gestation) Variable name: Q2MEDPS Construct: EPDS score in second measurement (in third trimester: 34–38 weeks gestation)	Variable name: MAEPDS_Score Construct: EPDS score in first measurement (during recruitment: <27 weeks of gestation) Variable name: MBEPDS_Score Construct: EPDS score in second measurement (in 14–26 weeks of gestation for those participants who were 0–13 weeks of gestation during the recruitment) Variable name: MCEPDS_Score Construct: EPDS score in third measurement (in 27–40 weeks of gestation for those who were 0-26 weeks of gestation during recruitment)	• Complete matching construct • Partial matching in terms of number of measurements and measurement time during pregnancy (week of gestation) Action taken: In both datasets, using the recorded week of gestation at first, second and third measurements, 3 variables of EPDS score for each trimester were created. • EPDS score in first trimester • EPDS score in second trimester • EPDS score third trimester	Three combined variables for depression during pregnancy • EPDS score in first trimester • EPDS score in second trimester • EPDS score third trimester
Anxiety during pregnancy	Variable name: Q1MSSAI Construct: anxiety score in first measurement (during recruitment: <24 weeks of gestation), measured by STAI-20 Variable name: Q2MSSAI Construct: anxiety score in second measurement (in third trimester: 34-38 weeks gestation), measured by STAI-20	Variable name: MASCL_Score Construct: anxiety score in first measurement, measured by SCL-90 (during recruitment: <27 weeks of gestation) Variable name: MBSCL_Score Construct: anxiety score in second measurement, measured by SCL-90 (in second trimester: 14–26 weeks of gestation for those participants who were 0–13 weeks of gestation during the recruitment) Variable name: MCSCL_Score Construct: anxiety score in third measurement, measured by SCL-90 (in third trimester: 27–40 weeks for those who were 0–26 weeks of gestation during recruitment)	• Completely un-matching variable Action taken: • Harmonized anxiety score measured by each scale for each trimester using the same process for depression during pregnancy. Accordingly, three separate variables for anxiety during pregnancy by trimester (as for depression) for each anxiety scale were created. • Overlapped participants and their anxiety data measured by both scales identified.	Anxiety data measured by two different scales were pooled as two different variables • For 231 participants who participated both studies, each variable contained anxiety data. • For independent participants, each variable contained missing values if they did not have anxiety data measured by the same scale.
Anxiety during pregnancy, measured by EPDS-3A	Variable name: Q1MEDPS Construct: EPDS score (comprising EPDS-3A anxiety score) in first measurement (during recruitment: <24 weeks of gestation)	Variable name: MAEPDS_Score Construct: EPDS score (comprising EPDS-3A anxiety score) in first measurement (during recruitment: <27 weeks of gestation)	• Complete matching construct • Partial matching in terms of number of measurements and measurement time during pregnancy (week of gestation)	Three combined variables for anxiety during pregnancy • EPDS-3A score in first trimester • EPDS-3A score in second trimester • EPDS-3A score third trimester

(Continued)

Table 1: Continued

Variables	AOF cohort dataset	APrON cohort dataset	Harmonization process	Variables combined (and recoded if needed)
	Variable name: Q1MEDPS Construct: EPDS score (comprising EPDS-3A anxiety score) in first measurement (during recruitment: <24 weeks of gestation)	Variable name: MBEPDS_Score Construct: EPDS score (comprising EPDS-3A anxiety score) in second measurement (in second trimester: 14-26 weeks of gestation for those participants who were 0-13 weeks of gestation during the recruitment)	Action taken: In both datasets, we created the compatible anxiety variables, by extracting the data on three items of the EPDS (i.e., anxiety items 3, 4, and 5) measured by both studies. The three items comprise the anxiety subscale (EDPS-3A)	
		Variable name: MCEPDS_Score Construct: EPDS score (comprising EPDS-3A anxiety score) in third measurement (in third trimester: 27-40 weeks for those who were 0-26 weeks of gestation during recruitment)	In both datasets, using the recorded week of gestation at first, second and third measurements, 3 variables of EPDS-3A score for each trimester were created. • EPDS-3A score in first trimester • EPDS-3A score in second trimester • EPDS- 3A score third trimester	

Note: AOF: All Our Families; APrON: Alberta Pregnancy Outcomes and Nutrition; EPDS: Edinburgh Postnatal Depression Scale; STAI-20: State-Trait Anxiety Inventory-State 20-item scale; SCL-90: Symptoms Checklist-90; EPDS-3A: Edinburgh Postnatal Depression scale- anxiety subscale.

Variable harmonization

Study documentation from the AOF and APrON studies (such as study protocols and standard operating procedures, questionnaires and instrument calibration procedures, data dictionaries, and published papers) were accessed and reviewed. Conversations between our research team and the AOF/APrON research teams enabled an understanding of the level of substantive heterogeneity (i.e., study methodologies and equivalence of variables to be harmonized) and data management systems across studies [6]. Agreement on data access and intellectual property from each study and ethics approval from the Conjoint Health Research Ethics Board at the University of Calgary were obtained before data harmonization. We also performed preliminary exploration of each dataset before initiating the actual harmonization to further understand the constructs, questions, responses, variables available in the datasets, data distributions and value labels, or the data quality and comparability [6]. These strategies facilitated the identification and selection of variables to consider for harmonization and helped decide harmonization strategies to be employed.

Variables pertinent to address our research objectives were selected to consider for harmonization (Supplementary Table 2). These variables were harmonized in each dataset considering multiple features of the data, as recommended by previous authors [1–3, 5, 17]. These features included whether the variables were completely or partially identical regarding: (a) the construct measured; (b) question asked and response options; (c) the measurement scale used; (d) the frequency of measurement; (e) the timing of the measurement (i.e., when in pregnancy the variable was measured); and (f) the coding features of variables. The coding features of variables considered for data harmonization included: variable name, definition, type, format, and response categories; variable value label; and missing values, including response categories "not applicable", "not stated", and "don't know".

