Canadian Journal of Nursing Research, 2000, Vol. 32, No. 3, 109–129

Secondary Analysis: Theoretical, Methodological, and Practical Considerations

Sean P. Clarke and Sylvie Cossette

L'analyse secondaire, qui implique l'utilisation d'ensembles de données existantes pour répondre à de nouvelles questions de recherche, constitue un choix méthodologique de plus en plus utilisé chez les chercheurs désireux de se pencher sur des questions spécifiques mais ne disposant pas de ressources permettant d'effectuer une cueillette de données primaires. Toutefois, cette voie peut entraîner une perte de temps et une frustration importantes chez les chercheurs qui entament des analyses secondaires sans être conscients des défis méthodologiques et pratiques distincts présents dans une telle démarche. Cet article met en lumière les difficultés qui peuvent survenir lorsque les chercheurs utilisent des données provenant de projets de recherche clinique antérieurs, y compris des questions théoriques et des problématiques nécessitant une démarche d'échantillonnage, de mesurage et de validation externe et écologique. Il offre également des suggestions pratiques pour entreprendre une analyse secondaire et présente les critères d'évaluation qui s'y rattachent.

Secondary analysis, which involves the use of existing data sets to answer new research questions, is an increasingly popular methodological choice among researchers who wish to investigate particular research questions but lack the resources to undertake primary data collections. Much time loss and considerable frustration may result, however, if researchers begin secondary analyses without an awareness of the distinctive methodological and practical challenges involved. This article highlights difficulties that may arise when researchers use data from previous clinical research projects, including theoretical issues and problems involving sampling, measurement, and external and ecological validity. It also offers practical suggestions for undertaking a secondary analysis and criteria for evaluating secondary analyses.

Secondary analysis is "any further analysis of an existing data set which presents interpretations, conclusions or knowledge additional to, or different from, those presented in the first report on [an] inquiry and its main results" (Hakim, 1982, p. 1). Put simply, secondary analysis involves the use of existing data sets to answer new questions. Given the huge human and financial investment in data collection by funders, investigators, and subjects, it seems reasonable if not imperative, ethically speaking, to make the maximum use possible of research data

---

Sean P. Clarke, RN, PhD, CRNP, is a postdoctoral fellow, Center for Health Outcomes and Policy Research, University of Pennsylvania School of Nursing, Philadelphia, USA. Sylvie Cossette, RN, PhD, is Assistant Professor, Faculty of Nursing, Université de Montréal, Quebec, Canada.

(Hakim; Hyman, 1972). New uses of data may take many forms. Secondary analysts may select dependent variables from the data set that were not the focus of the initial analyses and examine their correlates, or they may explore further predictors of the primary outcome in the original study, or they may choose to examine relationships among variables in subgroups of subjects in the data set. Secondary data can have many other uses — for instance, validation of instruments and testing of complex theories, such as the buffering effect of social support on health in data sets where indicators of both concepts have been included.

The possible economies of money and time, compared with those entailed in mounting a primary study, have made secondary analysis an increasingly popular methodological choice, especially for doctoral students and researchers starting out in a particular area of inquiry. Nonetheless, much time loss and considerable frustration may ensue if researchers begin secondary analyses without an awareness of the distinct challenges involved. Unfortunately, limited guidance is available in the form of published discussions of the practical aspects of planning and conducting secondary analyses. Authors have touched on the advances in nursing research and knowledge development that might result from increased sharing of research data by nurses (Estabrooks & Romyn, 1995). A number of authors have discussed the use of large, publicly available databases in health research (Jacobson, Hamilton, & Galloway, 1993; Lange & Jacox, 1993). Still others have discussed the general way in which secondary analyses can be conducted (Mainous & Hueston, 1997; McArt & McDougal, 1985). Little attention has been paid, however, to difficulties that may arise in using data from previous clinical research, which is the major source of secondary data for nurse researchers.

To address this gap, this article will examine theoretical, methodological, and practical aspects of the planning and execution of secondary analyses. While qualitative data can be re-analyzed (Szabo & Strang, 1997), the focus here will be quantitative data sets. Furthermore, although secondary analysis may be conducted by the researchers responsible for a primary data collection, this discussion will assume that the secondary analyst was not involved in the design, conduct, or analysis of the original study. To provide a context for the discussion, occasional reference will be made to the first author's doctoral research, a study of psychosocial correlates of health outcomes in patients with serious heart disease that involved secondary analysis of data from a series of randomized controlled drug trials (Clarke, 1998).

