Published in final edited form as:

West J Nurs Res. 2019 January ; 41(1): 78–95. doi:10.1177/0193945917749481.

Creation of Data Repositories to Advance Nursing Science

Joseph Perazzo, PhD, RN^{1,*}, Margaret Rodriguez¹, Jackson Currie¹, Robert Salata, M.D.², and Allison Webel, PhD, RN¹

¹Frances Payne Bolton School of Nursing, Case Western Reserve University, 2120 Cornell Road, Cleveland, OH, 44106 USA

²Case Western Reserve School of Medicine, 2109 Adelbert Rd, Cleveland, OH 44106 USA

Abstract

Data repositories are a strategy in line with precision medicine and big data initiatives, and are an efficient way to maximize data utility and form collaborative research relationships. Nurse researchers are uniquely positioned to make a valuable contribution using this strategy. The purpose of this article is to present a review of the benefits and challenges associated with developing data repositories, and to describe the process we used to develop and maintain a data repository in HIV research. Systematic planning, data collection, synthesis, and data sharing have enabled us to conduct robust cross-sectional and longitudinal analyses with more than 200 people living with HIV. Our repository-building has also led to collaboration and training both in and out of our organization. We present a pragmatic and affordable way that nurse scientists can build and maintain a data repository, helping us continue to make to our understanding of health phenomena.

Keywords

Biological Research; Biorepository; HIV; Data

Decades of exploratory and intervention research have led to a better understanding of the health of individuals across diverse populations. Effective prevention and treatment of disease and the promotion of health require an in-depth understanding of the biological, clinical, behavioral, psychosocial, and ecological actors that shape health (National Institutes of Health, 2017). Researchers, particularly those in academic settings, often amass a large amount of valuable data, but face barriers to maximizing their use of that data (Fecher, Friesike, & Hebing, 2015). These barriers include: limitations in the scope, purpose, and generalizability of individual research projects; lack of experience or discomfort regarding data sharing; and the high costs associated with longitudinal research (Acord & Harley, 2012; Fecher et al., 2015; van Panhuis et al., 2014).

Corresponding Author: Allison Webel, PhD, RN, 2120 Cornell Road, Cleveland, OH 44106, (216) 368-3939, arw72@case.edu.

*Author is currently with the University of Cincinnati College of Nursing; 3110 Vine Street, Cincinnati, OH 45219

The authors declare no conflicts of interest

Data Repositories: A Solution for Nursing Research

Creating data repositories is a strategy to maximize data utility that is in line with the national precision medicine and big data initiatives. These initiatives aim to efficiently utilize data to better understand the personal and contextual factors impacting the health of individuals over time (National Institutes of Health, 2017). The majority of such studies have included biological “data banking” (biorepositories) in the fields of genetics and medicine (Siwek, 2015), but nurse researchers study a wide range of phenomena (American Association of Colleges of Nursing, 2006) and can make a valuable contributions through creation of, and contribution to, data repositories.

Data repositories are built through a systematic process of planning, collecting, and archiving data (Siwek, 2015). Large national and international data sets that follow this framework (e.g. Framingham Heart Study; National Heart Lung and Blood Institute and Boston University, 2017; Centers for Medicare and Medicaid Services, 2017) are repeatedly and consistently used to understand different phenomena up to decades after the initial data collection. Careful development of data repositories can benefit researchers and the populations they serve (Acord & Harley, 2012; Siwek, 2015; Tenopir et al., 2011). First, researchers can maximize their use of their data. For example, the collection and storage of tissue samples analyzed alongside observational or self-report measures (e.g. functional tests, questionnaires) can generate new knowledge in the form of new projects or secondary analyses (Siwek, 2015). Second, researchers focusing on a particular phenomenon can use consistent measures over multiple projects to enhance their investigation of a phenomenon. For example, they can compare their findings to similar or differing populations and examine changes in a phenomenon over time (Hamilton et al., 2011). This is particularly beneficial to health scientists, as it allows them to determine how variables change as a result of changes in treatment in structural conditions (Hamilton et al., 2011). Finally, data repositories present opportunities to share data and collaborate with other scientists. Data repositories allow researchers to efficiently compare findings, combine and analyze like data across various projects and institutions, conduct systematic reviews guided by research questions related to archived data (Coady et al., 2017), and present training opportunities for students and novice scientists (Fecher et al., 2015).

Despite the benefits of data archiving and repository development, researchers are reluctant to archive and share data. Their hesitation may relate to cost, concerns about participant privacy and data use and ownership concerns (Acord & Harley, 2012; Fecher et al., 2015; Tenopir et al., 2011). The cost of developing and maintaining a data repository is difficult to estimate, as some studies may require more resources than others. Common costs associated with data repositories include: personnel hired for data management, laboratory supplies, shipping, and storing where applicable, and software costs. However, some of these costs can potentially be offset by the use of free data capture systems (e.g. REDCap), and the integration of research students and postdoctoral fellows who can gain invaluable experience with data collection, management, and analysis in collaboration with principal investigators. In other cases in which data do not include overhead costs, costs are relatively minimal outside of the time required to manage the data.

