Answering Research Questions Using an Existing Data Set

Authors:

Daniel M. Doolan RN, PhD

Jennifer Winters RN, PhD

Sahar Nouredini RN, PhD

Affiliation:

California State University East Bay; CA, USA

Authors' e-mail addresses:

daniel.doolan@csueastbay.edu

jennifer.winters@csueastbay.edu

sahar.nouredini@csueastbay.edu

Correspondence author:

Daniel M. Doolan RN, PhD

daniel.doolan@csueastbay.edu

.

Abstract:

It is often advisable for researchers to use an existing data set to answer research questions. In particular, using an existing data set can help a researcher obtain results much more quickly, at a lower cost, and without exposing new research subjects to many of the potential harms associated with research participation. However, the many researchers seeking to use an existing data set face a variety of challenges specific to this research methodology. This article reviews some of the key differences associated with using an existing data set as compared with those conducting research by recruiting research subjects. Advantages and disadvantages associated with the use of existing data sets are discussed as are ethical issues, strategies to obtain an optimal data set, and special considerations associated with this methodology. Additionally, suggestions are given relevant to reporting results when conducting research using an existing data set or a "secondary analysis".

Key Words:
Secondary Analysis
Data Sets
Research Methodology
Research Techniques

1.1 Existing Data

Technological breakthroughs, especially over the last 20 years, have made sharing and analyzing data sets much easier. Increasingly, research studies are being conducted using existing data sets. Research methods publications have been slow to highlight challenges and techniques needed to conduct research using existing data sets. In this article, the authors review and update previous work done on this topic and incorporate lessons learned from research conducted using existing data.[1] [2]

1.2. Terminology/Definitions

For this publication, "data set" will be defined as any set of existing data that could be used to answer important new research questions and/or provide further evidence relevant to ongoing research questions.[3] Hence, a data set might be subject information from a previous research study, or existing data from other sources. Other sources include, but are not limited to, hospital charts, academic course records, quality improvement records, news media or social media. If a data set originated from a previously orchestrated research study, the original research study will be referred to as the parent study.[1] "Primary Investigator" will refer to the individual conducting research using an existing data set, unless otherwise specified. Research using an existing data set will sometimes be referred to as a secondary analysis.

1.3. Major Methodological Differences in Approach

Despite the common practice of researchers answering new research questions using existing data sets, there remains a scarcity of publications to provide guidance to those using existing data sets.[4] Resources for scientists conducting research are usually written with the underlying assumption that the researcher designs the study after identifying research questions.[1] This is not the case for those using an existing data set. Conducting research using an existing data set still requires a familiarity with concepts relevant to recruiting subjects and gathering data for research, but it also requires a variety of other skills specific to the challenges of conducting research with existing data.[1]

TABLE I: Checklist for Secondary Analysis as Compared With Designing a Prospective Study

Iterative Process During Secondary Analysis	**Iterative Process During a Prospective Study**
1. Perform literature review	1. Perform literature review
2. Find gaps in the research and find research opportunities	2. Find gaps in the research and find research opportunities
3. Identify and obtain permission from the original PI to analyze a data set	3. Pose research questions that could be answered given a prudent sample measures, and, if applicable, follow-up
4. Refine research questions	4. Write a research proposal including how subjects will be recruited, what data will be collected. and what safeguards will be in place to protect subjects' safety
5. Evaluate the appropriateness of the original sample, design. and measures[a]	5. Obtain approval from the organizations IRB or equivalent and committee
6. Establish appropriate safeguards to protect data and consider legal and ethical implications of the analysis	6. Recruit consenting subjects
7. Obtain approval from the organization 's IRB or equivalent committee	7. Obtain predetermined measures from subjects
8. Perform secondary analysis of the data	8. Perform the analysis of the data
9. Disseminate findings to the research community	9. Disseminate findings to the research community

a: Existing publications and discussions with the parent study's PI are useful here since access to the raw data will not occur until Institutional Review Board (IRB) approval is granted (step 7 of secondary analysis section). This table reproduced with permission of Springer Publishing Company via Copyright Clearance Center. [1]

Parent study primary investigators (PIs) have tasks that differ from investigators using an existing data set. Parent study PIs (Table I) have the ability to prospectively determine study variables, subject recruitment strategies, and other research protocols. Recruitment of study subjects might be able to be extended until the number meets or exceeds those called for in the power analysis. When using an existing data set, a PI is unable to prospectively engage in these activities and is, thus, faced with different challenges relevant to working with a data set. [1] One of the most important challenges faced by a PI using an existing data set is associated with selecting a data set that is a good fit for the research question(s). A subsequent section further discusses data set selection strategies.