Theoretical Issues

The theoretical questions at the core of a secondary analysis are no different from those of any other type of research project: What are the variables of interest to the researcher? How did they come to be selected? What relationships between the variables of interest will be examined, and what conceptual framework or frameworks define these relationships? What means are available to measure them? (Polit & Hungler, 1995) How do different measures of the same concept interrelate? How do these measures overlap, how are they different, and which measure is preferable for a given study, and why? (Pedhazur & Schmelkin, 1991) The secondary analyst who reaches the point of data analysis — or, worse yet, manuscript preparation — without having answers to these questions will face serious difficulties. The theoretical underpinnings of any study are critical to every aspect of its planning and execution.

Concerns are sometimes voiced that secondary analysis is undertaken because it is somehow "easier" than primary research. Some feel that secondary analysts bypass the long process of framing research questions and that "ready-made" data sets enable researchers to skirt the issues of definition and operationalization of major concepts (Kasl, 1995). In avoiding the many difficulties that plague original data collection (such as securing the necessary funds and gaining access to subjects and obtaining reliable information from them, especially in health-care contexts), Kasl states, secondary analysts are too often forced to make unacceptable compromises.

Perhaps some of the concerns stem from differences of opinion on the nature of the research process. Many investigators have been taught to view research as a linear process whereby the researcher begins with a literature review that reveals gaps in knowledge about a phenomenon. After reflecting on a theoretical framework to guide his or her study, the researcher chooses a population, a set of variables, and a series of relationships; selects measures; plans data collection and analysis; then gathers and analyzes the data and writes for publication (Brink & Wood, 1994; Polit & Hungler, 1994; Wilson, 1989). At least one version of this worldview posits that in "good research" the study's methodology flows from research questions, not the other way around.

In practice, however, the planning of studies involves simultaneous consideration of the state of the literature, what can be measured, what subject pools might be available for data collection, and what analytic methods exist and are within the researcher's expertise and budgets of

time and money. A linear approach, in which each "step" of the research process or portion of the research protocol is completed before the next one begins, is rarely possible. Research questions may be stimulated not only by reflecting on theories, and inconsistencies and gaps in the literature, but also by considering ways in which existing measurement tools or data sets can be put to use. Reflection on a theoretical framework may or may not occur at the outset of a primary study or secondary analysis. To be sure, using existing data places serious constraints on the secondary analyst. However, in primary research as well, compromises and deviations from methodological ideals, simultaneous rebalancing of multiple theoretical and design factors, and repeated recasting of research questions are nearly always required. In any research project, the methodological compromises must not invalidate the researcher's conclusions and the researcher must situate the finished product within a substantive or theoretical frame of reference. The order in which theoretical and methodological concerns were resolved in the development of a particular study is probably far less important than the theoretical integrity of the overall project.

When the dissertation research referred to in the introduction was being planned, the preliminary questions dealt with the relationship of psychosocial factors, such as social support, psychological distress, and socio-economic status, to patient well-being, clinical events, and mortality in heart failure (Clarke, 1998). The questions were stimulated by research suggesting that psychosocial characteristics predict health outcomes in forms of cardiac disease other than heart failure (Bucher, 1994; Frasure-Smith, Lespérance, & Talajic, 1993; Williams et al., 1992). The guiding theoretical perspective was the biopsychosocial model, which proposes that there are interconnections between physical health and the psychological and social worlds of individuals (Shaver, 1985). The original intent was to assemble a cohort of several hundred heart failure patients from local clinics, administer a baseline questionnaire, and follow the patients for approximately 1 year. Early on, however, a data set became available from a major clinical trial, Studies of Left Ventricular Dysfunction (SOLVD), involving more than 6,000 patients with left ventricular impairment, including a sizeable number with established heart failure (SOLVD Investigators, 1990, 1991). In SOLVD, patients completed an extensive series of paper-and-pencil questionnaires addressing psychosocial factors in a quality-of-life battery at the outset of their follow-up in the trial and at regular intervals thereafter (Rogers et al., 1994). An impressive array of biomedical variables were measured for each patient (including a variety of laboratory measures), and all subjects were meticulously followed for hospitalization, clinical

deterioration, and death over a median period of 3 years. Although the SOLVD data were not perfect for testing all of the hypotheses of interest, they were nonetheless far more comprehensive than would be possible in a primary data collection for a doctoral research study.

Methodological Issues

All departures from methodological ideals, particularly threats to internal and external validity, weaken confidence in the results of any study (Abelson, 1995; Campbell & Stanley, 1966; Pedhazur & Schmelkin, 1991). This is certainly true of secondary analyses. While it is difficult to generalize across all secondary analyses and data sets, three types of methodological concern are of special interest to secondary analysts and readers of their work. None of the three is unique to secondary analysis per se; however, all are quite common when previously collected data sets are used. They are: sample biases, measurement issues, and aspects of the original study (conditions of data collection) that may compromise generalizability to the populations of interest.