Participant privacy concerns relate to the risk of re-identification and/or disclosure when data are shared between multiple researchers or institutions (Tucker et al., 2016). Concerns about disclosure have been important considerations for researchers collecting tissue for genetic studies, as scientific advancement may lead to discovery of new diseases or risks unknown at the time of informed consent (Christenhusz, Devriendt, & Dierickx, 2013). In both cases, the National Institutes of Health Office of Extramural Research (2003) and the National Science Foundation (2017) developed policies and guidelines to help researchers develop immediate and long-term plans for data sharing. These guidelines can help researchers maximize the benefits of data sharing while minimizing the corresponding risks, such as privacy breaches. Furthermore, they can also give guidance on clinically actionable disclosures when future findings warrant contacting participants.

There are several reasons why investigators may be reluctant to share their data. These concerns relate to relinquishing control of the data, which can lead to misuse, or misinterpretation of the data in subsequent publications (Acord & Harley, 2012; Fecher et al., 2015; Tenopir et al., 2011; van Panhuis et al., 2014). Furthermore, data ownership concerns (especially in publication) are prevalent and related to competing professional obligations between investigators. Acord and Harley (2012) point out that when it comes to sharing data, researchers contend with competing priorities both within the academic institution and within the scientific community. Researchers may not share their data because (a) they won't receive institutional credit toward their own career advancement (e.g. promotion, tenure), or (b) it may create competition with other scientists for publication and recognition of scientific contributions (Accord & Harley, 2012). Some grant mechanisms (e.g., federal grants) provide instructions about data ownership and data sharing policies (National Institutes of Health Office of Extramural Research, 2003). However, the related data sharing plans are often vague and ultimately left up the principal investigator of the grant. A generally accepted data sharing method involves a period of exclusive ownership by a principal investigator followed by data sharing with mutually agreed-upon terms of acknowledgement in future work (National Institutes of Health Office of Extramural Research, 2003). However, in projects that are not subject to federal mandates regarding data sharing, individual investigators and their collaborators will benefit from developing data sharing policies congruent with ethical standards of publication (Acord & Harley, 2012; Fecher et al., 2015). As repositories become ready for data sharing, it is advisable for scientists to use websites that can integrate online, bi-directional, survey-style data requests and proposals similar to those used to apply for conference presentations. Such a system will send the request to an appointed individual responsible for approval and export, and will create a standardized approach that allows for rotation and succession planning for data sharing responsibilities.

Scientists also face the challenge of collecting and maintaining data from various sources and across various sites. A major limitation addressed by data repositories is the ability to analyze data with a more representative population than a scientist (or team) is able to obtain on their own. This may involve the collection of data from multiple health systems or data archives. When accumulating data in this way, however, scientists must be vigilant of inherent differences in the way variables are defined, measured, and documented across data sources. To address this challenge, scientists can detail data abstraction efforts in their

protocols and make data dictionaries available to other scientists that provide definitions of specific terms, outline collection procedures, and define specific parameters and data collection measures for specific variables.

Another potential challenge involves data linking and matching across multiple sources (e.g. medical records, repository records). Particularly in longitudinal investigations, a single participant (of potentially thousands) may have dozens of study visits or healthcare encounters that an investigator would like to analyze. Often, there are only certain identifiers disclosed in a data set (Dusetzina et al., 2014). To manually navigate through large data sets to match data from multiple sites is overwhelming, and in some cases, completely infeasible. Statistical analysis programs with merge capability and informatics tools like linkage software (Dusetzina et al., 2014), have algorithms that allow scientists to assign linkage criteria. These software programs then do the work of finding matches and eliminating non-matches. While invaluable to consolidating participant data, it does not supplant the need for thorough data cleaning, in which transcription errors, misspellings, and erroneous data entry must be corrected. The more automated the process of data entry and abstraction is, the less likely we introduce human error into our work.

When working with multiple sites, scientists face the challenge of data collection and management from a distance, and pragmatic challenges including data analysis responsibilities, authorship, and data use agreement (Dusetzina et al., 2014; Holzemer, 2007). Scientific teams will benefit from developing a clear protocol that outlines the expected responsibilities of all sites and team members during and after a study. An example of this model is seen in the International Nursing Network for HIV/AIDS Research (Holzemer, 2007). This group is composed of nurse researchers across more than 15 sites worldwide. The team works together to develop and conduct multi-site collaborative studies. At two annual meetings, the team develops a study protocol (see Holzemer (2007) for a list of studies to date), often driven by the proposal of an individual who will be the central study principal investigator. Investigators from specific sites sign on to be a participating site depending on their availability. The team decides on overall research questions that will be reserved for cross-site analyses and publications that include all site principal investigators, while each site retains ownership of their site-specific data for other analyses. All site personnel inform the network of planned publication via email listserv to prevent inadvertent duplication of research questions and analyses reserved by the collective network and/or other sites within the network. For each study, the central principal investigator and a central study coordinator assume responsibility for overseeing central data linkage, cleaning, and management. For example, a research-secure cloud drive or electronic data capture system (e.g. InForm, REDCap, Oracle, Clinsys, DATATRAK) provide an efficient and automated way to manage data (Shah et al., 2010).