1.4 Advantages and Disadvantages of Using an Existing Data Set

Good science is about finding credible answers to important research questions. Many research questions can be answered well either by using an existing data set or by engaging in a prospective research study. It has been erroneously implied that research using an existing data set is inherently more suspect than other research.[5] As a general guideline, it is highly preferable to answer research questions using an existing data set if one is available to answer the research questions. This is preferable for several reasons.

First, use of an existing data set tends to save a great deal of time, as opposed to designing a new research protocol to recruit subjects and gather data. This benefit of obtaining results more quickly is especially true for longitudinal studies, in which subjects need not only be recruited, but also followed over months or years. Second, it tends to be much less expensive to use data that has already been collected. Third, as most prospective research studies include at least some level of risk to the subjects, a researcher using an existing data set has the benefit of not putting subjects in harm's way, as those subjects have already experienced any burdens associated with participating in the research.[3] Table II contains several successful examples of researchers using an existing data set to answer new research questions.[6] [7] [8] [9] [10] [11] [12] [13]

Table II: Examples of Secondary Analysis Studies

Investigator/Name of Parent Study	Design/Year/ N	Sample	Research Question	Intervention	Key Items Assessed	Results
Example 1 Parent Study Saitz et al./Screening and brief intervention for drug use in primary care: the ASPIRE randomized clinical trial.	RCT/2014/528	Patients who screened positive for illicit drug use at an urban primary care clinic in Boston, Massachusetts, USA	To test the efficacy of 2 brief counseling interventions for unhealthy drug use (any illicit drug use or prescription drug misuse)	A brief negotiated interview (BNI) and an adaptation of motivational interviewing (MOTIV)- compared with no brief intervention	Main drugs used and severity of drug use (Alcohol, Smoking, and Substance Involvement Screening Test[ASSIST]) at baseline and 6 months	Brief interventions did not have efficacy for decreasing unhealthy drug use in primary care patients identified by screening.
Example 1 Secondary Analysis Park et al./Changes in health outcomes as a function of abstinence and reduction in illicit psychoactive drug use: A prospective study in primary care	Statistical analysis using ANOVA, Spearman's correlation, and linear regression models/2015/528	Same as above	(1) test the associations between longitudinal illicit drug use patterns including abstinence and health outcomes and (2) test whether these associations vary by drug type among illicit drug users in primary care.	Same as above	Same as above	Abstinence, but not decreased drug use without abstinence, was associated with improvement in drug use consequences, compared to those with continued or increased drug use. This relationship was found among those whose main drug was cocaine and opioids, but not marijuana

Example 2 Parent Study Kopansky-Giles D & Papadopoulos C./Canadian chiropractic resources databank (CCRD): a profile of Canadian chiropractors	Survey/2011/ 2,529	Members of the Canadian Chiropractic Association (CCA)	To profile practice and concerns of DCs	Survey	Professional activities, education, research and teaching activities, main sectors of activity, care provided to patients, chiropractic techniques used, type of conditions treated, and referral practices	Resource data bank profiling Canadian chiropractors
Example 2 Secondary Analysis Blanchette, M., Cassidy, J. D., Rivard, M., & Dionne, C. E./Chiropractors' characteristics associated with their number of workers' compensation patients	Cross sectional analysis/2015/ 2,529	Same as above	To describe the characteristics of Canadian DCs who tend to treat more workers' compensation cases	Same as above	Same as above	Canadian DCs with practices oriented toward the treatment of injured workers that collaborate with other health care providers and facilitate workers' access to care reported more workers' compensation patients
Example 3 Parent Study Algase, D. L. et al./Are wandering and physically nonaggressive agitation equivalent?	Cross-sectional correlational design/2008/ 181	Ambulatory residents of 22 SNFs and 6 ALFs meeting DSM-IV criteria for dementia	Examined equivalence of wandering and physically nonaggressive agitation (PNA) as concepts	Video-tapes for up to twelve 20-minute observations per participant	Wandering and PNA behaviors	Wandering and PNA as overlapping, but nonequivalent phenomena