Table 1 *Major Methodological Issues in Secondary Analysis*

Sampling

- Selection biases: inclusion/exclusion criteria
- Representativeness of original sample
- Sufficient variability on key concepts
- Reductions in analytic sample due to missing data
- Selection of subgroups from a larger primary data set

Measurement

- Shortcomings of original researchers' instruments
- Conceptual slippage in proxy measures

External and ecological validity

- Original conditions of measurement may be at odds with conditions of interest
- Time elapsed since original data collection

Sampling

With the exception of data sets from studies that have enrolled probability samples from general or patient populations, sampling bias is a near-universal problem in health research. It is a basic methodological principle that all inclusion and exclusion criteria limit the generalizability of results and that such limitations are exchanged for increased ease of data collection and analysis, especially in clinical trials (Friedman, Furberg, & DeMets, 1985). In secondary analysis, generalizability is driven by the original researchers' decisions about entry criteria, and these may well have been based on concerns different from those of the secondary analyst. For instance, in clinical trials many conditions and patient characteristics are measured and controlled to derive a relatively homogeneous sample, in order to facilitate the drawing of conclusions about the effectiveness of a given intervention.

The secondary analyst must be thoroughly acquainted with the methods used to accrue the original sample. Whether or not a research sample is representative of the population may be relatively unimportant for a primary researcher evaluating the effectiveness of a medication, yet it could be pivotal for a secondary analyst interested in the relationships of one measured variable (e.g., social support) to other measured variables in general populations of patients and families. By the same token, variability across subjects on psychosocial or clinical characteristics in a data set may be relatively unimportant to the primary researcher but crucial to the secondary analyst. The characteristics of convenience samples of patients and families will be strongly influenced by the types of individuals who visit or seek care at the clinics, hospitals, or other recruitment sites used by the primary researcher.

A good example of potentially serious sampling issues can be found in the first author's doctoral dissertation. A design feature of the original clinical trials was the screening and exclusion of subjects who did not reliably take study medications (defined as taking less than 75–80% of the medication dispensed at particular visits) (SOLVD Investigators, 1990). This process enabled the original researchers to consider adherence a relatively minor factor in the interpretation of their analysis. However, in the secondary analysis (Clarke, 1998), in which psychosocial factors were examined as possible correlates of a variety of health outcomes, having both compliant and non-compliant subjects might have been key to identifying psychosocial predictors. Social support and depression are believed by some scholars to influ-

ence health outcomes through their influence on adherence to medication and other disease regimens. Depressed and socially isolated individuals, for instance, have been shown to take their medications and engage in condition-related self-care less consistently than others (Carney, Freedland, Rich, & Jaffe, 1995; Cohen, 1988; Horwitz & Horwitz, 1993) and thus may be at higher risk of adverse health outcomes. By removing non-compliant patients from their final sample, the original researchers may have removed a high proportion of depressed and socially isolated individuals from their pool of prospective subjects. The results of the secondary analysis suggest that there are psychosocial correlates of health outcomes even among highly compliant patients with left ventricular dysfunction. However, the loss of non-compliant patients from the original sample might well have reduced the magnitude of the observed effects.

Representative sampling can also be compromised by missing data. Of course, it is to be expected that fewer than the full complement of the original subject pool will be available for any particular analysis because of missing data for certain variables. Many large data sets contain observations from more than one time point; in any longitudinal data collection, however, there will nearly always be at least some loss of subjects over time (Polit & Hungler, 1995). Subjects who fail to complete all measures in a study may differ significantly from those who have data for the entire set of variables. In individuals with medical illnesses, for instance, failure to complete measures is sometimes thought of as a proxy measure of disease severity (Fairclough & Gelber, 1996). If missing data forces the secondary analyst to select a subset of the original subjects, there will be consequences for sample size and statistical power, and there may well be an exacerbation of problems with selection bias and a further limiting of generalizability. Practical issues regarding the handling of missing data will be addressed in the next section.

The secondary analyst's own decisions regarding which subjects from the original study to include may also influence generalizability. Not all of the subjects may be suitable for the analyst's specific concepts of interest. However, drawing subsets of the original sample can have consequences for sample size and representativeness. For instance, when only the control or the experimental group of a clinical study is relevant to the secondary analysis and subjects have been randomly allocated to these groups, statistical power (and not sample bias) will be of concern. Obviously, all subject losses, from those screened out in the primary study to those excluded in the secondary analysis, must be

carefully reviewed. There are risks in drawing conclusions about trends and effects in general populations from samples that are skewed by one or more forms of selection bias. Particular caution is required in drawing conclusions from analyses of subgroups of patients from samples that are non-representative in the first place (Oxman & Guyatt, 1992). For instance, in analyses of data on female patients that originate from convenience samples in which men are overrepresented, the conclusions may not be generalizable to women in the larger community.

