1. Multiple uses of large data sets

The research enterprise is both human and financially resource intensive, whereas the funding for research is highly competitive. How then can an investigator find ways to begin and sustain a viable program of research in a focus area of science absent of a significant infusion of funding? One option that many investigators do not initially consider is conducting secondary analysis using large health care data sets. Such large data sets are available from federal and state agencies and associations for nominal fees, if not for free. They include patient-level data or aggregated data for a particular type of patient or payor. These data are used by clinical and social scientists as well as by health services researchers to conduct analyses to inform quality agendas, health policy, and health care coverage determinations. Conducting secondary analysis using these large data sets can be the focus of an investigator’s entire program of research or can lead to intervention or translational research in the clinical setting.

Large data sets facilitate replicating generalizable findings and understanding patient outcomes. By far, most large database analysis for researchers involves secondary analysis of existing data sets to explore clinical phenomena and answer new research questions. In part, this is achievable because existing databases offer the investigator rich sources of data to answer these research questions when funding opportunities may be limited or scarce and when primary data collection may not be economically or logistically feasible. Secondary analysis is a legitimate form of scientific inquiry in which nurses should be actively engaged (Castle, 2003).

Although large data sets, ranging from administrative data to large-scale surveys, are collected for a specific purpose, they can be used for multiple purposes. Data elements are extensive and range from clinical and charge data for procedures to demographic data about each patient. In some cases, there can be more than 100 variables for each patient case. Researchers who want to investigate a policy question linked to clinical and cost outcomes can do so with existing sources and can even track these outcomes related to policy changes over time. The size of the data depends on scope and the number and frequency of variables. Only a few years ago, very few researchers had the computing capacity to use large data sets on mainframes. Today, many large data sets can be managed and analyzed on personal computers and laptops.

2. Advantages of large data sets

2.1. Large numbers of cases

Depending on the data set, there can be literally millions of cases in a large administrative data set. For instance, the Health Cost and Utilization Project (HCUP) of the Agency for Healthcare Research and Quality (AHRQ) contains a collection of patient-level data submitted by participating states each year from millions of hospital discharges. Each year of data contains more than 100 variables for each patient or case and includes demographic, clinical, and charge data.

2.2. Cost

As previously stated, the cost of conducting primary data collection for a phenomenon of interest can be extraordinary and beyond the reach of some investigators. These data sets can be purchased for nominal fees or, in some cases, are free. For example, each HCUP data “year” can be purchased from AHRQ for US$200.00. The investigator need only complete data agreement training that provides specific guidelines for how the data can be conveyed and complete the appropriate forms of use. Graduate students may purchase the data at a significantly lower fee.
2.3. Linking patient outcomes with policy changes

Health care in the United States is a dynamic process that is primarily impacted by decisions made outside of the health care industry, specifically at the policy level. Many important research questions can be analyzed using large databases using a pre/post design. An investigator may want to determine how a policy change at a given point in time influenced patient-level variables such as clinical or cost outcomes or numbers and types of procedures. In fact, if a researcher has interest in the temporal changes of a particular health care policy, large databases (e.g., national survey databases such as the Medical Panel Expenditure Survey and administrative databases such as Medicare, Parts A through D) are an excellent way to trend the influence of such policy changes. Although the investigator may not report findings as “causality,” there are ways to show relationships between policy and health care consumption and outcomes.

2.4. May reduce time from proposal to dissemination

Investigators who conduct secondary analysis will find the process of securing institutional review board (IRB) approval significantly reduced when requesting review of their proposal. Because all IRBs are not the same, the research proposal using large data sets may be deemed expedited or exempt, which typically can translate into a much shorter start-up time. In addition, once the data set has been secured and the analysis has been completed, the results can be prepared for publication in the time it might take another investigator to obtain initial IRB approval. This shortens the length of time from the determination of the research question to the dissemination of the findings to the scientific community or other stakeholders.

3. Disadvantages of large data sets

3.1. Data quality

By definition, secondary analysis means that the researcher who wants to use this type of data is not involved in the development of the operational definitions for study variables or the data collection process. Furthermore, because every database has limitations, researchers must be diligent about data quality during each phase of analyzing the data. Secondary data must be used with the same rigor that one would use in collecting and using primary data. This may include testing the quality of the data by comparing sample size and response rates with published findings. Additionally, variables need to be fully understood in terms of coding procedures and conventions as well as potential biases. Schwartz et al. discussed data-quality issues that are extensive to include variability among persons who abstracted the medical records for entry into the large data set as well as incongruence between physicians with International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) coding. These authors also address the “race” variable, which can be challenging to use particularly in light of persons of mixed heritage when confronted with limited response options (Schwartz, Gagnon, Muri, Zhao, & Kellogg, 1999).

Most groups that collect and manage data for these large data sets have worked extremely hard over the years to reduce variability in data collection and reporting by providing assistance and clarity to those who contribute to the data set. Moreover, the advent of the electronic medical record may improve the overall quality of data by limiting the need to manually abstract data for such large data repositories (Iezzoni, 1997). Ultimately, it is up to the investigator who uses the data set to fully appreciate the limitations of the data and to report this in publications or presentations.