Example 3 Secondary Analysis Lee, K. H., Boltz, M., Lee, H., & Algase, D. L./ Is an Engaging or Soothing Environment Associated With the Psychological Well-Being of People With Dementia in Long-Term Care?.	Cross-sectional correlational design/2017/ 177	Same as above/ participants who completed more than three emotional expression observations that evaluated psychological well-being	To examine the relationship between environmental ambience and psychological well-being of persons with dementia	Same as above/ observed displays of positive and negative emotional expressions	Emotional Expressions	An engaging environment was associated with more positive emotional expressions, however, a soothing environment was associated with neither positive nor negative emotional expressions.
Example 4 Parent Study Verbeek, H., et al./ Dementia care redesigned: Effects of small-scale living facilities on residents, their family caregivers, and staff	Quasi-experimental/ 2009/404 (229 Family caregivers, 259 residents, 305 nursing staff)	Residents, their family caregivers and nursing staff from small and large scale dementia facilities	The effects of small-scale living facilities in dementia care	Questionnaire/ scales	Residents: quality of life, neuropsychiatric symptoms, and agitation Family Caregivers: perceived burden, satisfaction, and involvement with care Nursing: Job satisfaction and motivation.	Residents: No effects were found for total quality of life, neuropsychiatric symptoms, and agitation. Family Caregivers: in small-scale living reported significantly less burden and were more satisfied with nursing staff than family caregivers in regular wards. No differences were found in their involvement with care.

						Nursing: No significant differences were found for staff's job satisfaction and motivation, although subgroup analyses using contrast groups (revealed more job satisfaction and motivation in small-scale living compared with regular wards.
Example 4 Secondary Analysis Adams, J., Verbeek, H., & Zwakhalen, S. G./ The Impact of Organizational Innovations in Nursing Homes on Staff Perceptions: A Secondary Data Analysis	Cross sectional/2017/138	Nursing staff from small and large scale dementia facilities	(a) explore staff perceptions about skills warranted in both small and large scale dementia facilities (b) determine differences in job satisfaction, motivation, and job characteristics of staff between the two care settings	Same as above	Job satisfaction, motivation, and job characteristics	Job satisfaction, job motivation, Job autonomy, and social support were significantly higher in typical small-scale nursing homes, compared to those in typical large scale (traditional) nursing homes.

Conducting research using an existing data set is not always advisable, and there are several major areas of potential concern that should be considered. First, existing data sets, inevitably, will not be a perfect match with what the researcher might have collected had the study been designed prospectively, so, limitations relevant to what data are available must be considered.[14] Second, research questions best answered by conducting a randomized controlled clinical trial cannot be answered by an existing data set since there is no way to randomize participants into an intervention group after the data set has already been collected. Third, especially in the social sciences, history bias presents a potential problem, particularly if the data set is more than a few years old. Fourth, at times, there may be no available existing data set to answer a particular research question, in which case a prospective study would be the only way to reasonably obtain results relevant to the research question.

1.5. Selecting a Data Set

Increasingly, making data sets available to researchers is seen as an ethical mandate. Nursing organizations, such as Quality & Safety Education for Nurses (QSEN) and various oncology agencies and groups, recognize the importance of utilizing available data.[15] Because it can facilitate results more quickly than when recruiting subjects is required, using existing data reduces disparities in care by demonstrating best practices more quickly. Also, as previously mentioned, the avoidance of new research participants being subjected to risks provides another major ethical benefit of using existing data. [4 14 15 16]