Studies leveraging data repositories have typically been in the fields of medicine, genetics, and big data science, in which medical record data and biological samples are automatically collected and stored (Siwek, 2015). These automated methods have allowed for important medical investigations. However, nurse scientists often study phenomena at the intersection of biological and social science (American Association of Colleges of Nursing, 2006). With advanced planning and rigorous implementation, nurse scientists are uniquely positioned

create data repositories and initiate interdisciplinary collaborations that will contribute to a novel understanding of health-related phenomena. In this paper we present our process for developing and maintaining a longitudinal data repository of adults living with HIV.

Purpose

The purpose of this article is to discuss the benefits of developing data repositories in academic nursing and to present the process used to develop and maintain a longitudinal data repository of people living with HIV (PLWH).

Methods

An Overview of the Data Repository

The HIV research team at the Frances Payne Bolton School of Nursing conducts exploratory and intervention research examining how to improve self-management and health promotion behaviors of people living, and aging, with HIV (Webel, 2016). Completed and ongoing protocols have focused on: health behaviors (e.g. sleep hygiene, planned exercise and lifestyle physical activity, diet, and substance use), the factors that influence these behaviors (e.g. social capital, mental health, ecological factors), and the impact of these behaviors on key health outcomes in people living with HIV (e.g. cardiovascular and immune health). These research protocols span more than five years and include data and tissue samples from more than 220 people living with HIV (Table 1), mean age 49.95($\pm$ 7.18), on antiretroviral therapy an average of 14.45 years ($\pm$ 5.79 years), with average recent CD4+ t cell count of 671.93 ($\pm$ 396.36). The creation of a data repository encompassing these research protocols has been a central component to planning, data collecting, data analysis, and data sharing within the team. Herein we will provide an overview of the process we used to develop the data repository, including the: (a) planning phase, (b) data collection phase, (c) data synthesis phase, and (c) maintenance and data sharing phase.

The Planning Phase

Dr. Allison Webel, an HIV self-management expert and Assistant Professor of Nursing at Case Western Reserve University, established the HIV Biobehavioral research office in 2011. She is the Principal Investigator or the Responsible Investigator of the completed and ongoing research protocols discussed in this article. Planning the creation of a data repository began during her post-doctoral fellowship and required several important steps: proposing the data repository; obtaining IRB approval, strategically designing each study; creating a central point of contact, recruitment and retention of eligible participants; creating an infrastructure for data collection and storage; and developing a data sharing plan.

Proposal for the Creation of the Data Repository and Obtaining IRB Approval

—Our first step in creating a data repository was the development of a proposal outlining the plan to create a data repository and to archive collected data. We approached this similar to any research study, in which we summarized the scientific importance and potential benefits of data repositories and the methods we would use to execute this idea. We incorporated a

plan into our proposal allowing us to collect data while protecting human subjects. In our application to the IRB, we also included the following procedures:

1. Our consent process includes a discussion of the types of
 - Future contact we would have with participants (e.g., notifications about new protocols or study visits outside of the current protocol),
 - Information about the data repository (e.g., where it is stored and who has access to it) and data sharing, and
 - Participant preferences related to disclosure of study data.
2. During the informed consent discussion, participants are told about the potential for future studies and provide written informed consent to be contacted for future research, beyond the current protocol. All participants have the option to request no further contact from the study team.
3. During the informed consent discussion, participants are told about the purpose of the data repository and the potential for data sharing. Only participants who provide written, informed consent for their data to be archived in the repository are included. All participants have the option to specify that they do not want their data deposited into the data repository.
4. During the informed consent discussion, participants are told about the potential to share deidentified data with other researchers (not involved in the current research study), and asked about their preferences related to disclosure of study data (e.g. findings from genetic studies using their stored blood sample). Only those providing written informed consent for data sharing would have their data shared with others. Additionally, on the consent form participants are given three options for disclosure (i.e., “do not contact me”, “contact me any time you use my blood”, and “contact me only if you find something significant”). Participant initial their preferences.
5. Participants who consent to have their data in our repository are given a repository-specific ID (not linked to any subject characteristics) that is used across protocols to ensure accurate collection of data from participants.