Measurement

Measurement issues are the second category of concern facing secondary analysts and the readers of their research. Although construction of "new" scales or indicators from the items available in a data set may be possible, the secondary analyst must generally settle for the original research group's choice of measurement tools. In the data set analyzed in the first author's dissertation research, the original research team measured psychological well-being using the Profile of Mood States (POMS) (McNair, Lorr, & Droppleman, 1981). When the original data collection was planned, in the early 1980s, this well-validated tool was commonly used in studies of psychosocial aspects of illness. However, when the secondary analysis was being planned, in 1994, more comprehensive measures of clinically important psychological distress, better reflecting the state of the art in psychological assessment, would probably have been selected for a primary data collection. Choices would have included instruments measuring anxious and depressive symptomatology (e.g., the Spielberger anxiety scales, the Beck Depression Inventory, the Center for Epidemiologic Studies–Depression scale) (Boyle, 1985), or even standardized clinical interview tools. Nonetheless, in the secondary analysis the POMS scores were able to shed light on some issues of interest. Unfortunately, null (non-significant) results that were in conflict with those of other studies raised questions about whether the absence of some of the expected associations was attributable to the use of a depression measure that was not strictly comparable to the ones employed by most other researchers.

Imperfections are inevitable, and it is unlikely that the best and most recent measurement tools for addressing all of the concepts of interest will have been used in the primary study. The researcher considering the use of an existing data set must make a judgement about whether the instruments that were used reflect current thinking about

a concept and its measurement, and must ensure that the psychometric properties of the instruments are strong. If measures from sources outside the original study's data collection will be added to a data set (i.e., from other studies, new data collections, or special databases), these measures must also be reliable and valid.

The secondary analyst may be forced to construct alternative means of tapping important variables. In the data set used in the first author's dissertation research, for instance, marital status was not formally assessed in approximately half of the patients. A proxy measure of marital status (whether or not a question regarding marital satisfaction was answered) was developed. As another example, if no information about income or education is available in a data set, but the patient's zip or postal code is on file and socioeconomic status is of interest, publicly available census data could be used to locate information about average income and education in subjects' neighbourhoods and thereby serve as a stand-in measure (Krieger, 1992). The drawback of using proxy measures is that they may contain inaccuracies and they often incompletely address the concept of interest. The degree to which an available measure is only a partial or tangential measure of a concept has been called "conceptual slippage," and it has serious repercussions for the interpretation of both positive and negative (or non-significant) findings (Hyman, 1972).

New variables can often be added to an existing data set by using identifying data (for instance, names or health insurance or social security numbers) to gather information on health-care use or long-term mortality. There are, of course, ethical issues involved in using unique identifiers (like government- or provider-issued client numbers) to find out more about subjects than they had agreed to share upon entering the study. These ethical concerns will be outlined in the next section. From a methodological point of view, the reliability and validity of any measures added to the data set must be carefully assessed. It is particularly important that any new variables be collected in a consistent manner for all subjects. If, for instance, a specific hospital was used as a patient recruitment site for a primary study and some time later that same institution's medical records are used to gather data to extend the follow-up of the subjects, it must be remembered that patients change their venues of care over time. There will be far more comprehensive data available for individuals who continued to seek care at that institution than for those who went elsewhere for medical attention at the conclusion of the study, creating a potential bias in the secondary data set.

External and Ecological Validity

A study's results are generalizable only if they were found in a sample comparable to the population of interest and involved measurements obtained under conditions similar to those seen in practice. In assessing the external and ecological validity of a secondary analysis, one must consider the overall context of the data collection in the original study. Under what conditions and for what purposes were the data collected? By whom? When? What was subject burden in the original study? Heavy demands on patients in terms of interview time, lengthy questionnaires, or frequent data collection (of other types) may lead to subtle forms of selection bias, with only especially healthy or strongly motivated patients having valid data on the measures of interest. Lengthy data collection may also jeopardize the validity of the individual measures in the sample. A tired or bored subject may not exercise particular care in filling out the last pages of a 30-page questionnaire.

In determining whether the conditions of the original data collection limit external and ecological validity of results from the secondary analysis, several further questions must be asked. What features of the subjects' experience in the original study differed from those of conventional follow-up, and, by extension, in what frame of mind were the patients when they completed the measures? Were they seriously ill individuals who hoped that their participation in a clinical study would enhance their well-being or prolong their lives? Intervention is part of many clinical research studies, and being randomly assigned to one of several study conditions is a significant departure for patients being followed under non-research conditions. The patients in the first author's dissertation data set were followed at academic medical centres; all received intense surveillance, medical management, and specialty referrals that may not be representative of patient follow-up in the general population. Also, patients questioned at various points with respect to a treatment, an experimental manipulation, or a major life event may give quite different responses to questionnaires or interviews in each case. Reflection on the circumstances under which measures were performed in the original research protocol is pivotal to avoiding erroneous interpretations of secondary data.

