Feasibility of Combining Common Data Elements Across Studies to Test a Hypothesis

Elizabeth Corwin, Emory University
Shirley M. Moore, Case Western Reserve University
Andrea Plotsky, Emory University
Margaret M. Heitkemper, University of Washington
Susan G. Dorsey, University of Maryland
Drenna Waldrop-Valverde, Emory University
Donald E. Bailey, Jr., Duke University
Sharron L. Docherty, Duke University
Joanne D. Whitney, University of Washington
Carol M. Musil, Case Western Reserve University

Only first 10 authors above; see publication for full author list.

Journal Title: Journal of Nursing Scholarship
Volume: Volume 49, Number 3
Publisher: Wiley: 12 months | 2017-05-01, Pages 249-258
Type of Work: Article | Post-print: After Peer Review
Publisher DOI: 10.1111/jnu.12287
Permanent URL: https://pid.emory.edu/ark:/25593/s9ht8

Final published version: http://dx.doi.org/10.1111/jnu.12287

Copyright information:
© 2017 Sigma Theta Tau International

Accessed December 13, 2019 5:33 AM EST
Feasibility of Combining Common Data Elements Across Studies to Test a Hypothesis

Elizabeth J. Corwin, PhD, RN, FAAN¹, Shirley M. Moore, PhD, RN, FAAN², Andrea Plotsky, MPH³, Margaret M. Heitkemper, PhD, RN, FAAN⁴, Susan G. Dorsey, PhD, RN, FAAN⁵, Drenna Waldrop-Valverde, PhD⁶, Donald E. Bailey Jr., PhD, RN, FAAN⁷, Sharron L. Docherty, PhD, RN, FAAN⁸, Joanne D. Whitney, PhD, RN, FAAN⁹, Carol M. Musil, PhD, RN, FAAN¹⁰, Cynthia M. Dougherty, PhD, APRN, FAAN¹¹, Donna J. McCloskey, PhD, RN, FAAN¹², Joan K. Austin, PhD, RN, FAAN¹³, and Patricia A. Grady, PhD, RN, FAAN¹⁴

¹Alpha Epsilon, Associate Dean for Research and Professor, Emory University Nell Hodgson Woodruff School of Nursing, Atlanta, GA, USA
²Alpha Mu, Professor, Case Western Reserve University, Bolton School of Nursing, Cleveland, OH, USA
³Research Informatics Analyst, Emory University Nell Hodgson Woodruff School of Nursing, Atlanta, GA, USA
⁴Psi-at-Large, Professor, University of Washington School of Nursing, Seattle, WA, USA
⁵Pi, Professor, University of Maryland School of Nursing, Baltimore, MD, USA
⁶Associate Professor, Emory University Nell Hodgson Woodruff School of Nursing, Atlanta, GA, USA
⁷Beta Epsilon, Associate Professor, Duke University School of Nursing, Durham, NC, USA
⁸Beta Epsilon, Associate Professor, Duke University School of Nursing, Durham, NC, USA
⁹Psi-at-Large, Professor, University of Washington School of Nursing, Seattle, WA, USA
¹⁰Alpha Mu, Professor, Case Western Reserve University, Bolton School of Nursing, Cleveland, OH, USA
¹¹Psi-at-Large, Professor, University of Washington School of Nursing, Seattle, WA, USA
¹²Clinical Advisor/Contractor, National Institutes of Nursing Research, Washington, DC, USA
¹³Alpha, Consultant, National Institutes of Nursing Research, Washington, DC, USA
¹⁴Tau, Director, National Institutes of Nursing Research, Washington, DC, USA

Abstract

Purpose—The purpose of this article is to describe the outcomes of a collaborative initiative to share data across five schools of nursing in order to evaluate the feasibility of collecting common
data elements (CDEs) and developing a common data repository to test hypotheses of interest to nursing scientists. This initiative extended work already completed by the National Institute of Nursing Research CDE Working Group that successfully identified CDEs related to symptoms and self-management, with the goal of supporting more complex, reproducible, and patient-focused research.

**Design**—Two exemplars describing the group’s efforts are presented. The first highlights a pilot study wherein data sets from various studies by the represented schools were collected retrospectively, and merging of the CDEs was attempted. The second exemplar describes the methods and results of an initiative at one school that utilized a prospective design for the collection and merging of CDEs.