These steps are repeated as participants enroll in each new protocol to ensure that they are reminded of their rights as participants. While we have not experienced active study withdrawal up to the time of this publication, a crucial component to the planning phase is a clear understanding of procedures that follow study dropout or voluntary withdrawal of consent. We sought guidance from the Office for Human Research Protections on this, and have integrated their recommendations and good clinical practice (GCP) guidelines into our protocol (Gabriel & Mercado, 2011; United States Department of Health and Human Services, 2017). The following guidelines are consistent with the 2017 changes to the final Common Rule governing research with human subjects (Agency for International Development, 2017).

1. If a participant is terminated from the study: the study team will meet with the participant to explain why, obtain permission for further contact (often by phone), and remind the participant of plans for data usage.
2. If a participant voluntarily leaves the study, withdrawing their consent for participation: the study team remains compliant with GCP, HIPAA, OHRP and FDA regulations on record retention. Participant data preceding the date of withdrawal will remain part of the retained record. Data collected to that point, particularly if analyses have already been run, must be retained to protect the integrity of the data. From the withdrawal date forward, no new data are added to the record, nor will any further procedures be completed on biological specimens. No further access of the participant's medical records or other sources of protected health information (PHI) will be conducted.
3. The principal investigator of the study will respectfully attempt to contact the participant for better understanding of withdrawal, placing the participant under no obligation to answer, and will confirm continued consent for further contact. In the event that the participant no longer wants to be contacted, notation will be made in the registry and participant information will be retained for compliance purposes.

Those considering developing their own data repository should first check with their local IRB in order to see if there are local requirements.

Strategic Study Design: An important consideration when initiating studies with an intent to archive data is a selection of variable measures that are valid, have been used in other investigations with similar populations, and (if applicable) can be used to compare findings across different populations (United States National Library of Medicine, 2017). Biological variables often translate across studies easily, while behavioral or psychosocial measures may have multiple different options for measurement. This has been a limitation in behavioral research due to conflicting opinions about the “best” measure, leading to inconsistencies in measurement across studies (Redeker et al., 2015). Researchers can leverage published protocols from other scientists and validated toolkits for guidance on measures that will best fit their research questions. For example, the PhenX Toolkit (PhenX, 2017) is an online resource that provides a discrete catalog of standardized measures (e.g. demographics, diet and nutrition, anthropometric measures, substance use) that researchers can use that will allow them to combine results across studies and populations, thereby maximizing the utility of their data.

The National Institute of Nursing Research (NINR) developed an initiative for the use of common data elements (CDE)s that researchers can use that improve the efficiency of data collection, facilitate cross-study analyses and data sharing, create standards for data collection, and produce high quality data (NINR, 2015). CDEs are quickly becoming a requirement for studies funded by NINR, and include an evolving list of specific tools that can be used in a variety of populations to measure a certain phenomenon (e.g. symptoms) (United States National Library of Medicine, 2017). It is highly recommended that researchers use CDEs whenever possible (and appropriate) to maximize the use of their data

(NINR, 2015). However, researchers can also use measures or subscales from instruments that are tailored to their target population. For example, across our studies, we examined symptom experiences and psychological variables in PLWH using both common data elements (i.e., PROMIS29 [Cella et al., 2010]) and validated measures of specific symptoms (e.g., Beck Depression Inventory [Beck, Steer, & Brown, 1996], HIV-Related Fatigue Scale [Barroso & Lynn, 2002]). When possible, it is helpful to use validated and objective measurements to promote longitudinal analysis and data sharing. In our research on physical activity and fitness in people living with HIV, for example, we used a consistent protocol for strength (i.e., hand-grip dynamometry) and physical activity (i.e., seven-day accelerometry) over time and across study protocols. These measures are simple to collect and can easily be used across different studies and sites to facilitate comparison. In doing so, researchers can dynamically measure phenomena in ways that allow them to draw conclusions about their target population, while also positioning themselves to compare findings across other populations (e.g., breast cancer survivors and those newly diagnosed with cancer). Researchers will have an enhanced ability to draw broader conclusions from their data and to data share with other scientists who have used these measures. Tables 2 outlines variables examined in across our various protocols.

Recruitment and Retention of Participants: Creating a Central Point of Contact

—We recruit participants using several methods. We have used common recruitment methods such as flyers and advertisements at local HIV care and community resource venues, as well as recruitment through word-of-mouth (i.e. snowball sampling) and physician referral. Following our consenting procedure, we created an access-restricted registry that allows us to contact previous participants about participating in new research protocols. In the context of the data repository, we have ongoing contact with subjects who participated in previous protocols and can invite them to participate again, while still recruiting new subjects (i.e., those who have not previously participated in our protocols). After enrollment, most of the study visits occur in a small research office at the Frances Payne Bolton School of Nursing, which is also our site of subject contact and data storage (except for tissue samples which are stored in a secure clinical research unit freezer). Having this *consistent, central point* of contact has been critical to our ability to recruit and retain participants over time. A researcher may not have consistent office space available or their research takes place in a clinical or community setting. In this case, creating telephone and email contacts dedicated to research are inexpensive and effective ways to give participants a central point of contact for ongoing and future research. While universities may provide this service to faculty members, free services (e.g. Google Voice) are available to create a telephone contact point. Investigators should be cautious with regard to exchanging sensitive information, and only use properly protected email accounts when exchanging information about a study using this medium. For example, we never say anything about HIV in our voicemails, letters to subjects, or emails in order to protect the subject's privacy.