As per Table I, researchers doing secondary analysis should seek out a data set early in the process, after reviewing the literature and determining potential research questions based on gaps in existing research literature. By this point, the researcher should have adequate expertise in the subject area to be familiar with key concepts associated with the general research area. As research questions become more clearly defined, a more thorough analysis of existing literature is advisable to optimize expertise in the specific area and to generate potential sources of existing data sets. Having a firm understanding of the research questions will help guide the PI to data sets that include appropriate variables, samples, and study designs. [1]

A strong theoretical basis for the proposed research is extremely beneficial when determining the best possible data set. Researchers obtaining an original data set need to justify the importance of their research prior to their recruitment of the first subject. Similarly, those using an existing data set to conduct research need to have a firm understanding of the theoretical underpinnings that justify the research when identifying the best possible data set.

Table III: Possible Sources of Existing Data Sets

Name	Address
The United States Census	https://www.census.gov/
University of California San Francisco	www.ucsf.edu
The Mayo Clinic	www.mayoclinic.org
The Cleveland Clinic	https://my.clevelandclinic.org/
QSEN	www.qsen.org
National Center for Education Statistics	http://nces.ed.gov
Science and Engineering Statistics	www.nsf.gov/statistics
Cal-PASS Plus	www.calpassplus.org
California Health Interview Survey	http://healthpolicy.ucla.edu/chis/Pages/default.aspx
Behavioral Risk Factor Surveillance System	https://www.cdc.gov/brfss/index.html
National Health & Nutrition Examination Survey	https://www.cdc.gov/nchs/nhanes/index.htm
Partners in Info. Access for… Pub. Hlth Wkforce	https://phpartners.org/index.html
Data.gov	https://www.data.gov/

Table III contains some potential sources of data sets. Upon identifying a potential data source, the PI should consider the overall quality of the data set. Areas of interest may include whether the data set has the appropriate variables to evaluate the research question(s); sampling methodology; how variables are defined and measured; whether the sample size is sufficient such that the inquiry would be adequately powered; whether there is excessive missing data and/or a high loss to follow up in a longitudinal study.[1 3 4] The PI should consider specific aspects of the data, including original research proposals, recruitment plans, and research protocols.

The PI of the secondary analysis will need to build a persuasive case that the original data set is a good fit. The measures used in the original data set, most often, will not be a perfect match with what the PI may have hoped for, but they must be sufficient to potentially make meaningful scientific discovery. If the PI determines that a data set is of insufficient quality or lacking in variables required to answer the research question, it is best to seek out a more appropriate data set. Attempting to make a data set work if it is not well suited to answer the research questions is futile. However, it may be possible to fine-tune the research questions to better match the available data set, so long as the revised research questions are still relevant and important.

For proprietary data sets, the parent study PI may be saving certain analyses for other researchers. Thus, it is important to discuss clear guidelines with the original PI to ensure the specific analysis of data being requested is permissible.[1] If what is available to the PI to analyze is insufficient, then alternative plans should be made to avoid wasting time.[14]

When focusing in on a data source, the PI should collaborate with individuals that are expert in the specific data source being considered. Especially for large data sets, it is possible that coding and other data entry processes change over time, and the PI will need to make sure that the data is being correctly deciphered. Erroneous information could be reported if the PI has not gained sufficient knowledge of how to correctly interpret information from the data source. Ideally, the PI will work closely with the investigator from the original study, receiving codebooks and other information that can be used to correctly interpret the data. Some successful studies have occurred without substantial collaboration with the PI of the parent study; however, this should be done with some caution to avoid misinterpreting information.[3]

Certain types of databases tend to have specific strengths and weaknesses. Administrative data sets, which are commonly used in oncology research, tend to lack the level of clinical detail that might be useful since they tend to be originally obtained for non-research purposes. However, they are often well suited for examining healthcare practice variation and other questions for which an overview of trends would be enlightening, whereas patient registries are well suited for more disease and/or treatment specific data.[4] Indeed, a registry tends to be best suited for research questions relevant to the population that is predominant within the registry.

There is the potential to merge multiple data sets together. However, exercise caution as data corruption is an issue when manipulating data sets in this way. Specifically, there is a risk of merging variables that may not have been measured as reliably in some of the merged studies.