**Methods**—Methods for identifying a common symptom to be studied across schools and for collecting the data dictionaries for the related data elements are presented for the first exemplar. The processes for defining and comparing the concepts and acceptable values, and for evaluating the potential to combine and compare the data elements are also described. Presented next are the steps undertaken in the second exemplar to prospectively identify CDEs and establish the data dictionaries. Methods for common measurement and analysis strategies are included.

**Findings**—Findings from the first exemplar indicated that without plans in place a priori to ensure the ability to combine and compare data from disparate sources, doing so retrospectively may not be possible, and as a result hypothesis testing across studies may be prohibited. Findings from the second exemplar, however, indicated that a plan developed prospectively to combine and compare data sets is feasible and conducive to merged hypothesis testing.

**Conclusions**—Although challenges exist in combining CDEs across studies into a common data repository, a prospective, well-designed protocol for identifying, coding, and comparing CDEs is feasible and supports the development of a common data repository and the testing of important hypotheses to advance nursing science.

**Clinical Relevance**—Incorporating CDEs across studies will increase sample size and improve data validity, reliability, transparency, and reproducibility, all of which will increase the scientific rigor of the study and the likelihood of impacting clinical practice and patient care.

The discovery and dissemination of knowledge that improves the health of individuals, families, and populations around the world is a fundamental goal of nursing science. While the work of individual researchers is essential to achieving this goal, individual efforts are often limited by sample size, participant characteristics, generalizability, lack of replication samples, and sometimes access to cutting-edge technology or appropriate tools and measures. Thus, a key strategy to accelerate the advancement in nursing science lies in expanding data-sharing methods and methodologies among research teams both within and across institutions.

Some large-scale data-sharing collaborations are well established and functioning, for example, those run through the National Cancer Institute and the Human Genome and Microbiome Initiatives, while other collaborations are in progress. Over the past 5 years, the National Institute of Nursing Research (NINR) has been spearheading efforts to advance nursing science by guiding researchers in the collection of data in ways that will allow
sharing with other investigators, including across universities and practice settings. A core strategy of the NINR initiative has been their work towards defining sets of common data elements (CDEs).

**Common Data Element**

A CDE is a combination of a defined variable paired with a specified set of similarly coded responses to questions that are common to multiple data sets or used across different studies. CDEs are used in research where measurement, reproducibility, and comparison across studies is important. They can be structured as a single data element, or may be included in a collection of data such as a survey scale. Use of CDEs can facilitate data sharing and standardization to improve data quality and enable data integration from multiple studies (Sheehan et al., 2016). To develop CDE standards specifically within the symptom and self-management research community, in 2013 the NINR convened the NINR CDE Working Group. This group remains active and includes representatives from NINR extramural program staff and NINR-supported P20 and P30 center directors. The defined goals of the NINR CDE Working Group are to increase the efficiency and effectiveness of research studies, increase data quality, and facilitate data sharing.

The NINR CDE Working Group began by identifying data elements that were common across P20 and P30 research centers focused on symptom and self-management science. Through a consensus process, the CDE Working Group reviewed, agreed on, and developed a set of CDEs to be commonly collected in all symptom studies—sleep, fatigue, pain, and cognitive or affect—regardless of diagnosis (Redeker et al., 2015), and to evaluate self-management of these symptoms (Moore et al., 2016).

Recommendations regarding the selection of CDEs for use in nursing research protocols involving biological samples are in progress, as is the development of a common data repository for nursing research, the latter also under the guidance of the NINR. Ultimately, by incorporating CDEs into research protocols, and depositing data generated in an accessible, common data repository, nursing researchers and others will have the opportunity to address more complex questions and advance the field in a way not possible when individuals work alone or in silos (Cohen, Thompson, Yates, Zimmerman, & Pullen, 2015; Corwin et al., 2014).

**Purpose**

The purpose of this article is to describe the efforts of a group of nursing scientists from five schools of nursing across the country (Case Western Reserve University [CWRU], Duke University, Emory University, University of Maryland, and University of Washington) working collaboratively to advance nursing science and bring momentum towards advancing methods for data sharing. Over the past 2 years, our group engaged in efforts to test the feasibility of data sharing with the goal of developing a common data repository directed towards hypothesis testing in an area of mutual interest.