Multiple features of data were checked through the review of the documentations of the primary studies, the conversations with primary study research teams, and preliminary exploration of variables in the datasets. If the variables were found to have an exact match for each of these features, they were considered completely matching. If the variables were the same in terms of what construct was measured, but were different in terms of frequency of measurement, the timing of measurement, and variable response options and coding features, these variables were considered partially matching. These partially matching variables were harmonized or processed under a common format and, if needed, to the same frequency and timing of measurements across the datasets. Finally, some important variables did not match, and required a different approach (Table 1 and Supplementary Table 2).

If the construct was not measured in one of the datasets or if different measurement scales that emphasize the different components were used to measure the same construct across the datasets, the variables were deemed completely unmatching (Supplementary Table 2). In particular, the AOF dataset had data on anxiety during pregnancy measured by the State-Trait Anxiety Inventory-State 20-item scale (STAI-20), and the APrON dataset had anxiety data during pregnancy measured by the anxiety subscale of the Symptoms Checklist-90 (SCL-90). The variables comprising the anxiety data measured by these two different scales were important for our research that intended to compare the performance of multiple anxiety scales in measuring anxiety during pregnancy. Hence, we created anxiety data measured by two different scales as two different variables. We identified that there

were participants who participated in both cohort studies (n = 231) and their anxiety data measured by both scales (Table 1).

Anxiety data with a large sample size was critical for our research that aimed to examine effect modification between anxiety and/or depression status during pregnancy and neighborhood socioeconomic status on the risk of preterm birth. Since harmonization of direct measures of anxiety into a single variable was not feasible, we created comparable anxiety variables across studies by extracting data on three items of the Edinburgh Postnatal Depression Scale (EPDS) [18], which was used in both studies (Table 1). Specifically, items 3, 4 and 5 of the EPDS comprise an anxiety subscale (EPDS-3A), which has been suggested by previous studies as a measure of anxiety in the obstetric population [19, 20].

Documentation was created for variables across two datasets in terms of a variable name (a unique identity of the variable, e.g., smoking), variable definition (a short description of the variable, e.g., smoking status before pregnancy), variable value label (a short description of the response attributed to the underlying numerical values, e.g., "no" for 0, "yes" for 1), variable type (continuous or discrete), variable format (numeric or character), and missing value coding ("." or "999"). Once the selected variables in each dataset were harmonized and documented, the datasets were organized such that the same number of appending variables appeared in the same order for both datasets. Hence, the datasets were vertically identical by appending variables. Then, the two harmonized cohort datasets were concatenated into a single dataset (n = 5,538).

We used quality control procedures to test and describe the quality of harmonized data. Cross-tabulation or five-number summary (as appropriate to the data type) of each harmonized variable was done in each dataset to evaluate the consistency of those variables and the distribution of participants across the datasets (Supplementary Table 3). Variable formatting and descriptive statistics or distribution of participants were also assessed on the harmonized, combined datasets to explore any discrepancies with the variables on study-specific datasets.

Data harmonization procedures and the descriptive statistics of study-specific and combined data were documented as described above, and discussed with our research team and the broader AOF and APrON study teams. The discussion with the teams provided a qualitative validation of the data harmonization strategies used, a key step to make sure that the data harmonization process maintained the integrity of the original data and the original data were not lost. The discussion also facilitated to fix the errors (related to original variable coding or data entry) that were observed during the data harmonization process.

The final, harmonized data set was then used to answer our research objectives. Analytic approaches included regression analyses, structural equation modeling, and prediction model development and evaluation [11–13]. We imputed missing values, for the study variables that were not measured (thus contained missing values) in one cohort/dataset and also for those that were measured in both dataset with $\geq$5% missing data, from the predictive distribution based on the observed data.

Results

A total of 20 variables were considered for harmonization, and of those, 18 variables (90.0%) were successfully harmonized. Of 20 variables, three variables (15.0%) were completely matching and 14 (70.0%) were partially matching. These variables were successfully harmonized across the datasets and pooled/combined (i.e., appended into a single variable). One variable (5.0%) was completely unmatching across the datasets due to the different measurement scales used to measure the same construct, this variable was harmonized across the datasets for the purpose of data merging (i.e., pooling data as two different variables). Two variables (10.0%) were only available in one dataset; thus, variable harmonization was not applicable (Supplement Table 2). Characteristics (or distribution) of participants across the studies were similar in harmonized data, except drug abuse and smoking status (Supplementary Table 3). There were discrepancies in the proportion of missing data for some variables, particularly body mass index and gestational age at delivery. These discrepancies also existed in the original datasets; thus, they were not related to the data harmonization process.

Several partially matching variables such as marital status, ethnicity, income, parity, depression, and smoking were successfully harmonized (Supplement Table 2). For example, one variable, current marital status, was partially identical across the datasets as the construct measured (or question asked) was completely identical across both datasets but the variable response categories and the value level coding were different across the datasets. As the variable response categories were collapsible to identical and meaningful categories across the datasets, the variable response was re-organized into three identical categories in both datasets. Another variable, depression symptoms during pregnancy - which was measured in both datasets using the same scale, the EPDS - was not compatible in terms of frequency of measurement and gestational age at each measurement. Accordingly, the depression variables were harmonized by creating three unique variables in each dataset that indicated the depression score in each trimester of pregnancy.

Similarly, the EPDS-3A-based anxiety variables, which were made by extracting data on three items of the EPDS, were harmonized by creating three unique variables in each dataset that indicated the anxiety score in each trimester of pregnancy. The anxiety variables measured by two different anxiety scales (i.e., STAI-20 and SCL-90) were harmonized by creating three unique variables in each dataset that indicated the anxiety score (measured by different scales across the datasets) in each trimester of pregnancy (Table 1).