1.6. IRB Approval

Institutional Review Board (IRB) approval provides a mechanism for the researcher to demonstrate a plan to protect subjects' private information. Also, obtaining IRB approval prior to analyzing data from an existing data set is one important way to make sure that the answers being sought from the data set are prudent. Analysis of existing data sets should be purposeful. Given an alpha of 0.05, common in social sciences, we would expect 1 in 20 significant findings to be based on pure coincidence. So, it would be a major mistake to aimlessly compare variables of an existing data set simply because the data are available. Having a sound research plan based on a competent literature review and approved by the IRB is a good way to make sure that comparisons and analyses being conducted with an existing data set have merit and are appropriate. Additionally, the researcher should do a power analysis to make sure that the data set is likely to find a significant result for areas of interest if such a result really exists.[5]

1.7. Special Circumstances

Using an existing data set may require the PI to transfer data from one storage format to another. Increasingly, existing data sets are digitally formatted. However, when transferring a data set into another statistical software program and/or otherwise manipulating the storage mechanism, there is a risk of corrupting the data. The PI should engage in rigorous quality controls, with backup files, to ensure that the data remains accurate and that no human error corrupts the data set.[1] Sound practice involves running descriptive analyses and making sure that post-data transfer results match those obtained prior to transferring the data set. As a general guideline, it is inadvisable to attempt to manually transpose data line by line. Such manual transcriptions are notoriously prone to entry error issues, and current technology tends to have much more reliable methods for data conversion.

PIs conducting a prospective study recruit as many participants as appropriate given the power analysis and then tend to stop recruiting. However, PIs using existing data sets may have extremely large sample sizes, far larger than would be required by a power analysis. Samples this large can show statistically significant differences with very small effect sizes. If this is the case, the researcher must consider what is a relevant effect size so as not to overemphasize trivial differences that may be significant predominantly as a result of an extremely large sample. Similarly, readers of a study with an extremely large sample size must get past the tendency to applaud statistical significance alone in favor of higher scrutiny of the effect size. In healthcare research, this entails considering if the effect size is clinically meaningful.

Just as those gathering an original data set need to strongly consider sampling bias, problems associated with a biased sample can be a particular problem for those using an existing data set to answer new research questions. [16] It is important for the PI to know the response rate, inclusion and exclusion criteria, sampling pool, methods, and other relevant aspects of the original study. Often, several decisions made during the original study can adversely affect what population the sample most represents in a secondary analysis. These issues include: study samples do not tend to be true random samples; population subgroups may (purposefully) be oversampled in the original study; historical bias among sampling subgroups; stratification of variables systematically alters the samples representativeness as it pertains to the general population; and non-respondents might be excluded from analysis, further altering findings. [16] Fortunately, there are statistical methods that can address several of these problems. Further information is published elsewhere as to how to do so. [16] When taking steps to address sampling problems, it is important to adequately outline the steps that were taken when publishing research results so that the reader can make appropriate determinations regarding the perceived veracity of the research findings.

1.8. Reporting Results of a Secondary Analysis

A well reported study that uses an existing data set will include sufficient context such that journal editors and readers can reasonably determine that the PI has competently addressed the challenges associated with using an existing data set. Specifically, the PI should explicitly state why the chosen data set is an appropriate fit for the research questions. If applicable, the researcher should describe steps taken to make sure that the data set was interpreted correctly, highlighting relevant collaboration with those most familiar with the original data set. Additionally, clearly state limitations and challenges associated with study variables including recruitment plans that are not a perfect match with what the PI might have planned.

1.9. Conclusion

A researcher using an existing data set to answer new research questions needs to take specific steps relevant to this research methodology. Among those, it is particularly important that the PI be able to build a strong case that the data set selected is a good match to answer the proposed research question. When reporting results, the PI should discuss steps taken to ensure appropriate rigor.

Acknowledgements:

The authors wish to thank their Department Chair Lynn Van Hofwegen RN, DNP, Dean Jason Singley PhD, and California State University East Bay for being supportive of this work. Additionally, the authors would like to thank all their peers and mentors, particularly those from California State University East Bay and the University of California San Francisco. No specific funding was provided to the authors to complete this work.