We first present as an exemplar a pilot study in which retrospective data from previously completed studies at each school were collated. We contrast this with a second exemplar of
CDE use in a center grant located in one of our five schools of nursing. The successes and challenges we encountered are described, as are recommendations for future initiatives.

**Methods and Results**

**Exemplar 1: A Retrospective Approach to Building a Common Data Repository**

The first exemplar began by identifying a completed study at each institution that included at least one of the symptoms supported by the NINR in its program announcements: sleep, fatigue, affect/cognition, or pain. The studies selected for inclusion in Exemplar 1 were not randomly selected, but rather, a consensus was reached to focus on understanding the characteristics and distributions of depression, the most common symptom included in most studies (i.e., n = 6) at the five institutions. Each site was asked to submit the matching data dictionary and one or more published papers from that study, and provide an example of the raw data. This first stage was conducted with the intent to compare the depression instruments used and the demographic measures collected. Once we determined these elements to be comparable, we planned to move forward to gather the details of the studies (e.g., design and procedures).

This approach allowed us to simulate the steps one would follow if placing data into a shared data repository so that a hypothesis, or hypotheses, could be tested. We began by examining the data dictionary for each study. The data dictionaries contained each variable (or concept) in the study, and the acceptable values that were allowed for each of the variables. For example, the “sex” concept may have had male or female as acceptable values, coding as M or F, respectively. In another study, the “sex” concept may have had values of male or female, but coding as an integer (e.g., male = 1; female = 2). After reviewing the demographic variables from the data dictionaries of each study, key variables were chosen to be included in a merged data set. These included age, gender (defined as identity), sex (biologically defined), education level, race, ethnicity, marital status, employment status, income, and health insurance status.

The next critical and complex step in our work involved determining if the values for each demographic variable were amenable to harmonization. Data harmonization is the process of combining variables from disparate sources, using an equivalent coding scheme such that they can be combined and compared. We analyzed the degree to which our data elements aligned with each other, as well as how well they aligned with the CDEs suggested by the NINR or other standardized vocabularies. The National Institutes of Health (NIH) common data element resource portal was used as the authority for determining candidate concept alignments (https://www.nlm.nih.gov/cde/). Where available, NINR concepts and values were considered the gold standard for comparison. If the NINR did not have a set of values for a particular concept (e.g., gender), then the CDEs from another standard source recognized by the NIH as having a major focus on demographics were used.

When comparing studies, we found that each differed on one or more of the following aspects necessary for complete harmonization: (a) the variables chosen to operationalize the study concepts, (b) their definitions, (c) the values assigned to them, or (d) the way the values were coded. For our purpose, a decision was made to ignore how variables were
coded, since these could be recoded easily using any statistical software—that is, male or 1 could always be recoded to M if necessary; female or 2 could be recoded to F, as in the example given above. However, it was noted that such inconsistencies would inevitably slow down the harmonization process and carry the possibility of introducing errors. For the CDEs under investigation, a subset of these comparisons is shown in Table 1 and discussed below.

**Variables available**—Given that each of the identified studies was designed, initiated, and in some cases completed prior to this data harmonization effort, not all studies included data on all preselected demographic variables; for example: (a) In two studies, sex was not included as a variable because all subjects were female; and (b) In one study, marital status was not collected, as the investigators were interested only in the status of partnered versus nonpartnered participants, since those who had a same-sex partner at that time could not have been married. While sex could have been imputed in the first example, a valid marital status value could not have been imputed for the second, since whether the subjects were married or in a nonmarital relationship was not known. If this was an essential variable (i.e., intended to be used for comparison or stratification when combining data), this study would have been disqualified.

**Concept definition**—The largest discrepancies in concept definitions were between race or ethnicity and sex or gender.