The harmonized combined cohort dataset (n = 5,538) contained several important variables, including maternal age, gestational age at delivery, marital status, ethnicity, duration of stay in Canada, body mass index, parity, smoking, and anxiety (measured by EPDS-3A), and depression during pregnancy for each trimester. Additionally, variables that were important for our research but were only available in one of the datasets (previous preterm birth and prenatal care visits) or measured by different anxiety measurement scales (anxiety during pregnancy) were included in the combined dataset.

The anxiety data measured by two different scales across the datasets were pooled as two different variables, with

missing values recorded for measures on the scale not included in the original study. Anxiety data or values were available for both anxiety-related variables for participants who participated in both cohort studies (overlapping study cohort, n = 231) (Table 1). Similarly, the combined dataset contained missing values for the cohort with no measurement of previous preterm birth and prenatal care visits variables.

Discussion

This study describes data harmonization strategies, which helped create comparable datasets across two cohort studies and enabled the datasets pooling. The combined dataset created unique research opportunities to answering our clinically relevant research questions, by providing a large sample size (thus increased study power and efficiency), additional variables, and data measured by multiple different scales [11–13]. The use of the harmonized, combined dataset facilitated statistical analysis to answer our research questions and added comprehensiveness to our research, which would have been less feasible using either of the datasets alone.

For example, the large sample size provided an opportunity to analyze the risk of preterm birth (relatively a rare outcome) across the several strata of risk factors, such as anxiety alone, depression alone, and both anxiety and depression and their stratification across socioeconomic variables [13]. Similarly, we evaluated the performance of multiple anxiety scales in measuring anxiety during pregnancy: the suitability of STAI- 20 and SCL-90 anxiety screening scales in the individual study cohort and the comparability of these scales (correlation between the anxiety scores measured by two scales) restricting our analysis in the overlapping study cohort [12]. We performed analyses including those variables that were available in both datasets [11–13]. We also performed sensitivity analyses using the additional variables available in one dataset [11–13].

The harmonized data are stored in a secure data repository (SAGE - Secondary Analysis to Generate Evidence) which also houses the cohort-specific datasets. The dataset may be available upon request from the AOF and APrON data custodians. The harmonization strategies described are applicable to generate comparable data across administrative databases, survey cycles, jurisdictions (provincial, national or international), and measures repeated over time. However, the strategies may not be necessarily directly applicable to different contexts, such as harmonizing data from a larger number of studies or data sources. Heterogeneity across studies or datasets becomes more persistent and data harmonization process becomes complex as the number of datasets or data sources increases. Nevertheless, it may be worthwhile to evaluate the utility and applicability of these strategies in subsequent studies.

The success or the scientific impact of any data harmonization and integration research project depends on the quality of the data harmonization process, the quality of the information collected by the primary studies, and the ability to access the data collected [1–6, 17]. Hence, a series of procedures should be considered as a part of the data harmonization and synthesis initiatives to ensure the quality and validity of the harmonized databases created. To illustrate, the potential to harmonize and integrate information depends on homogeneity across a range of study-specific factors. These include the study design, target population, time period, and duration of follow-up; the type of information and samples collected; the specific tools and standard operating procedures used to collect or generate data; and the data coding and data management systems employed. The incompatibility of these study-specific factors can affect whether variables recorded in different data sources are actually measuring the same construct. Access to documentation from the primary studies, dialogue with their research teams, and preliminary exploration of the dataset before the actual harmonization allow researchers to understand the level of substantive heterogeneity across studies [5, 6]. These strategies ultimately facilitate the selection of variables to be harmonized or combined and helps decide harmonization strategies to be employed.

Additionally, agreement on data access and intellectual property from each study and ethical approval must be obtained before data harmonization. Finally, it is important that the person(s) involved in data harmonization always create new files for the harmonization and document the harmonization process [17]. This facilitates the evaluation of the data harmonization process and reproducibility.

While the need for additional statistical power has often led investigators to employ data harmonization and data pooling, there are several other benefits as well [2–4]. These include increased use of existing data, strengthening the scientific impact of individual studies, and optimal return on investments. To illustrate, compared to building new studies involving thousands of participants, employing data harmonization on existing data can permit the generation of research projects relatively rapidly and at a lower cost, with timely knowledge translation opportunities. This also allows researchers to properly explore similarities and differences across time and place. Ultimately, data harmonization initiatives leverage national and international collaborations, facilitates the emergence of leading-edge collaborative and cross-disciplinary research initiatives and innovations, and thereby minimizes the duplication of research efforts [21].

While data harmonization is an important component in research, its application (harmonization process and harmonized data) has some challenges and limitations. Recent publications provide high-level guideline on harmonization [5, 6], but literature on how to perform data processing and evaluate harmonization quality (practical approaches) is limited [5, 6]. The data harmonization process is resource intensive. It involves a repetitive/iterative and time-consuming process, requires thorough preparatory work, and has many elements that must be worked through carefully and systematically with rigorous documentation.

To illustrate, data processing and integration in a systematic manner requires a comprehensive understanding of previous studies (study-specific designs, standard operating procedures, data collection devices, data format and data content, and quality of study-specific data) and requires research content knowledge and analytical skills [5, 6]. Even if harmonization procedures (variable selection and pairing rules definition and data processing) are done under the consensus and advice from experts, there is inevitably an element of

subjectivity in harmonization procedures. Evaluation of the quality of the harmonized data is required to understand its scientific performance [6]. At least two independent individuals are needed to evaluate inter-coder agreement with regard to their data harmonization procedures or processes (such as Cohen's k statistic) [5]. Furthermore, data harmonization may lead to limited use of information (in terms of the number of variables, variable categories) collected by specific primary studies.