3.2. Missing data

In large administrative data sets, one may have significant numbers of missing data on certain variables. For instance, if different hospitals contribute information to a particular database, they may or may not collect or contribute a certain variable for institutional reasons, such as those mandated by their respective IRB. Investigators must gain an understanding of missing data and determine if this limits the findings to such an extent to reduce the usefulness of the research altogether. As long as the researcher understands the complexities of missing data and handles them properly through the use of appropriate statistical methods, these issues associated with missing data can be minimized (E-Masri & Fox-Wasylyshyn 2005; Patrician, 2002). However, it is important to note that even in primary data collection, there may be significant missing data. It is a universal problem regardless of study design.

3.3. Health Insurance Portability and Accountability Act

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) can be a double-edged sword for any investigator using large databases. On the one hand, the data agreements typically required by database owners or agencies prohibit reporting findings in a way that would reveal not only the patient involved, if names are even part of the data set, but also the institutions that are reporting the data. On the other hand, if there is such a heightened sense of concern to not violate HIPAA, there can be issues of securing data from organization or associations that may translate into major barriers for securing a database.

4. Other considerations

Secondary data analysis demands expert knowledge of research designs. Before any analysis begins, researchers must understand the distinctive methodological and practical challenges involved. There can be problems involving sampling, measurement, and external and ecological validity. It is also important to understand key factors in
maintaining the integrity of the data files as they are moved, merged, and transformed. One way to assure that a research project conforms to scientific standards is to work with an experienced researcher who has previously and successfully used large data sets and to assemble a team that can ensure the scientific integrity of the project from start to finish (Zhan & Miller, 2006).

5. Main sources of data

Large databases other than those from federal sources may be available. Associations may be compelled to work with an investigator using their data with the appropriate assurances to report findings in a way that both meets the needs of the researcher and protects the proprietary nature of the data. A good start at such collaboration would be if an investigator has been an active member of the organization or association for some time. These groups are always in pursuit of leveraging resources to extend the work of their organization. The symbiotic nature of this type of arrangement can be beneficial for both parties.

6. Handling the data: Some tricks of the trade

6.1. Understanding the data

Just because one was not involved in the collection of the data does not preclude one from being required to fully understand each variable. Any variable used without a full understanding of how the variable was defined and collected will adversely impact one’s research. For example, it is essential to understand whether a variable is independent or dependent. It is also important to always expect anomalies (e.g., only a three-digit code for diagnosis for some patients or sites of care, instead of the more useful five-digit code), zeros being used to fill blank cells, and decimal points missing (e.g., cost information that includes both dollars and cents but no decimal point is included in the data field). The usability of each variable will determine its usefulness in research. As such, one must assess each variable in terms of potential threats to internal and external validity as well as to reliability and generalizability. This is also important when variables from different large data sets are combined.

6.2. Cleaning the data and creating an analytic data set

Rarely can one use a secondary data set without engaging in several key steps to create an analytic data set for research purposes. First, one must select which variables in the data set will be needed, primarily because the data set was most likely created for other purposes and with different research questions in mind. For example, Medicare data contain some variables for retrospective payment (e.g., provider number) that are not useful in patient outcome analysis if the researchers are not able to use identifiable variables in their analysis. Second, the analytic data set is initially created from selected variables in the secondary data set, and then, frequencies need to be assessed to determine if each variable was appropriately represented based on how it was defined and on how it was intended to be used. For example, to assess the incidence of infectious diseases, the ICD-9-CM-coded variables that should begin with a “V” may not be properly represented because the variables were only coded as numeric. Third, because the data set may not contain all the information the researcher wants, it is often necessary to create analytical variables that can be derived from those in the data set to create the analytic data set for research analysis. Lastly, it is important to evaluate the quality of the data set itself, in a stepwise approach, to address possible concerns about bias and potential confounding factors and to determine if case-mix adjustment is necessary before data analysis begins (Wunsch, Linde-Zwirble, & Angus, 2006).

6.3. Statistical methods

Researchers must understand both basic and advanced statistical methods when using large data sets. Unlike assessing initial descriptive statistics with small data sets, it is important to test the actual statistical code using a small sample of the data before running the entire data set. This is because some data analysis runs take awhile to be processed by the computer because they may demand thrice the space taken by the actual analytic data set—a common issue in using large data sets—to perform the test and may be costly or may slow one down if one does not test the accuracy of the statistical code beforehand. It is important to use statistical software packages that are capable of analyzing large data sets (e.g., SAS), especially when the data are derived from complex sampling designs. Most national health care surveys involve inherent sampling design issues that require the application of sample and variance estimation weights when conducting statistical analyses.

Initial analysis of the data involves basic statistics (e.g., frequency distributions) and correlation coefficients to assess the association between two continuous distributions. Then, while there are multiple statistical tests that can be employed, both t-test and regression models are appropriately (Chen, 2002) and commonly used to analyze large data sets, focusing on assessing associations and trends. For example, multiple regression analysis is appropriate for multivariate analyses with a continuous outcome variable (e.g., length of hospital stay) and logistic regression methods are appropriate for binary outcomes (e.g., yes or no).

Using large data sets for research requires an open and inventive mind. Secondary analysis is not an easy process. When using large data sets and performing secondary data analysis, the quality of the research and its findings depend on using the most appropriate study design and analytic methods. Appropriately using secondary data sets can enable nurse researchers to make significant contributions to nursing science.
References