**Race or ethnicity**—According to most U.S. data standards, the concept “race” refers to the standard five categories for self-identification defined in 1997 by the Office of Management and Budget (OMB): White, Black or African American, Alaska Native, Asian, Native Hawaiian, or Other Pacific Islander. This standard further provides values for the ethnicity concept, defined as Hispanic or Latino or not Hispanic or Latino. With separate concepts for race and ethnicity, it is possible to define a subject as White Hispanic, Black Hispanic, White non-Hispanic, etc. In several of our studies, however, options for race included “Hispanic” as a possible value, thereby precluding defining a subject by both concepts (e.g., White Hispanic). In this case, because a participant would have had to choose either “White” or “Hispanic,” the ability to use these data in a merged data set would be impossible if race was an important study variable. Furthermore, in one study subjects were allowed to mark more than one race, while in the others, selection of more than one value was not possible—thereby eliminating the ability to combine the data in a meaningful fashion. In addition, one study named the concept “ethnicity” but included values for the “race” concept. While this issue would not totally invalidate a data set, it could complicate down the process of compiling the data, and increase the chance for error. In five of the six studies, the values for race and ethnicity were different enough from the OMB or the NINR standard to preclude them from being harmonized into one data set as either race or ethnicity.

**Sex and gender**—According to the standard set by the World Health Organization, and adopted by many U.S. data repositories, the concept of sex refers to biological and physiological characteristics of a person. Gender refers to the sociological identification of a
person. At the time of this pilot study, the NINR had not included sex as a concept in the CDEs. They have since added clarification to their definition for gender to use it for both “biological sex distinctions and/or cultural gender role distinctions” (U.S. National Library of Medicine, 2015). In the current study, these two terms were used interchangeably across the various studies, making the data difficult or impossible to harmonize, if one wanted to delineate sex from gender. None of the studies, including the NINR CDE’s, included a “transgender” option.

Age—At the time of this pilot study, the NINR had chosen date of birth (DOB; specified as month, day, year) as the CDE for age; however, the Health Insurance Portability and Accountability Act (HIPAA) precludes the sharing of DOB or dates of medical visits. Four of the six studies used the HIPAA-preferred age variable (e.g., 21, 60) while the other two had the month, date, and year of birth. However, in these two studies the date of entry into the study was not a variable; thus, an absolute age of a subject at enrollment was not known. The NINR has since changed the coding of this variable to age in years (U.S. National Library of Medicine, 2015).

Education and income—The concepts of education and income had highly variable ways in which the data were gathered. For example, in one study, the education question was phrased “What was your highest level of education?” with options including “completed high school,” “some high school,” “completed college,” and “some college.” This question may not have yielded the same result as questions used by other studies, including “What’s the highest grade completed?” or “How many years did you attend school?” The latter question was problematic as it would not differentiate if the subject skipped a grade and completed high school in 3 years, or had been held back and took additional years to complete elementary or high school. With these and other inconsistencies, we were unable to harmonize the data for these concepts.

The “income” concept was also variable in its wording. Were the subjects asked to provide annual income, monthly income, household income, or after-tax income? In none of the data dictionaries was the operationalization of the variable identified. While it may have been possible in studies of low-income subjects for the research team to assume that a stated income of $30,000 was an annual income, in studies that included a wide socioeconomic distribution of participants, making such an assumption could have been incorrect, reducing the rigor and reproducibility of the data set. Also, use of different breakpoints for income groupings precluded harmonization.

Remaining concepts—Similar to the issues described above, none of the values used for the concepts marital status, health insurance, or employment status allowed the data to be fully harmonized across studies due to ambiguities in definitions of their values. This was especially clear for marital status, where choices across studies varied from married or single, to married, divorced, partnered, widowed, etc. For example, trying to harmonize the data into a possible common denominator “married” versus “single” would have forced us as investigators to choose how to categorize the individual who had identified himself as a widower; was he married or was he single? This would have negatively impacted data integrity and scientific rigor.
Depression measures—When depression measures were compared, other harmonization issues arose. The measures used were the Center for Epidemiologic Studies Depression Scale (CES-D)-10 (Radloff, 1977), the CES-D-20 (Roberts, 1980), Edinburgh Postnatal Depression Scale (EPDS; Cox, Holden, & Sagovsky, 1987), Patient-Reported Outcomes Measurement Information System (PROMIS) Depression 8a instrument (Reeve et al., 2007), and Beck Depression Inventory-II (BDI-II; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). While the CES-D-20 and the BDI-II have been validated against the PROMIS Depression instrument, the EPDS and the CES-D-10 have not. Therefore, either those studies that used a depression tool that had not been validated against the PROMIS instrument would have been eliminated, or the studies would have needed to be compared utilizing cutoff scores for depression. This would have been doable, although it would have forced a less granular analysis of the subjects’ data, since the data point would be binary (Depressed? Yes or No) instead of indicating degree of depression. In addition, although some studies had baseline and post-treatment scores, others had only post-treatment scores, since they had been referred for the study because they were diagnosed with depression, but actual scores prior to treatment were not known. Having a field for delineating pretreatment and post-treatment scores would be necessary to make meaningful use of these data in comparisons across studies of intervention effects.