For example, in our research context, maternal ethnicity and household income variables were categorized differently across datasets, broad categories vs. specific categories. Using harmonized data, we had to analyze the data by broad categories. We also had to analyze the anxiety data on the subsample. In contrast, the use of a single study or dataset is less resource-intensive, with more flexibility on using the information collected by primary studies, but has other limitations as described. Additionally, the harmonization strategies used in one context may not necessarily be directly applicable to different contexts due to the variation in heterogeneities, such as large number of studies or data sources and heterogeneous target population and data collections and management systems across studies.

Conclusion

Data harmonization is an important aspect of conducting research using multiple datasets. It generates comparable data across different data sources and facilitates pooling of relevant data across data sources, leading to unique opportunities for research. Data harmonization and pooling augment the utility and scientific impact of existing data or individual studies, creates a collaborative research environment, minimize the duplication of research, and increase research feasibility. Hence, data harmonization is a very promising avenue to support advancement in population health research that can result in improvements to the health and well-being of populations.

Ethical statements

Ethics approval for this study was obtained from the Conjoint Health Research Ethics Board at the University of Calgary (REB16-2548). This study used secondary data and all the data were anonymized; therefore, did not require informed consent.

Competing interests

The authors declare that they have no competing interests.

References

1. Roberts G and Binder D. Analyses Based on Combining Similar Information from Multiple Surveys. Section on Survey Research Methods Joint Statistical Meetings (JSM); 2009. p.2138-47.

2. Rao SR, Graubard BI, Schmid CH, Morton SC, Louis TA, Zaslavsky AM, et al. Meta-analysis of survey data: application to health services research. Health Services and Outcomes Research Methodology. 2008;8(2):98–114.

3. 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):261–4; author reply 5-6.

4. Fortier I, Doiron D, Wolfson C, Raina P. Harmonizing data for collaborative research on aging: Why should we foster such an agenda? Canadian Journal of Aging. 2012;31:95–99.

5. Fortier I, Doiron D, Little J, Ferretti V, L'Heureux F, Stolk RP, et al. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. Int J Epidemiol. 2011; 40:1314–1328.

Box 1: Data harmonization best practices or key lessons

1. Appraisal of published and unpublished documents of the primary studies and conversations with the primary study teams, to gain in-depth understanding regarding the primary study methodologies and facilitate the judgement around the homogeneity of study- or data sources-specific factors.
2. Agreement on data access and intellectual property from each study and ethical approval before data harmonization.
3. Preliminary exploration of variables in the datasets before initiating the actual harmonization to understand the variables available in the datasets, the variable coding or data management, the data distributions, or the data quality and comparability.
4. Identification of completely unidentical and completely or partially identical variables across data sources, and variable harmonization, considering multiple features such as construct measured, measurement scale used, cross-sectional or longitudinal measurement, data coding, and overlapping samples.
5. Establishing the consistency of variables across two datasets before data combination.
6. Preserve the integrity of the original data and ensure that the original data are not lost, while seeking to harmonize variables to address own research purposes and exploring unique research opportunities such as overlapping samples and data measured by multiple scales.
7. Documentation of data harmonization procedures and sharing/discussing it with the primary study teams, seek suggestions on data harmonization procedures used and solutions for data errors observed in the original datasets during the data harmonization process.

6. Fortier I, Raina P, Heuvel ERVd, Griffith LE, Craig C, Saliba M, et al. Maelstrom research guidelines for rigorous retrospective data harmonization. Int J Epidemiol. 2017;46 (1):103–105.

7. Kaplan BJ, Giesbrecht GF, Leung BM, Field CJ, Dewey D, Bell RC, et al. The Alberta Pregnancy Outcomes and Nutrition (APrON) cohort study: rationale and methods. Matern Child Nutr. 2014;10(1):44–60.

8. Leung BM, McDonald SW, Kaplan BJ, Giesbrecht GF, Tough SC. Comparison of sample characteristics in two pregnancy cohorts: community-based versus population-based recruitment methods. BMC Med Res Methodol. 2013;13:149.

9. McDonald SW, Lyon AW, Benzies KM, McNeil DA, Lye SJ, Dolan SM, et al. The All Our Babies pregnancy cohort: design, methods, and participant characteristics. BMC Pregnancy Childbirth. 2013;13 Suppl 1:S2.

10. Tough SC, McDonald SW, Collisson BA, Graham SA, Kehler H, Kingston D, et al. Cohort Profile: The All Our Babies pregnancy cohort (AOB). Int J Epidemiol. 2017;46(5):1389–90k.

11. Adhikari K, Patten SB, Williamson T, Patel AB, Premji S, Tough S, Letourneau N, Giesbrecht G, Metcalfe A. Does Neighbourhood Socioeconomic Status Predict the Risk of Preterm Birth? A Community-based Canadian Cohort Study. BMJ Open. 2019;9:e025341. https://doi.org/10.1136/bmjopen-2018-025341

12. Adhikari K, Patten SB, Williamson T, Patel AB, Premji S, Tough S, Letourneau N, Giesbrecht G, Metcalfe A. Assessment of Anxiety during Pregnancy Using Multiple Anxiety Scales: Do Anxiety Scales Differ in Their Ability to Assess Anxiety During Pregnancy? Journal of Psychosomatic Obstetrics & Gynecology. 2020:1–7.

13. Adhikari K, Patten SB, Williamson T, Patel AB, Premji S, Tough S, Letourneau N, Giesbrecht G, Metcalfe A. Neighbourhood Socioeconomic Status Modifies the Association between Anxiety and Depression during Pregnancy and Preterm Birth: A Community-based Canadian Cohort Study. BMJ Open. 2020;10::e031035. doi:10.1136/ bmjopen-2019-031035.13.