Another element of context is time frame. Considerable time may elapse between the collection of data, the primary analysis, and the availability of a data set for a secondary analysis. If the data collection predates the introduction of a major clinical advance for a particular patient population, the relevance of older data to current practice must be addressed. Even in the small number of cases where a new treatment

eliminates the pool of patients with a given illness or complication, there may be important theoretical reasons for studying a particular set of relationships in an older data set. For instance, while new treatments change the natural histories of many diseases, the fact that illness is a major psychosocial stressor remains unchanged. Data collected some time previously can still yield important insights about, for instance, the process of patient and family adaptation to illness. That being said, the secondary analyst should be aware that opportunities to publish analyses of older data sets will be limited if peer reviewers see the data as excessively dated.

Summary

From the preceding discussions, it should be obvious that a methodologically "spotless" design is unlikely to be achieved in secondary analysis. However, assessment of potential problem areas is more complex than simple identification of the issues in the first place. In reports of secondary analyses, nearly every section dealing with limitations contains a sentence to the effect that conclusions are at best tentative because the data were not collected with the specific research questions in mind. While this is usually true, it is not especially informative. The reader is better served if the author pinpoints specific sampling, measurement, contextual, and other issues that are critical to weighing the credibility of the results and then discusses the directions that the results might have taken as a result of these problem areas.

Practical Issues

Having outlined the theoretical concerns in secondary analysis and addressed the various methodological issues that arise in this research approach, we now turn to some practical issues. While none of the steps in a secondary analysis is especially complex or unique, many are extremely time consuming. Fortune and McBee's (1984) comprehensive checklists of elements of the process are worth review; only the major phases and the pitfalls will be highlighted here.

Once the secondary analyst has located a prospective data set and its owner, he or she must ask some preliminary questions. At a minimum, the analyst needs to know the number and characteristics of the individuals in the data set (usually subjects or patients in clinical research) and which variables are contained in the data set. Once a data set is determined to be of potential interest, the conditions of use must be clarified with its owner. Research data are a scarce resource; thus the

Table 2 *Practical Issues in Secondary Analysis*

- Identifying a potentially useful data set and negotiating access to it
- Assessing data-set quality before making commitments/investments
- Obtaining ethics clearance, especially to seek further data on subjects
- Preparing the data for use and performing appropriate checks
- Dealing with missing data
- Interpreting results of exploratory analyses and analyses involving large sample sizes
- Reporting results of a secondary analysis in the proper context

owners of a primary data set may be quite protective of their interests when approached by outside investigators. Primary researchers are usually most concerned with credit and professional recognition such as authorship on papers, but control over the use of the data can be equally important. When the data in question are unique or of particularly high quality, the primary researcher may prefer to collaborate only with experienced colleagues in order to increase the odds that timely publication will result from the secondary analysis.

Considerable discussion has taken place in both Canada and the United States regarding planned regulations that would oblige researchers to share data from federally supported research after a "reasonable" amount of time (usually 5 years) has elapsed since data collection (Azar, 1999; Estabrooks & Romyn, 1995). Because there is currently no precedent for funding agencies to force researchers to hand over their data to other teams, data sharing remains voluntary. Even if options for obtaining data increase in the future, the prospective analyst should be aware that efforts to secure and analyze a data set will likely be fruitless without the cooperation of the primary research team.

Issues to be clarified with the original researcher include willingness to share the entire data set or only a certain portion of the variables, and payment or coverage of any expenses incurred in data sharing such as those for material or personnel resources. Access to the full range of variables collected for all subjects should not be taken for granted. The original researchers may offer the analyst a limited set of

variables, and may directly or indirectly limit the type of analyses permitted. For instance, they may have reserved specific analyses for themselves or their associates, or may not allow the secondary analyst to reexamine the original study's primary outcome, or may not permit analyses that question the premises of the original study or of their research program. It is equally crucial for the analyst to clarify the primary researchers' position regarding the right to publish from secondary data, and of course credit, especially authorship, on any publication that results from the secondary analysis. At a minimum, citing of the original investigators and their funding sources in any publications or presentations is generally required. Many owners of primary data sets will expect to share authorship credit, and may or may not be willing to assist with the analysis and preparation of abstracts and manuscripts in exchange. As with any research collaboration, directly addressing difficult questions at the outset may prevent conflict later on. From a number of standpoints, a written agreement stipulating the expectations of each party will be useful. Although sensitivity to the primary researchers' concerns is essential (Sieber, 1991), secondary analysts should exercise a healthy pragmatism in deciding whether access to the data is worth the conditions being imposed.