Conclusions and recommendations from Exemplar 1—In conclusion, this exemplar highlights the challenges in establishing a common data repository by merging data sets from studies already initiated or completed. Our experience made clear that harmonization plans and data repository characteristics need to be planned during study design phases well in advance of study initiation. Such an effort is presented in Exemplar 2.

Exemplar 2: A Prospective Approach to Building a Common Data Repository

The second exemplar describes the experience of developing CDEs for use across all pilot studies in a P30 center of excellence at the school of nursing at CWRU. The Self-Management Advancement through Research and Translation (SMART) Center is focused on examining the brain-behavior connections underpinning effective self-management behavior. Development of CDEs for the Center studies began in 2014 in preparation for the submission of the Center grant application. In response to the call for proposals and the requirement within the call for the incorporation of designated CDEs, a team of investigators at the school began to identify CDEs to be used across all the proposed pilot studies for the Center. In the grant application, all participants agreed to ensure common measurement and analysis strategies across not only the proposed pilot studies but to collaborate with other NINR-funded self-management centers in data sharing using those CDEs. The long-term goal was to be able to pool data across all of proposed Center studies to increase the knowledge gained about the effectiveness, mechanisms, and important contextual factors of a variety of self-management interventions targeted to different chronic illness populations.

Variables chosen—The first step in selecting CDEs involved developing a common model to be used for all of the Center pilot intervention studies. The focus of this research center was on uncovering the brain-behavior connections specific to the self-management of health and illness. Thus, the model incorporated biological, psychological, and social...
mechanisms, and contextual variables, as well as a set of proximal and distal outcomes of self-management interventions. The variables in this model became the framework for the design of the pilot studies and selection of CDEs. The Center leaders and the principal investigators of the first two pilot studies met to make decisions about the measures of the CDEs to be used. The group reviewed each proposed CDE, its conceptual definitions of constructs, and strengths and weaknesses of tools to measure each construct. Also taken into consideration were burden on subjects and study staff, level of data or dimensions captured, use across different populations, availability in different languages, and cost. As a beginning step to creating a manualized protocol for CDE use across the Center studies, a table listing each selected CDE variable, its measure, and details of its acquisition and use was developed. When the first pilot studies were started, the investigators were asked to develop a common code book (data dictionary) containing the CDEs. Although all Center pilot study investigators were expected to use the CDEs, they were also assured that they could use other measures of the same concept if they desired.

**Concept definition**—Seventeen CDEs were initially identified that were categorized using the Center model as: (a) outcomes (proximal behavioral outcomes of diet, activity, and medication adherence, and distal outcomes of quality of life and cost); (b) mediators (brain activation of specific target areas, hypothalamic-pituitary-adrenal (HPA) axis function, biological and perceived stress, self-efficacy, decision-making, self-regulation, and patient activation); and (c) moderators or confounders (social support, depression, anxiety, cognitive functioning, and demographic variables).

A year following the beginning of data collection for the initial two pilot studies, the first attempt at harmonization of data across the studies was made. The investigators found that even though one study had copied the questionnaires and codebooks from the other study, differences existed in the data sets that prohibited easy combination of the data sets. A frequent problem was the difference in the variable labels despite the use of what was thought to be a common data dictionary. Examples of these small, but important, differences in the data coding labels were a missing letter or underscore and the use of capitalization. Neither study data set included a variable for the data collection point (a restriction built into the vertical structure of the database system we used, which does not permit the use of unique labels in the database for the same variable, even when collected at different times). Some investigators had tried to manage this by using a question number as part of a variable label, but given that each study had a different number of questions, the use of question numbers in the variable name was problematic when data combining was attempted. Although it is obvious that the codes for missing data and “not applicable” responses should be identical across studies, they found that investigators handled these codes differently, and there was a need to standardize these codes as well.