14. Pampalon R, Raymond G. A deprivation index for health and welfare planning in Quebec. Chronic Dis Can 2000;21:104–13

15. Alberta Health Services. How to use the Pampalon Deprivation Index in Alberta: Research and Innovation, Alberta Health Services, 2016.

16. Statistics Canada. 2011 Census Program. Retrieved on February 01, 2021, from: 2011 Census Program: Topics (statcan.gc.ca)

17. Kveder A, Galico A. Guideline for cleaning and harmonization of generations and gender survey data. Retrieved on February 01, 2021, from: http://www.unece.org/pau/ggp/materials.htm.

18. Cox JL, Holden JM, Sagovsky R. Detection of postnataldepression. Development of the 10-item EdinburghPostnatal Depression Scale. Br J Psychiatry. 1987;150:782–786.

19. Matthey S. Using the Edinburgh postnatal depression scale to screen for anxiety disorders. Depress Anxiety 2008;25:926–31.

20. Matthey S, Fisher J, Rowe H. Using the Edinburgh postnatal depression scale to screen for anxiety disorders: conceptual and methodological considerations. J Affect Disord 2013;146:224–30

21. Bergeron J, Rachel M, Stephanie A, Alan B, William F, Isabel F. Cohort Profile: Research advancement through cohort cataloguing and harmonization (ReACH). Int J Epidemiol. 2021;50(2):396–397.

Supplementary Table 1: Characteristics of All Our Family (AOF) and Alberta Pregnancy Outcomes and Nutrition (APrON) cohort studies

Study characteristics	**AOF study**	**APrON study**
Study design	Prospective cohort study	Prospective cohort study
Target population	Pregnant women	Pregnant women
Study location	Calgary, Alberta	Calgary and Edmonton, Alberta (approximately 85% of the sample came from Calgary)
Inclusion criteria	Age: $\geq$18 years Gestational age: <24 weeks Language: able to complete questionnaires in English	Age: $\geq$16 years Gestational age: <27 weeks Language: able to complete questionnaires in English
Recruitment period	2008–2011	2009–2012
Recruitment strategies	Community-based: face-to-face recruitment in maternity clinics, posters, and word of mouth, provincial laboratory services etc.	Community-based: face-to-face recruitment in maternity clinics, poster, pamphlets and word of mouth, newspaper, television etc.
Sampling strategies	Community-based, non-stratified sampling	Community-based, non-stratified sampling
Sample size	3,388 pregnant women	2,200 pregnant women
Retention rate until postpartum period	Approximately 90%	Approximately 85%
Follow-ups and data collection (using maternal survey questionnaire) on variables related to this research	– During recruitment, <24 weeks of gestation: socio-demographics, lifestyle, social support, depression, anxiety – Follow-up in third trimester, 34–38 weeks gestation: anxiety and depression measurement – Follow up survey at 4 months of postpartum period: gestational age at delivery for index pregnancy	– During recruitment, <27 weeks of gestation: socio-demographics, lifestyle, social support, depression, anxiety – Follow-up in 14-26 weeks of gestation for those participants who were 0–13 weeks of gestation during the recruitment and in 27–40 weeks for those who were 0–26 weeks of gestation during recruitment: anxiety and depression measurement – Follow up survey at 4 months of postpartum period: gestational age at delivery for index pregnancy

Supplementary Table 2: List of AOF and APrON variables related to our research and selected for harmonization

Variables considered	AOF cohort	APrON cohort	Variable matching	Harmonization success
1. Maternal age	Construct: Maternal age at recruitment Type of data: continuous Missing: . (period)	Construct: Maternal age at recruitment Type of data: continuous Missing: 999	• Complete matching of construct • Complete matching of response or data type and coding, except missing value coding	• Successfully harmonized • Data pooled into one variable (appended)
2. Marital status	Construct: Current marital Data type: Categorical • 1 Single • 2 Single with partner • 3 Married • 4 Common-law • 5 Divorced • 6 Separated • . Missing	Construct: marital status Data type: Categorical • 0 Single • 1 Married • 2 Divorced • 3 Common-law • 4 Widowed • 5 Separated • 999 Missing	• Complete matching construct • Partial matching of response & coding	• Successfully harmonized • Data pooled into one variable (appended)
3. Maternal ethnicity	Construct: Ethnic origin Data type: Categorical • 0 Others • 1 White/Caucasian	Construct: Ethnic origin? Data type: Categorical • 1 Caucasian • 2 Chinese • 3 Filipino • 4 Japanese • 5 Korean • 6 Latin American • 7 Aboriginal/Native • 8 South Asian • 9 South East Asian • 10 Arab • 11 West Asian • 12 Black • 13 Others	• Complete matching construct • Partial matching of response & coding	• Successfully harmonized • Data pooled into one variable (appended)
4. Duration of stay in Canada:	Construct: Time in Canada • 0 "Born in CA/lived 5+years" • 1 "Lived in CA less than years"	Construct: Born in Canada Yes/No. If no, how long have you lived in Canada? • 0 Less than 1 year • 2 1-3 years • 2 4-5 years • 3 over 5 years • 888 valid skip	• Complete matching construct • Partial matching response & coding	• Successfully harmonized • Data pooled into one variable (appended)
5. Body mass index	Construct: Pre-pregnancy weight in kg Construct: Height in cm Data type: continuous Missing: .	Construct: Pre-pregnancy weight in kg (records in two variables) Construct: Height (cm) (records in two variables) Data type: continuous Missing: 999	• Complete matching construct • Partial matching of data coding or management system	• Successfully harmonized • Data pooled into one variable (appended)
6. Parity	Construct: Parity (birth to a fetus >24 weeks) • 0 No previous births • 1 Previous birth to a fetus (at least once) • . Missing If previous birth to a fetus, number of live births • 1 to 7 • Missing (.)	Construct: Live born children have you had • 0 to 4 • missing (999)	• Complete matching construct • Partial matching response and coding	• Successfully harmonized • Data pooled into one variable (appended)
7. Intended pregnancy	Construct: When you became pregnant, were you trying to get pregnant? • 1 Yes (Intended) • 2 No (Unintended) • . Missing	Construct: were you purposely trying to become pregnant? • 0 No (Unintended) • 1 Yes (Intended) • 999 Missing	• Complete matching construct • Complete matching response, with unmatching response value coding (partial matching)	• Successfully harmonized • Data pooled into one variable (appended)