Creating an Infrastructure for Data Collection and Storage—Underpinning a data repository, is an effective data organization and storage system that allows researchers to conduct primary and secondary data analyses. Electronic data capture systems provide a user-friendly way to collect and store data that then be exported in ways that best fit current

and future analyses. We used Research Electronic Data Capture (REDCap ©) (Obeid et al., 2013), a dynamic, free data management system that allows for creation of multiple “events”. These events may correspond to multiple visits in a single study, but can also be used to enter data from multiple protocols. We created unique identifiers for each participant and entered all biological, clinical, behavioral, psychosocial, and ecological variables used in our protocols into the new project, allowing us to conduct longitudinal analyses on variables over time. It may be that no such data management system is provided or available. Many statistical analysis programs can communicate with Microsoft Excel®. Researchers can use Excel to create their data repositories, using rows for specific participants and columns to report variable outcome data (e.g. questionnaire scores, lab values). These programs promote organized data storage, longitudinal data analyses, and data sharing. Creating such an infrastructure enables scientists to easily abstract, export, analyze, and to share their data with others. When researchers participate in multi-site collaboration, they can use the tools we’ve described and develop protocols for site-specific and centralized data collection and storage.

Creating a Protocol for Data Sharing—A motivating reason for developing a biorepository that includes biological, clinical, behavioral, psychosocial, and ecological data is the ability to collaborate with investigators and answering important clinical research questions. This collaboration can be within one institution (e.g., trainees) or across institutions with others investigating similar phenomena. By choosing to include common data elements (United States National Library of Medicine, 2017) and obtaining specimens using standard protocols, data can be pulled across institutions to increase power to answer a research question, improve generalizability, and conduct subgroup analyses that will improve knowledge and ultimately help to provide better clinical care.

A transparent and fair data sharing protocol should be developed early in the process. When developing ours, we examined the data sharing protocols of other related data repositories in HIV including the CFAR Network of Integrated Clinical Symptoms (<https://www.uab.edu/cnics/>) and Women’s Interagency HIV Study (<https://statepi.jhsph.edu/wihs/wordpress/history/>) cohorts. The process generally includes an interested investigator reviewing the available variables (either on a central website or by personal communication to the project director) and determining what questions have already been answered by that dataset. If the question the proposing researcher wants to answer is novel, he or she is usually asked to make a formal written proposal to the executive committee which is submitted using REDCap. This concise proposal includes the research question, background/justification, hypotheses, data analytic methods, anticipated outcomes, and the specific data requested. A verbal presentation to senior investigators may be required and further refinement of the proposal may be requested. At this stage, the executive committee is charged with ensuring there is no duplication with others using the data, that the research question can be answered with the available data, that the methods and statistical approach are valid and appropriate, and if approved, that the requesting investigator receives the de-identified data in a timely manner. After approval, the executive committee (or a delegated member) should follow up with investigator to ensure they are making progress in their specified timeline, answer any questions about the primary data that the investigator may have, and when the outcome/

manuscript is complete, help to guide that person through the publication committee process. This person may also serve a limited mentoring role, and in the case of students, may serve on his or her dissertation committee.

We were mindful of common concerns related to authorship and data ownership when designing our data sharing protocol. The principal investigator creates a list of planned analyses and publications related to primary and secondary research aims of their protocol. Then, in collaboration with co-investigators, retains decision-power regarding subsequent publications proposed by new collaborators. Authorship order and decisions regarding publication (e.g. choice of journal, delegation of writing) are made in writing prior to manuscript preparation. Similar to our above-described process of reviewing proposed analyses and data usage, all manuscripts created using these data are reviewed by the principal investigator, study team members, and associated collaborators prior to submission to keep the team informed of the registry output. Approval is not required but the authors are required to acknowledge the primary study in the resulting publication.

A final consideration for data sharing, particularly for those repositories containing genetic or proteomic data, is whether some or all of the data should be deposited into a subject-focused, public data archive. Many of these archives exist and a good list of can be found at https://professional.heart.org/professional/ResearchPrograms/UCM_461443_AHA-Approved-Data-Repositories.jsp. Increasingly, funding bodies ask grant recipients to consider depositing relevant data into these archives.