The next set of revisions to the CDEs was done 2 years after the first two pilot studies began, when they were planning for the start of four new pilot studies. During training of the new study investigators on the Center CDEs, it was found that large differences existed in the way the demographic variables had been defined, coded, and labeled across the first two studies. Although they had specified the demographic variables to be obtained in each study, they did not identify the exact way to define and measure the demographic variables. Similar
to the description given in Exemplar 1, study investigators used different definitions, measures, and coding for age, gender, race, ethnicity, education, marital status, and employment. Following considerable and difficult discussions, review of the literature, and review of other national CDE sets, they developed a common approach to the demographic variables.

Also, at this time, new CDEs were added to the Center CDE list. For example, two variables were added because new neurocognitive literature suggested that subject handedness (left, right, ambidextrous) and decentering (a cognitive referential process) are important to include. We also added intervention dose (definition, measure, and coding labels). CDEs for other measures of intervention fidelity were considered but ultimately left as recommendations, rather than as required CDEs.

Challenges and Recommendations from Exemplar 2—One challenge in the attempt at data harmonization across one university’s Center studies was the existence of multiple versions of instruments. For example, an instrument can have different versions that may create confusion among investigators. Investigators found this with the measure of self-regulation, which used the Index of Self-Regulation (ISR). The ISR originally had 16 items (Fleury, 1988), but the revised version of the ISR has 9 items (Yeom, Choi, Belyea, & Fleury, 2011). There also can be different versions of a measure that address either a general population or a specific population, which challenges instrument consistency across studies. For example, the initial studies used the original 13-item Patient Activation Measure (PAM; Hibbard, Stockard, Mahoney, & Tusler, 2004) but in the second round of studies that focused on caregiver self-management, the 10-item Caregiver PAM (Sadak, Korpak, & Boorson, 2015) was more desirable.

In addition, common problems, such as mistyping questions (“I can’t …” instead of “I can …”) or reversing response options, which are easily handled in a single study, require close vigilance and coordination to ensure harmonization of data and to prevent perpetuating mistakes across studies, such as miscopying into other files or into a template. The investigators recommend that at least two people reconcile questions and codebooks by reading aloud each question (and comparing it to the published instrument), response option, variable label, and variable description in order to catch even trivial mistakes that could be problematic at the data set combination stage.

The investigators also learned that more experienced investigators had strong preferences about operationalizing variables using favored instruments and measurement strategies. Interestingly, this was particularly true for the demographic CDEs. However, once investigators understood the usefulness of data harmonization, they were generally agreeable and willing to work toward group consensus.

The group recommends the creation of a master data dictionary that includes variable names and coding of the CDEs and that it be provided to investigators with clear instructions that any changes must be discussed with the cross-study data management team. They also recommend the development of a common manual of procedures for data collection and management. However, although in the ideal world codebooks and data dictionaries are
created a priori, the real world of research is more dynamic, requiring close attention and coordination over time. Developing and using CDEs is an iterative and evolutionary process.

The CWRU Center of Excellence to build the SMART Center is using CDEs in seven pilot studies. They have learned to view the use of CDEs as a developmental area in the conduct of research. Several new needs in research infrastructure have emerged, including the need for more knowledge about building data repositories and learning about how to define the metadata that go with putting data into a data repository (i.e., type of study, study design, study context). There is also a need for a centralized data manager and statisticians with skills in analysis of pooled data. Last, they are currently addressing the need for policies about the rights of an individual investigator’s ownership of their data deposited into a repository. There is a need for a system that acknowledges that sharing data across self-management studies is rewarding. These investigators are now poised to begin analyses of the pooled data, with the goal to advance knowledge related to self-management faster and explore more directly differences among interventions, populations, and cultures.

**Discussion**

Although our group encountered challenges in our attempts at using CDEs across studies and institutions, we are of the unanimous consensus that the important advantages to including CDEs in nursing research make their use compelling. These include the practical advantage of reduced start-up time and cost if data collection tools are already available and coded, rather than needing to be developed. Less practical but perhaps more importantly, the use of CDEs will promote standardized, consistent data collection that will improve data quality and facilitate data sharing. This will lead to improved opportunities for meta-analysis and comparison of results from different studies, thereby increasing sample size, data rigor, and reproducibility, and, ultimately, the clinical impact our science will have on patient and population health.