References

1. Doolan, DM & Froelicher, ES. Using an existing data set to answer new research questions: A methodological review. *Research and Theory for Nursing Practice: An International Journal*. 2009;23(3)203-15.
2. Doolan, DM, Stotts, NA, Benowitz, NL, Covinsky, KE, & Froelicher, ES. The Women's Initiative for Nonsmoking (WINS) XI: Age-related differences in smoking cessation responses among women with cardiovascular disease. *The American Journal of Geriatric Cardiology*. 2008;17(1)37-47.
3. Abeysekera, I. Secondary analysis of two environmental practice studies: Do empirical variables represent expressed theoretical constructs? *Journal of Cleaner Production*. 2014;79, 7-17.
4. Rice, HE, Englum, BR, Gulack, BC, et al. Use of patient registries and administrative datasets for the study of pediatric cancer. *Pediatr Blood Cancer*. 2015;62(9)1495-1500. doi:10.1002/pbc.25506.
5. Marler, JR. Secondary analysis of clinical trials-A cautionary note. *Progress in Cardiovascular Disease*. 2012;54(4)335-337. DOI: 10.1016/j.pcad.2011.09.006
6. Adams, J, Verbeek, H, & Zwakhalen, SG. The impact of organizational innovations in nursing homes on staff perceptions: A secondary data analysis. *Journal of Nursing Scholarship*. 2017;49(1), 54-62. DOI:10.1111/jnu.12271
7. Algase, DL, Antonakos, C, Yao, L, Beattie, ER, Hong, GR, & Beel-Bates, CA. Are wandering and physically nonaggressive agitation equivalent? *American Journal of Geriatric Psychiatry*. 2008;16(4) 293–299.
8. Blanchette, M, Cassidy, JD, Rivard, M, & Dionne, CE. Chiropractors' characteristics associated with their number of workers' compensation patients. *J Can Chiropr Assoc*. 2015;59(3), 202-215.
9. Kopansky-Giles, D & Papadopoulos, C. Canadian chiropractic resources databank (CCRD): A profile of Canadian chiropractors. *J Can Chiropr Assoc*. 1997;41(3):155-191
10. Lee, KH, Boltz, M, Lee, H, & Algase, DL. Is an engaging or soothing environment associated with the psychological well-being of people with dementia in long-term care? *Journal of Nursing Scholarship*. 2017;49(2), 135-142. DOI:10.1111/jnu.12263
11. Park, TW, Cheng, DM, Lloyd-Travaglini, CA, Bernstein, J, Palfai, TP, & Saitz, R. Changes in health outcomes as a function of abstinence and reduction in illicit psychoactive drug use: A prospective study in primary care. *Addiction*. 2015;110(9), 1476-1483. DOI:10.1111/add.13020
12. Saitz, R, Palfai, TA, Cheng, DM, et al. Screening and brief intervention for drug use in primary care: The ASPIRE randomized clinical trial. *JAMA: Journal Of The American Medical Association*. 2014;312(5), 502-513. DOI:10.1001/jama.2014.7862
13. Verbeek, H, Zwakhalen, SMG, van Rossum, E, et al. Dementia care redesigned: Effects of small-scale living facilities on residents, their family caregivers, and staff. *Journal of the American Medical Directors Association*. 2010;11(9), 662–670.

14. Westin, GF, Dias, AL, & Go, RS. Exploring big data in hematological malignancies: Challenges and opportunities. *Current Hemotologic Malignancy Reports*. 2016;11,271-279. DOI: 10.1007/s11899-016-0331-4

15. The QSEN Institute. Quality and Safety Education for Nurses. www.qsen.org. Accessed August 1, 2017.

16. Bell, BA, Onwuegbuzie, AJ, Ferron, JM, Jiao, QG, Hibbard, ST, & Kromrey, JD. Use of design effects and sample weights in complex health survey data: A review of published articles using data from 3 commonly used adolescent health surveys. *Research and Practice*. 2012;102(7) 1399-1405.