Discussion

In this paper we have shared the process we used to create a dynamic data repository consisting of a multitude of biological, clinical, behavioral, psychosocial, and ecological variables. Our continued success is the result of planning and designing the repository based on the current state of the science, obtaining IRB approval, strategically designing studies, integrating key variables into our protocols, diligent efforts to recruit and retain participants, and creating an infrastructure for long term data collection and storage. We have also provided guidance on developing a protocol for data sharing.

This ongoing project aligns with national research priorities such as the precision medicine initiative and big data science (National Institutes of Health, 2017). These initiatives have been established to optimize the use of data collected by scientists across disciplines and research settings (National Institutes of Health, 2017; Siwek, 2015). In particular, nursing science has the potential to inform big data science through the use comprehensive theoretical frameworks and novel approaches to and understanding of health measures in the acute, chronic, and community health settings (Brennan & Bakken, 2015). In time, these efforts may promote a deeper and more complete understanding of health phenomena, and help us to develop interventions and treatments to best serve a wide variety of populations. It is not uncommon for researchers to have data from previous projects that do little more than “collect dust” as they move on to new projects (Acord & Harley, 2012; Fecher et al., 2015). Some of the reasons this occurs may include: simple lack of awareness of the possibilities of data archiving, sharing and collaboration, fear of relinquishing intellectual ownership,

perceiving data from negative outcomes (in a primary study) as useless or irrelevant, ethical concerns regarding data overuse (e.g. data splitting), or concerns related to publication (Acord & Harley, 2012; Fecher et al., 2015). This can result in a body of work with great, but unrealized potential. We contend, however, that as the breadth of knowledge in a given scientific area increases, data can take on new meaning and be analyzed in novel and useful ways. Furthermore, researchers may find that other investigators uncover meaningful results from data that a primary researcher had not considered. We see this often in the analyses of large national data sets that have resulted in many publications, each contributing to the current state of the science (Centers for Medicare and Medicaid Services, 2017; National Heart Lung and Blood Institute and Boston University, 2017). With careful planning and implementation, researchers across disciplines and settings can create data repositories that address major limitations to current knowledge.

Our early planning and ongoing research have allowed us to better understand the benefits and potential pitfalls associated with data repositories. For example, the specific steps we took to design the data repository reflect our efforts to adhere to contemporary ethical standards and place the participants' preferences at the forefront of each protocol. Our recruitment and retention methods have allowed us to accurately and consistently collect data from individuals from our target population and has resulted in strong and trusting relationships between our research team and those participants. Our infrastructure has made it possible for us to combine and analyze data in novel ways and to position ourselves to collaborate with other researchers.

Future directions for research include greater integration of technology into the research process. Today, technology (e.g. wearable devices, smartphone applications) allow for collection of data in real time, eliminating potential error due to retrospect and hindsight bias. For example, symptom tracking and diet/physical activity recall applications, actigraphy, microelectronic monitors for medication event monitoring (MEMS) are options for understanding phenomena like symptom trajectories and burden, physical activity and dietary habits, and treatment adherence (El Alili, Vrijens, Demonceau, Evers, & Hiligsmann, 2016; Granado-Font et al., 2015; Kelly et al., 2013). A major challenge that will accompany these efforts will be incorporating ways to promote their use among populations traditionally difficult to reach with technology (e.g. older adults, low income individuals) (Pew Research Center, 2013; Schnall, Cho, & Webel, 2017; Zickuhr & Madden, 2012). However, in addressing these challenges and integrating these resources into our research we create opportunities to better serve our participants by enhancing our understanding of their health. Another potential area for future research is the sharing of qualitative data, including transcribed interviews and focus groups, as well as photographs, and observational field notes. Researchers can carefully document the processes they used in collecting and analyzing qualitative data, thereby creating the possibility of qualitative data sharing. Raw qualitative data (that include original interview questions and contextual notes) can provide rich secondary investigation opportunities to obtain detailed descriptive data on many phenomena.

While our work is dedicated to enhancing the health and quality of life in people living with HIV, we believe the methods we have described can be used with many different, vulnerable

populations. Nurse scientists, regardless of institution, position, or area of interest, have a great deal to contribute through their research. We encourage nurse scientists across clinical and research settings to maximize the utility of their own data. This work will enhance our understanding of health-related phenomena across settings, ultimately leading to improved health of populations and advancing scientific knowledge.

Acknowledgments

Funding: This work was supported by:

National Institutes of Health [KL2RR02499, P30NR015326, 5T32NR014213-03]; American Heart Association [14CRP20380259]; and the American Nurses Foundation [2016-ANF-NRG].