It is essential, however, that the barriers that exist to the adoption of CDEs be addressed, including the lack of knowledge and familiarity with CDEs among many researchers, an increase in perceived burden to subjects, and concerns related to data storage and managing large data sets. The NINR and the P20 and P30 center directors are seeking to improve knowledge regarding CDE use by building greater awareness through outreach at the Council for Advancement of Nursing Science national research meeting and at the four regional nursing research society meetings. It is hoped that these outreach initiatives will promote the culture change necessary for CDE adoption and use. Moreover, researchers may have data collection protocols that do not incorporate the specific measures that have been developed into CDEs. In such situations, it may be necessary to collect data using more than one measure to ensure compatibility with legacy data, as well as to allow data from multiple laboratories to be easily aggregated as the field moves forward. Incorporating CDEs does not preclude the use of hypothesis-driven measures or the inclusion of other CDEs that would be necessary for a project. The barrier of perceived burden to subjects might be addressed by developing or choosing CDEs such as the PROMIS measures or others that are brief and easy to administer. Hesitation by researchers to manage large data sets is understandable; however, we believe this will be a necessary skill for the future as more complex questions
are posed, some of which cannot be answered using small data sets. In addition, addressing such questions will require necessary storage and technology support.

**Lessons Learned**

Given our experiences with the use of CDEs and the development of a common data repository both retrospectively and with preplanning, we identify the following for consideration and discussion:

1. Plan data collection procedures and levels of measurement that facilitate simple harmonization. As the two cases above illustrate, a priori determination of data concepts including their definition, acceptable values, and coding schema are essential; without this, the barriers to the retrospective merging of data sets across studies described in Exemplar 1 were insurmountable. In addition, these illustrative cases show that collection of data values at the most precise level of measurement possible allows for flexibility in harmonization procedures such as categorizing of values into strata. The inclusion of a research team member with expertise in data science or complex data set management will assist in this goal.

2. Set standards for data quality. Preplanning effort needs to go into the development of guidelines and data set inclusion criteria to ensure that only data sets that meet these criteria will be included in a particular harmonization effort. This involves developing valid descriptions of the purpose for which the original data sets were collected. Such efforts will assist users of the data repository as they develop their rationale for the fit of the data to their research question.

3. Commit to using available, published, and supported CDEs (e.g., NINR, National Institute of Neurological Disorders and Stroke) in order to contribute to national initiatives towards data harmonization. Professional research society-supported CDEs often are augmented by procedures for data collection and harmonization. The NIH has established a CDE resource portal (http://cde.nih.gov/) to assist investigators in identifying NIH-supported CDEs when developing grant proposals, protocols, case report forms, and other instruments for data collection. The portal also provides guidance about and access to NIH-supported CDE initiatives and other tools and resources in NIH-funded research. Investigators are encouraged to consult the portal and describe in their grant applications any use they will make of NIH-supported CDEs in their projects.

4. Continue to connect with professional nursing research societies and regional research meetings to further support data harmonization work already underway by the NINR CDE Working Group. These organizations may be able to provide infrastructure through devoted time/space and networking opportunities. Sharing resources, such as frameworks, analytic model expertise, and data collection expertise, will advance efforts in this field. Nursing societies and related meetings could organize panels or lectures on CDEs, and harmonizing data with built-in opportunity to network with colleagues about opportunities for shared data, identification of expertise, and areas where combining data would allow research questions to be considered in innovative ways.
5. Collaborate and form partnerships with academic/industry/government to facilitate use of CDEs for data harmonization. Access to available data repositories that support and encourage data sharing will accelerate high-impact collaborative science occurring across multiple sites, including international partners. Early engagement and frequent communication will help to ensure that duplication of effort and infrastructure is minimized. These partnerships may also assist in alleviating or anticipating possible legal requirements for data sharing that may delay or deter data harmonization activities.

6. If a common data repository is established, ensure careful curation whenever new data are added. It is unrealistic to assume that new data sent to the repository will be uploaded automatically. Careful monitoring by the repository data managers will be essential, and clarifying discussions will be required between the curators and study investigators.