Once a data set with variables of interest has been located and a preliminary understanding about access has been reached, a more in-depth evaluation of the data set is in order. Adequate documentation is vital. If, however, labelling and record-keeping have been less than perfect (as is often the case), it is essential that the original study personnel be available to answer questions. Publications from the primary study will be useful, but detailed protocols and procedures, and (if available) variable lists and descriptive statistics for the variables of interest, must also be studied. If the secondary analyst was not involved in the primary data collection, evidence of careful and consistent data collection is key. Did the original data collection proceed as planned? Did sampling and recruitment strategies change mid-course? A more subtle issue is the relative importance of the measures of interest for the original research team. If the major variables in the secondary analysis were not core measures in the original study, the data may be of poor quality and a high proportion of data points may be missing. An estimate should be made of how many subjects with desired characteristics have data available.

By this point, the preparatory work in conducting a secondary analysis has been completed and the researcher's attention turns to writing a proposal and obtaining review board or ethics committee clearance. These tasks are not always part of the process, but, especially

for a graduate student, protocol and proposal writing becomes an exercise in thinking through and formally presenting a coherent plan for analysis, and forcing the researcher to confront the issue of how much detail about the original data collection to present. As for ethics clearance, most consent forms in primary research restrict access to the original data forms (containing identifiers) to members of the primary research team (Polit & Hungler, 1995). If the data set is denominalized (cleared of all identifying information), ethics clearance for secondary analysis is usually expedited. Otherwise confidentiality becomes an issue and the review board will be particularly interested in the content of the original consent form and the mechanisms for ensuring subjects' privacy.

Some consideration should be given to whether additional data on the subjects in the data set could or should be collected. Will it be possible, for instance, to contact subjects and re-interview them, in order to obtain new variables or extend the follow-up period? Could similar ends be pursued using medical records, large databases (e.g., medical insurance or census data), or other research data? Is an expanded data set affordable, practical, and ethically feasible? If other sources of information (databases) are to be consulted, or if subjects are to be recontacted in order to increase the number of variables, mechanisms will need to be put in place to ensure that linking of data will not compromise the subjects' rights. Regulatory bodies are growing increasingly concerned about violation of privacy and threats to well-being as a result of research. Once a study is completed, therefore, it is becoming difficult for researchers to obtain information on subjects beyond what is agreed to in the original consent form (McCarthy, Shatin, Drinkard, Kleinman, & Gardner, 1999). Committees will often insist that researchers seek further consent from subjects, but compromises are possible. Boruch and Cecil (1979) provide an extensive discussion of methods for ensuring confidentiality, such as using encrypted identification numbers and "information brokers" to interrupt the chain of connections between types of data and thus limit the linking of sensitive information to specific individuals.

After institutional ethics clearances have been obtained, attention shifts to preparation of data files for analysis. Floppy disks, CD-ROMs, or electronic mail attachments containing files with all the variables of interest are very convenient, but data may or may not come in the medium or format that the secondary analyst wishes to use. For instance, large data sets may be shared on data tapes for mainframe computers or may have been prepared using a different software

package (or a different version of the same package) from the one the analyst will be using. While excellent software for converting data-set files across different statistical packages is now available (e.g., Stat/Transfer), the conversion process could require the use of several computers and considerable technical assistance. Once the original data are in a format that permits analysis, further work may be necessary: if the data have been divided into several smaller files, these files will need to be merged using a subject number or code as the matching variable; and if additions to the data are planned, any additional variables will have to be prepared for inclusion — which could involve manual data entry.

The first step, after a complete data file has been prepared, is calculation of descriptive statistics for comparison with printouts and data dictionaries or reports of preliminary results from the primary study, in order to verify that the transfer across media has not resulted in data corruption. The secondary analyst may also wish to rerun analyses from the primary study to verify that the major variables have been correctly identified and that, at a minimum, the main results can be replicated. This provides a check on both the accuracy of the data-transfer process and the secondary analyst's understanding of the architecture of the data set and the system of variable labels.

Attention then turns to the mechanics of conducting the new statistical analyses. At this point, decisions about the handling of missing data need to be made. It may be possible to run analyses that leave out subjects with missing data (Allison, Gorman, & Primavera, 1993; Norman & Streiner, 1994), but the concerns about generalizability discussed in the previous section must be kept in mind. A rule of thumb in the methodological literature is that if 15% or more of the subjects in smaller studies are missing data for a variable, serious consideration should be given to dropping that variable from the analysis, unless certain assumptions can be met and acceptable methods of imputation can be found. Imputation is a possible solution when relatively small pockets of data are missing at random in a data set (which is not particularly common). This sometimes involves entering averages or extreme values for missing ones, but newer methods (for instance, multiple imputation) use multiple regression and other procedures to calculate probable values of missing data based on other available variables (Rubin, 1987, 1996). While such sophisticated statistical algorithms can sometimes be used to fill in higher proportions of missing data, consultation with an experienced statistician is usually necessary (Allison et al.; Hertel, 1976).