(Continued)

Supplementary Table 2: Continued

Variables considered	AOF cohort	APrON cohort	Variable matching	Harmonization success
8. Smoking	Construct: Smoking status before pregnancy (daily or occasionally) • 1 Yes • 2 No • . missing	Construct: Smoking status before pregnancy (daily or occasionally) • 0 No • 1 Yes • 999 missing	• Complete matching construct • Complete matching response, with unmatching response value coding (partial matching)	• Successfully harmonized • Data pooled into one variable (appended)
9. Alcohol consumption	Construct: Alcohol consumption before pregnancy • 1 Yes • 2 No	Construct: Alcohol consumption before pregnancy • 0 No • 1 Yes	• Complete matching construct • Complete matching response, with unmatching response value coding (partial matching)	• Successfully harmonized • Data pooled into one variable (appended)
10. Drug abuse	Construct: Street drug use before pregnancy • 1 Yes • 2 No	Construct: Recreational drug use before pregnancy • 0 No • 1 Yes	• Complete matching construct • Complete matching response, with unmatching response value coding (partial matching)	• Successfully harmonized • Data pooled into one variable (appended)
11. Maternal education	Construct: The highest level of education completed • 1 Some elementary or high school • 2 Graduated high school • 3 Some college, trade, university • 4 Graduated college, trade, university • 5 Some graduate school • 6 Completed graduate school • . Missing	Construct: The highest level of education completed • 1 Less than high school diploma • 2 Completed high school diploma • 3 Competed trade, technical diploma • 4 Completed university • 5 Completed post-grad • 999 Missing	• Complete matching construct • Partial matching responses and coding	• Successfully harmonized • Data pooled into one variable (appended)
12. Household income	Construct: Total income, before taxes and deductions, of all household members from all sources in the past 12 months • 1 Less than $10,000 • 2 $10,000- $19,999 • 3 $20,000- $29,999 • 4 $30,000- $39,999 • 5 $40,000- $49,999 • 6 $50,000- $59,999 • 7 $60,000 - $69,999 • 8 $70,000- $79,999 • 9 $80,000- $89,999 • 10 $90,000- $99,999 • 11 $100,000 or more • . Missing	Construct: Total income, before taxes and deductions, of all the household members from all resources? • 1 Less than $20,000 • 2 $20,000 to $39,999 • 3 $40,000 to $69,999 • 4 $70,000 to $99,999 • 5 $100,000 or more • 999 Missing	• Complete matching construct • Partial matching response and coding	• Successfully harmonized • Data pooled into one variable (appended)
13. Weeks of gestation	Construct: week of gestation in first measurement (during recruitment) Construct: week of gestation in second measurement (in third trimester) Data type: continuous	Construct: week of gestation in first measurement during (during recruitment) Construct: week of gestation in second measurement (in second trimester for those who had their first measurement in first trimester) Construct: Week of gestation in third measurement (in third trimester) Data type: continuous	• Complete matching Construct • Partial matching of frequency of measurement	• Successfully harmonized • Data pooled into one variable (appended)
14. Depression during pregnancy	Construct: EPDS score in first measurement (during recruitment: <24 weeks of gestation) Construct: EPDS score in second measurement (in third trimester: 34-38 weeks gestation)	Construct: EPDS score in first measurement (during recruitment: <27 weeks of gestation)	• Complete matching construct • Partial matching in terms of number of measurements and measurement time (week of gestation)	• Successfully harmonized • Data pooled into one variable (appended)

(Continued)

Supplementary Table 2: Continued

Variables considered	AOF cohort	APrON cohort	Variable matching	Harmonization success
		Construct: EPDS score in second measurement (in second trimester:14-26 weeks of gestation for those participants who were 0-13 weeks of gestation during the recruitment) Construct: EPDS score in third measurement (in third trimester: 27-40 weeks for those who were 0-26 weeks of gestation during recruitment)		
15. Anxiety during pregnancy	Construct: anxiety score in first measurement (during recruitment: <24 weeks of gestation), measured by SAI-20 Construct: anxiety score in second measurement (in third trimester: 34-38 weeks gestation), measured by SAI-20	Construct: anxiety score in first measurement (during recruitment: <27 weeks of gestation), measured SCL-90 Construct: anxiety score in second measurement (in second trimester:14-26 weeks of gestation for those participants who were 0-13 weeks of gestation during the recruitment), measured SCL-90 Construct: anxiety score in third measurement (in third trimester: 27-40 weeks for those who were 0-26 weeks of gestation during recruitment), measured SCL-90	• Completely un-matching construct, measured by different anxiety scales that emphasize different component of anxiety	• Successfully harmonized • Pooled anxiety data measured by two different scales as two different variables
Anxiety during pregnancy, measured by EPDS-3A	Construct: EPDS-3A score (from EPDS) in first measurement (during recruitment: <24 weeks of gestation) Construct: EPDS-3A score (from EPDS) in second measurement (in third trimester: 34-38 weeks gestation)	Construct: EPDS-3A score (from EPDS) in first measurement (during recruitment: <27 weeks of gestation) Construct: EPDS-3A score (from EPDS) in second measurement (in second trimester: 14-26 weeks of gestation for those participants who were 0-13 weeks of gestation during the recruitment) Construct: EPDS-3A (from EPDS) score in third measurement (in third trimester: in 27-40 weeks for those who were 0-26 weeks of gestation during recruitment)	• Completely matching construct. However, the EPDS-3A score variable was not readily available in the datasets. EPDS-3A is a part of the EPDS, where the 3 items (EPDS-3A: EPDS items 3, 5 and 6) are considered as items that measure anxiety during pregnancy, making an anxiety scale.	• Successfully harmonized • Data pooled into one variable (appended)
16. Social support during pregnancy	Construct: Social support status • 1 Adequate • 2 Inadequate	Construct: Social support status • Social support score	• Complete matching of construct • Partial matching response coding	• Successfully harmonized • Data pooled into one variable (appended)
17. Gestational age at delivery	Construct: Weeks gestation at delivery Data type: continuous Missing: . (period)	Construct: Weeks gestation at delivery Data type: continuous Missing: 999	• Complete matching of construct • Complete matching of response or data type and coding, except missing value coding	• Successfully harmonized • Data pooled into one variable (appended)
18. Prenatal care visit during pregnancy	Data available: number of prenatal care visit	Data unavailable	• Not applicable or completely unmatching	• Not applicable
19. History of preterm delivery	Data available: previous preterm delivery	Data unavailable	• Not applicable or completely unmatching	• Not applicable