Conclusions

Combining data sets across studies focused on a common theme, problem, and construct or measurement tool can bring greater understanding and more powerful external validity as it permits aggregation of information across samples and populations. Such data aggregation and harmonization can lead to data comparisons to detect subtler and more complex associations among variables, allow greater statistical power and more robust point estimates than those derived from an individual study, promote greater collaboration across researchers in a field, and exert greater impact to change clinical practice and improve patient outcomes. The studies selected in Exemplar 1 were from a preselected pool; future efforts would benefit from a larger collection of studies. Larger data sets will also increase the scientific rigor of the composite studies and, by having the data housed in an accessible common data repository, will increase data transparency. The present report sought to demonstrate the challenges inherent in data harmonization approaches initiated post-hoc, or after variable operationalization choices have been made and data collection has begun. In contrast, when data harmonization is preplanned, opportunities for collaboration and significant advancement to the field is increased. The NINR CDE Working Group will continue to work with NINR-funded centers and others to collect feedback from the community to identify new CDEs. In the coming years, it is anticipated that an NINR-supported data repository will further support the discovery of new knowledge and enhance impactful, patient-focused nursing science research.

References


Clinical Resources

<table>
<thead>
<tr>
<th>Gold standard</th>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 3</th>
<th>Study 4</th>
<th>Study 5</th>
<th>Study 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>African American</td>
<td>White</td>
<td>White, not Hispanic</td>
<td>African American</td>
<td>White</td>
<td>Caucasian</td>
</tr>
<tr>
<td></td>
<td>American Indian or Alaska Native</td>
<td>Black</td>
<td>African American, not Hispanic</td>
<td>American Indian/Alaska Native</td>
<td>African American</td>
<td>American Indian/Alaska Native</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>Hispanic</td>
<td>Hispanic</td>
<td>Native Hawaiian or Other</td>
<td>Native Hawaiian or Other</td>
<td>Hawaiian or Other</td>
</tr>
<tr>
<td></td>
<td>Black or African American</td>
<td>Asian/Pacific Islander</td>
<td>White, Hispanic</td>
<td>Pacific Islander</td>
<td>Other</td>
<td>Pacific Islander</td>
</tr>
<tr>
<td></td>
<td>Native Hawaiian or Other</td>
<td>American Indian/Alaskan Native</td>
<td>Hispanic, White</td>
<td>American Indian/Alaskan Native</td>
<td>Pacific Islander</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>Pacific Islander</td>
<td>Other (free text)</td>
<td>Hispanic</td>
<td>Asian</td>
<td>Native Hawaiian/other</td>
<td>Other (check 1 or more)</td>
</tr>
<tr>
<td></td>
<td>Other (free text)</td>
<td></td>
<td>Asian/Pacific Islander</td>
<td>Pacific Islander</td>
<td>Other</td>
<td>Asian/Pacific Islander</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Other (free text)</td>
<td>Did not answer</td>
<td>Other</td>
<td>Other</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Hispanic or Latino</td>
<td>[included with race]</td>
<td>[included with race, but could be separated]</td>
<td>[included with race]</td>
<td>[included with race]</td>
<td>[included with race]</td>
</tr>
<tr>
<td></td>
<td>Not Hispanic or Latino</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment status</td>
<td>[Are you currently working?]</td>
<td>[Do you work outside the home now?]</td>
<td>[What is your employment status?]</td>
<td>Retired</td>
<td>Full time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>(Select all that apply)</td>
<td>Occupation categories</td>
<td>Part time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>No</td>
<td>Employed full time for pay</td>
<td>Retired officially</td>
<td>Not employed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[If no:] On medical leave</td>
<td></td>
<td>Employed part time for pay</td>
<td>Certified disabled</td>
<td>Retired</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Retired</td>
<td></td>
<td>Full-time student</td>
<td></td>
<td>Other</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Disabled</td>
<td></td>
<td>Unemployed (not officially or certified</td>
<td></td>
<td>Disabled</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(permanent or temporary)</td>
<td></td>
<td>100% disabled by the Veterans Administration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Keeping house</td>
<td></td>
<td>Retired officially</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Student</td>
<td></td>
<td>Certified disabled</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other, specify</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression measure used</td>
<td>CES-D (10) (yes/no response; instead of scaled)</td>
<td>Edinburgh Postnatal Depression Scale</td>
<td>PROMIS Depression 8a instrument</td>
<td>CES-D (20), BDI-II</td>
<td>BDI-II</td>
<td>CES-D (20)</td>
</tr>
</tbody>
</table>

Note. BDI-II = Beck Depression Inventory-II; CES-D = Center for Epidemiologic Studies Depression Scale; PROMIS = Patient-Reported Outcomes Measurement Information System.