It will often be necessary to create new variables for analysis, especially if certain aspects of the original data were never fully explored. This may involve the construction of scales using previously established or new, analyst-generated scoring schemes, or it may involve the recoding of variables to provide contrasts or composite measures. Psychometric properties of all measures (both old and new) should also be independently assessed by the secondary analyst.

The details of specific analyses (univariate, multivariate, parametric, non-parametric, etc.) will obviously depend largely on the research questions and the levels of measurement of the individual variables. Analyses will often be highly exploratory, and multiple statistical tests may be performed to try to identify patterns in the variables. If the secondary analyst is not cautious, "fishing expeditions" capitalizing on chance may yield statistically significant results that reflect spurious or trivial differences or relationships among variables (Mills, 1993). Patterns of results, rather than individual statistical tests, should generally be used to determine whether or not hypotheses are supported.

In a secondary analysis, as in any study, sample size will influence the interpretation of statistical significance tests. Naturally, unless acceptable means of imputation can be implemented, statistical power is lost when many subjects have missing data. On the other hand, large sample sizes, often a key attraction of secondary data sets, will affect statistical significance in the other direction. The researcher must be careful to avoid mistaking very small, clinically meaningless but statistically significant effects for important findings (Deal & Anderson, 1995). Furthermore, if multiple tests are performed and the numbers of subjects represented in various analyses are uneven, variable statistical power across different tests must be considered when discussing patterns of significant and non-significant results.

The final phase of a secondary analysis, reporting the results, is in many ways similar to this phase of any research. In most cases, a brief description of the methodology used in the primary study will indicate the context of the original data collection. Referring the reader to publications from the original study is a good strategy for keeping this section brief. There must be a clear statement on the derivation of the final sample analyzed (i.e., the number of individuals screened in the original study, the number ultimately enrolled, the number completing follow-up, and finally the number meeting entry criteria for the secondary analysis and with data on the variables of interest). As discussed earlier, a comment in the Discussion about the limitations of the particular analysis that touch on the origins of the data will also be needed.

Conclusion: Contribution of Secondary Analysis to the Scientific Literature

By now it will be obvious that reduced expenditures of time and energy should not be the primary motivations for undertaking a secondary analysis. Experience will often show that obtaining a suitable data set, adapting it for use, and dealing with the many unexpected turns typical of secondary analysis is an extremely lengthy process. Before committing to a project, the analyst should review the appropriateness of a data set for the questions of interest. Missing data and subject attrition in a data set can temper enthusiasm for the broad ranges of variables on large numbers of subjects often available in secondary data sets. The secondary analyst must thoroughly understand the methods used in the original study, and must be especially vigilant about methodological issues with regard to sampling and measurement in order to accurately gauge the meaningfulness of his or her results.

Why, given all of these issues, would anyone choose secondary analysis? Put simply, for various reasons researchers often lack the resources (monetary and human) to assemble new databases that address their specific research questions. Secondary analysis becomes the methodology of choice when other researchers have collected data on similar variables and these older data sets might shed light concerning the new research questions.

The second half of the previous sentence is the critical one. As with all research appraisal, in reviewing the results of a secondary analysis the primary question is a simple one: Are the conclusions drawn by the investigators justified once the nature of the sample and the data analyzed are taken into account? (Abelson, 1995) Other questions include: Is the research grounded and presented in the context of relevant theory and empirical knowledge? Does the research add to existing knowledge on the phenomenon or issue? Is the research intellectually sterile, or "research for its own sake"? Has adequate attention been paid to potentially serious methodological issues? Have the researchers devised innovative solutions to such issues as missing data, instrumentation problems, or the absence of an appropriate control group (if relevant)? Some authors believe that the results of secondary analyses always require confirmation in future studies because of the extensive exploratory work involved, but in the quantitative paradigm *all* research results demand replication before being given credence.

Is secondary analysis as valid a form of scholarly inquiry as primary research? Given the premium placed on obtaining funding for

new data collections, the student or researcher embarking on a secondary analysis might encounter the view — particularly on the part of researchers who have spent years overcoming the many practical problems that arise in primary data collection — that this form of investigation deserves less recognition than primary research. The secondary analyst may well have to defend his or her choice of research to colleagues and, in the case of graduate students, faculty members who oversee and review their thesis or dissertation work.