Note: AOF: All Our Families; APrON: Alberta Pregnancy Outcomes and Nutrition; EPDS: Edinburgh Postnatal Depression Scale; STAI-20: State-Trait Anxiety Inventory-State 20-item scale; SCL-90: Symptoms Checklist-90; EPDS-3A: Edinburgh Postnatal Depression scale- anxiety subscale.

Supplementary Table 3: Characteristics of All our Family (AOF) cohort, Alberta Pregnancy Outcomes and Nutrition (APrON) cohort, and combined cohort

Variables	AOF (N = 3,351)		APrON (n = 2,187)		Combined (N = 5,538)	
	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI
Maternal age (years), n (m ± sd)	3245 (30.59 ± 4.56)	–	2182 (30.54 ± 4.51)	–	5427 (30.97 ± 4.56)	–
Maternal age	2612 (77.98)	76.54, 79.35	1700 (77.73)	77.94, 79.43)	4312 (77.13)	76..77, 78.95
<35yrs	633 (18.89)	17.60, 20.25	482 (22.04)	20.34, 23.83	1115 (20.13)	19.10, 21.21
≥ 35yrs	106 (3.13)	2.59, 3.78	5 (0.23)	0.09, 0.55	111 (1.99)	1.65, 2.39
Missing						
Marital status	185 (5.52)	4.80, 6.35	84 (3.84)	3.11, 4.73	269 (4.86)	4.32, 5.46
Single/divorced/separated	3131 (93.43)	92.54, 94.23	2017 (92.23)	91.02, 93.28	5148 (92.96)	92.25, 93.60
Married/common-law	35 (1.04)	0.75, 1.45	86 (3.93)	3.19, 4.83	121 (2.18)	1.83, 2.61
Missing						
Maternal ethnicity	2611 (77.92)	76.48, 79.29	1681 (76.86)	75.05, 78.58	4292 (77.50)	76.38, 78.58
White/Caucasian	705 (21.04)	19.69, 22.45	414 (18.93)	17.34, 20.63	1119 (20.21)	19.17, 21.28
Others	35 (1.04)	0.75, 1.45	92 (4.21)	3.44, 5.13	127 (2.29)	1.93, 2.72
Missing						
Duration of stay in Canada	2987 (89.14)	88.04, 90.15	1877 (85.83)	84.29, 87.22	4864 (87.83)	86.91, 88.66
Born/5 years+	319 (9.52)	85.71, 10.56	165 (7.54)	6.51, 8.73	484 (8.74)	8.02, 9.51
<5 years	45 (1.34)	1.00, 1.79	145 (6.63)	5.66, 7.75	190 (3.43)	2.98, 3.94
Missing						
Body mass index	24.34 ± 5.11	3.94, 5.36	24.20 ± 4.83	2.42, 3.89	24.29 ± 5.00	3.51, 4.54
Underweight (<18.5kg/m2)	154 (4.60)	58.22, 61.54	67 (3.06)	53.74, 57.90	221 (3.99)	56.98, 59.58
Normal weight (18.5–24.99)	2007 (59.89)	20.04, 22.82	1221 (55.83)	16.94, 20.20	3228 (58.29)	19.22, 21.34
Overweight (25-29.99	717 (21.40)	10.59, 12.77	405 (18.52)	8.65, 11.52	1122 (20.26)	10.13, 11.77
kg/m2)	390 (11.64)	2.00, 3.06	215 (9.83)	11.42, 14.22	605 (10.92)	5.91, 7.22
Obesity (≥30 kg/m2)	83 (2.48)		279 (12.76)		362 (6.54)	
Missing						
Parity	1615 (48.19)	46.50, 49.89	1184 (54.14)	52.04, 56.22	2799 (50.54)	49.44, 51.86
Primiparous	1689 (50.40)	48.71, 52.09	917 (41.93)	39.87, 44.01	2606 (47.06)	45.74, 48.37
Multiparous	47 (1.40)	1.05, 1.86	86 (3.93)	3.19, 4.84	133 (2.40)	2.03, 2.84
Missing						
Intended pregnancy	2668 (79.62)	78.22, 80.95	1708 (78.10)	76.31, 79.78	4376 (79.02)	77.92,80.07
Yes	649 (19.37)	18.06, 20.74	399 (18.24)	16.68, 19.91	1048 (18.92)	17.91, 19.98
No	34 (1.01)	0.73, 1.42	80 (3.66)	2.95, 4.53	114 (2.06)	1.72, 2.47
Missing						
Smoking before pregnancy	544 (16.23)	15.02, 17.52	571 (26.11)	24.31, 27.99	1115 (20.13)	19.09, 21.21
Yes	2776 (82.84)	81.52, 84.08	1530 (69.96)	68.00, 71.84	4306 (77.75)	76.64, 78.83
No	31 (0.93)	0.65, 1.31	86 (3.93)	3.19, 4.83	117 (2.11)	1.76, 2.53
Missing						
Alcohol consumption	2736 (81.65)	80.30, 82.92	1836 (83.95)	82.35, 85.43	4572 (82.56)	81.53, 83.53
Yes	592 (17.67)	16.41, 18.99	262 (11.98)	10.68, 13.41	854 (15.42)	14.49, 16.40
No	23 (0.69)	0.46, 1.03	89 (4.07)	3.31, 4.98	112 (2.02)	1.68,2.43
Missing						
Drug abuse	281 (8.39)	7.49, 9.37	483 (22.09)	20.39, 23.87	764 (13.80)	12.91, 14.72
Yes	3043 (90.81)	89.78, 91.74	1613 (73.75)	71.87, 75.55	4656 (84.80)	83.09, 85.01
No	27 (0.81)	0.55, 1.17	91 (4.16)	3.40, 5.08	118 (2.13)	1.78, 2.54
Missing						
Maternal education	118 (3.52)	2.95, 4.20	58 (2.65)	2.06, 3.41	176 (3.18)	2.74, 3.67
Less than high school	718 (21.43)	20.01, 22.85	200 (9.14)	8.01, 10.42	918 (16.58)	15.62, 17.58
Completed high school	2482 (74.07)	72.55, 75.52	1823 (83.36)	81.73, 84.86	4305 (77.74)	76.62, 78.81
≥High school	33 (0.98)	0.70, 1.38	106(4.85)	4.02, 5.83	139 (2.51)	2.13, 2.96
(trade/technical/university)						
Missing						
Household income	296 (8.83)	7.92, 9.84,	187 (8.55)	7.44, 9.79	483 (8.72)	8.01, 9.49
$<40,000	477 (14.23)	13.09, 15.46	279 (12.76)	11.42, 14.22	756 (13.65)	12.77,14.58
$40,000–70,000	789 (23.55)	22.14, 25.01	466 (21.31)	19.64, 23.07	1255 (22.66)	21.58, 23.78
$70,000–<100,000	1656 (49.42)	47.73, 51.11	1146 (52.40)	50.30, 54.49	2,802 (50.60)	49.28, 51.91
$≥100,000	133 (3.97)	3.36, 4.69	109 (4.98)	4.14, 5.98	242 (4.37)	3.86, 4.94
Missing						