References

- Acord SK, Harley D. 2012; Credit, time, and personality: The human challenges to sharing scholarly work using Web 2.0. *New Media & Society*. 15:379–397. DOI: 10.1177/1461444812465140
- Agency for International Development. Federal policy for the protection of human subjects. 2017. Retrieved from <https://www.federalregister.gov/documents/2017/01/19/2017-01058/federal-policy-for-the-protection-of-human-subjects>
- American Association of Colleges of Nursing. AACN position statement: Nursing research. 2006. Retrieved from <http://www.aacn.nche.edu/publications/position/nursing-research>
- Barroso J, Lynn MR. 2002; Psychometric properties of the HIV-related fatigue scale. *Journal of the Association of Nurses in AIDS Care*. 13(1):66–75. [PubMed: 11828861]
- Beck, AT, Steer, RA, Brown, GK. Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation; 1996.
- Brennan PF, Bakken S. 2015; Nursing needs big data and big data needs nursing. *Journal of Nursing Scholarship*. 47(5):477–484. [PubMed: 26287646]
- Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, ... Hays RD. 2010; Initial item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS) network: 2005–2008. *Journal of Clinical Epidemiology*. 63(11):1179–1194. [PubMed: 20685078]
- Centers for Medicare and Medicaid Services. Explore all datasets. 2017. Retrieved from <https://data.medicare.gov/data>
- Christenhusz GM, Devriendt K, Dierickx K. 2013; To tell or not to tell? A systematic review of ethical reflections on incidental findings arising in genetics contexts. *European Journal of Human Genetics*. 21(3):248–255. [PubMed: 22739341]
- Coady SA, Mensah GA, Wagner EL, Goldfarb ME, Hitchcock DM, Giffen CA. 2017; Use of the National Heart, Lung, and Blood Institute Data Repository. *New England Journal of Medicine*. 376:1849–1858. [PubMed: 28402243]
- Dusetzina, SB, Tyree, S, Meyer, A-M, Meyer, A, Green, L, Carpenter, WR. Linking data for health services research: A framework and instructional guide. Rockville, MD: Agency for Healthcare Research and Quality; 2014.
- El Alili M, Vrijens B, Demonceau J, Evers SM, Hiligsmann M. 2016; A scoping review of studies comparing the medication event monitoring system (MEMS) with alternative methods for measuring medication adherence. *British Journal of Clinical Pharmacology*. 82(1):268–279. [PubMed: 27005306]
- Fecher B, Friesike S, Hebing M. 2015; What drives academic data sharing? *PloS one*. 10(2):e0118053. [PubMed: 25714752]
- Gabriel AP, Mercado CP. 2011; Data retention after a patient withdraws consent in clinical trials. *Open Access Journal of Clinical Trials*. 3:15. [PubMed: 23935390]
- Granado-Font E, Flores-Mateo G, Sorlí-Aguilar M, Montaña-Carreras X, Ferre-Grau C, Barrera-Uriarte ML, ... Satué-Gracia E-M. 2015; Effectiveness of a Smartphone application and wearable

device for weight loss in overweight or obese primary care patients: Protocol for a randomised controlled trial. *BMC Public Health*. 15(1):531. [PubMed: 26041131]

Hamilton CM, Strader LC, Pratt JG, Maiese D, Hendershot T, Kwok RK, ... Pan H. 2011; The PhenX Toolkit: Get the most from your measures. *American Journal of Epidemiology*. 174(3):253–260. [PubMed: 21749974]

Holzemer W. 2007; University of California, San Francisco International Nursing Network for HIV/AIDS research. *International Nursing Review*. 54(3):234–242. [PubMed: 17685906]

Kelly LA, McMillan DG, Anderson A, Fippinger M, Fillerup G, Rider J. 2013; Validity of actigraphs uniaxial and triaxial accelerometers for assessment of physical activity in adults in laboratory conditions. *BMC Medical Physics*. 13(1):5. [PubMed: 24279826]

National Heart Lung and Blood Institute and Boston University. Framingham heart study: A project of the National Heart, Lung, and Blood Institute and Boston University. 2017. Retrieved from <http://www.framinghamheartstudy.org/>

National Institute of Nursing Research. Common Data Element (CDE) resource portal. 2015. Retrieved from <https://www.nlm.nih.gov/cde/>

National Institutes of Health. All of us research program. Research and training: Precision medicine initiative. 2017. Retrieved from <https://www.nih.gov/research-training/allofus-research-program>

National Institutes of Health Office of Extramural Research. NIH data sharing policy. 2003. Retrieved from https://grants.nih.gov/grants/policy/data_sharing/

National Science Foundation. Dissemination and sharing of research results. 2017. Retrieved from <https://www.nsf.gov/bfa/dias/policy/dmp.jsp>

Obeid JS, McGraw CA, Minor BL, Conde JG, Pawluk R, Lin M, ... Taylor R. 2013; Procurement of shared data instruments for Research Electronic Data Capture (REDCap). *Journal of Biomedical Informatics*. 46(2):259–265. [PubMed: 23149159]