(Continued)

Supplementary Table 3: Continued

Variables	AOF (N = 3,351)		APrON (n = 2,187)		Combined (N = 5,538)	
	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI
Gestational age (week of gestation), n (mean ± sd)[a]	3309 (16.36 ± 4.22)	–	2094 (17.02 ± 5.53)	–	5403 (16.61 ± 4.78)	–
Measurement 1	3031 (34.47 ± 1.49)		347 (18.96 ± 3.09)		3378 (32.88 ± 5.01)	
Measurement 2	NA		1807 (31.94 ± 1 .71)		1807 (31.94 ± 1.71)	
Measurement 3						
Anxiety measurement during pregnancy[a]	682 (20.35)	19.00, 21.76	392 (17.92)	16.34, 19.60	1074 (19.39)	18.36, 20.46
Trimester 1	2595 (77.43)	75.99, 78.85	1859 (85.00)	83.44, 86.47	4454 (80.43)	79.36, 81.46
Trimester 2	3027 (90.33)	89.28, 91.31	1810 (82.76)	81.11, 84.32	4837 (87.34)	86.43, 88.21
Trimester 3						
Depression measurement during pregnancy[a]	681 (20.32)	18.97, 21.72	391 (17.88)	16.29, 19.55	1072 (19.36)	18.32, 20.42
Trimester 1	2591 (77.32)	75.86, 78.73	1848 (84.50)	82.91 85.99	4439 (80.16)	79.08, 81.19
Trimester 2	3019 (90.09)	89.03, 91.08	1803 (82.44)	80.78, 84.01	4822 (87.07)	86.16, 87.94
Trimester 3						
Social support at any time pregnancy	2681 (80.01)	78.62, 81.32	1580 (72.25)	70.32, 74.08	4261 (76.94)	75.81, 78.03
Adequate social support	659 (19.67)	18.35, 21.05	520 (23.78)	22.04, 25.60	1179 (21.29)	20.23, 22.39
Inadequate social support	11 (0.33)	0.18, 0.59	87 (3.98)	3.23, 4.88	98 (1.77)	1.45, 2.15
Missing						
Gestational age at delivery, n (m ± sd)	2994 (39,056 ± 1.89	6.08, 7.80	2128 (39.28 ± 1.92)	5.48, 7.50	5122 (39.14 ± 1.91)	6.07, 7.39
Preterm birth (<37 weeks)	231 (6.89)	81.13, 83.70	140 (6.40)	89.62, 92.04	371 (6.70)	84.84, 86.68
No preterm birth	2763 (82.45)	9.65, 11.74	1988 (90.90)	2.09, 3.47	4751 (85.79)	6.84, 8.23
Missing	357 (10.65)		59 (2.70)		416 (7.51)	

Note: denominator is same for each variable as the missing value was included in the total sample.
[a]Total cell count is higher than denominator due to longitudinal nature of measurement.