Pew Research Center. Health topics. 2013. Retrieved from <http://www.pewinternet.org/fact-sheets/health-fact-sheet/>

PhenX. PhenX toolkit. 2017. Retrieved from <https://www.phenxtoolkit.org/index.php>

Redeker NS, Anderson R, Bakken S, Corwin E, Docherty S, Dorsey SG, ... Pullen C. 2015; Advancing symptom science through use of common data elements. *Journal of Nursing Scholarship*. 47(5):379–388. [PubMed: 26250061]

Schnall R, Cho H, Webel A. 2017; Predictors of willingness to use a smartphone for research in underserved persons living with HIV. *International Journal of Medical Informatics*. 99:53–59. [PubMed: 28118922]

Shah J, Rajgor D, Pradhan S, McCready M, Zaveri A, Pietrobon R. 2010; Electronic data capture for registries and clinical trials in orthopaedic surgery: Open source versus commercial systems. *Clinical Orthopaedics and Related Research*. 468(10):2664–2671. [PubMed: 20635174]

Siwek M. 2015; An overview of biorepositories: Past, present, and future. *Military Medicine*. 180(10S):57–66. [PubMed: 26444893]

Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, ... Frame M. 2011; Data sharing by scientists: Practices and perceptions. *PloS One*. 6(6):e21101. [PubMed: 21738610]

Tucker K, Branson J, Dilleen M, Hollis S, Loughlin P, Nixon MJ, Williams Z. 2016; Protecting patient privacy when sharing patient-level data from clinical trials. *BMC Medical Research Methodology*. 16(1):77. [PubMed: 27410040]

United States Department of Health and Human Services. Office of Human Research Protections: Final revisions to the common rule. 2017. Retrieved from <https://www.hhs.gov/ohrp/regulations-and-policy/regulations/finalized-revisions-common-rule/index.html>

United States National Library of Medicine. Common data element resource portal. 2017. Retrieved from <https://www.nlm.nih.gov/cde/>

van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, ... Burke DS. 2014; A systematic review of barriers to data sharing in public health. *BMC Public Health*. 14(1):1144. [PubMed: 25377061]

Webel, AR. The Webel research lab. 2016. Retrieved from <https://nursing.case.edu/research/labs-studies/webel-lab/>

Zickuhr, K; Madden, M. Older adults and internet use. Pew Internet & American Life Project. 2012.
Retrieved from <http://pewinternet.org/Reports/2012/Older-adults-and-internet-use.aspx>

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

FPB HIV Research Office Data Repository Variables

Domain	Variables
Biological	HIV-specific biomarkers (CD4+, HIV-Viral RNA); Inflammatory biomarkers (e.g., IL-6, hsCRP), cardiovascular biomarkers (lipid panel, insulin resistance, Coronary Calcium Scoring, waist-hip-ratio, BMI), maximal cardiopulmonary exercise tests, functional MRI, Endopat
Imaging	Coronary Computed Tomography Scan, fMRI of the brain, Diffusion Tensor Imaging of the Brain
Psychological/Neurological	Depression, Cognitive Function (gross), Attention, Cardiovascular Disease Risk Perception, Decision-making and Risk, Response Time, Self-efficacy
Behavioral	Physical Activity (actigraph, activity recall), Sleep (actigraph and PROMIS), Medication Adherence, Substance Use (tobacco, alcohol, and illicit drug use), Self-Management
Symptoms	Depression, Anxiety, Fatigue
Functional	Six Minute Walk Test, Hand-Grip Dynamometry, Flexibility
Psychosocial	Social Capital, Stigma
Ecological	Weather, Neighborhood Environment, Food Security

Table 2

Frances Payne Bolton School of Nursing HIV Research Office Data Repository Participant Characteristics

Variables	N (%)
Gender	
Male	102 (58.6%)
Female	67 (38.5%)
Transgender	5 (2.9%)
Monthly Income	
No Income	25 (14.4%)
Less than \$200	10 (5.7%)
\$200–399	8 (4.6%)
\$400–599	5 (2.9%)
\$600–799	65 (37.4%)
\$800–999	21 (21.1%)
\$1000+	39 (22.4%)
Educ.	
11 th Grade or Less	45 (25.9%)
High School or GED	44 (25.3%)
Some College or Technical School	48 (27.6%)
2 Years of College	12 (6.9%)
Bachelor's Degree	20 (11.5%)
Masters Degree	5 (2.9%)
Housing	
Permanent Housing	156 (89.7%)
No Permanent Housing	18 (10.3%)
HIV **	
History of Advanced HIV (AIDS)	70 (45.5%)
Undetectable Viral Load	121 (78.6%)

* Frances Payne Bolton School of Nursing HIV Research Office

** (N) Reflects: (a) Permission to access medical record for data abstraction; and (b) availability of data for medical record abstraction