Emotions aside, several things are clear. Secondary analysis extends the pool of knowledge to make maximal use of data that are often very expensive to collect. Secondary analysis forces researchers and the readers of their work to ask serious questions about the way in which conclusions are drawn from empirical data and are used to build knowledge in a discipline. The process involves considerable intellectual effort on the part of the secondary analyst, both in appropriately using statistical methods to answer clearly articulated, substantial (rather than trivial) research questions, and in situating the results in terms of both the research literature and the strengths and limitations of the data. Lastly, secondary analysis requires a considerable investment of time and expertise. It is unfair and unwise to generalize about the quality of such analysis. As with any research, judgements regarding methodological integrity and scholarly rigour in secondary analysis must always be made on a case-by-case basis.

References

Abelson, R.P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Allison, D.B., Gorman, B.S., & Primavera, L.H. (1993). Some of the most common questions asked of statistical consultants: Our favorite responses and recommended readings. *Genetic, Social, and General Psychology Monographs, 119*, 155–185.

Azar, B. (1999). Will you be forced to share your data? *APA Monitor, 30*(8). Available online: www.apa.org/monitor.

Boruch, R.F., & Cecil, J.S. (1979). *Assuring the confidentiality of social research data*. Philadelphia: University of Pennsylvania Press.

Boyle, G.J. (1985). Self-report measures of depression: Some psychometric considerations. *Journal of Clinical Psychology, 24*, 45–59.

Brink, P.J., & Wood, M.J. (1994). *Basic steps in planning nursing research: From question to proposal (4th ed.)*. Boston: Jones & Bartlett.

Bucher, H.C. (1994). Social support and prognosis following first myocardial infarction. *Journal of General Internal Medicine, 9*, 409–417.

Campbell, D.T., & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carney, R.M., Freedland, K.E., Rich, M.W., & Jaffe, A.S. (1995). Depression as a risk factor for cardiac events in established coronary heart disease: A review of possible mechanisms. *Annals of Behavioral Medicine, 17,* 142–149.

Clarke, S.P. (1998). *Psychosocial correlates of mortality, cardiac events, health care utilization, and quality of life in patients with left ventricular dysfunction.* Unpublished doctoral dissertation, McGill University, Montreal, Quebec.

Cohen, S. (1988). Psychosocial models of the role of social support in the etiology of physical disease. *Health Psychology, 7,* 269–297.

Deal, J.E., & Anderson, E.R. (1995). Reporting and interpreting results in family research. *Journal of Marriage and the Family, 57,* 1040–1048.

Estabrooks, C.A., & Romyn, D.M. (1995). Data sharing in nursing research: Advantages and challenges. *Canadian Journal of Nursing Research, 27*(1), 77–88.

Fairclough, D.L., & Gelber, R.D. (1996). Quality of life: Statistical issues and analysis. In B. Spilker (Ed.), *Quality of life and pharmacoeconomics in clinical trials (2nd ed.)* (pp. 427–435). Philadelphia: Lippincott-Raven.

Fortune, J.C., & McBee, J.K. (1984). Considerations and methodology for the preparation of data files. In D.J. Bowering (Ed.), *Secondary analysis of available databases* (New Directions for Program Evaluation Series, #22) (pp. 27–49). San Francisco: Jossey-Bass.

Frasure-Smith, N., Lespérance, F., & Talajic, M. (1993). Depression following myocardial infarction: Impact on 6-month survival. *Journal of the American Medical Association, 270,* 1819–1825.

Friedman, L.M., Furberg, C.D., & DeMets, D.L. (1985). *Fundamentals of clinical trials (2nd ed.).* Littleton, MA: PSG Publishing.

Hakim, C. (1982). *Secondary analysis in social research.* London: Allen & Unwin.

Hertel, B.R. (1976). Minimizing error variance introduced by missing data routines in survey analysis. *Sociological Methods and Research, 4,* 459–474.

Horwitz, R.I., & Horwitz, S.M. (1993). Adherence to treatment and health outcomes. *Archives of Internal Medicine, 153,* 1863–1868.

Hyman, H.H. (1972). *Secondary analysis of sample surveys: Principles, procedures and potentialities.* New York: Wiley.

Jacobson, A.F., Hamilton, P., & Galloway, J. (1993). Obtaining and evaluating data sets for secondary analysis in nursing research. *Western Journal of Nursing Research, 15*(4), 483–494.

Kasl, S.V. (1995). Strategies in research on health and aging: Looking beyond secondary data analysis. *Journals of Gerontology: Series B, Psychological Sciences & Social Sciences, 50*(4), S191–S193.

Krieger, N. (1992). Overcoming the absence of socioeconomic data in medical records: Validation and application of a census-based methodology. *American Journal of Public Health, 82*(5), 703–710.

Lange, L.L., & Jacox, A. (1993). Using large data bases in nursing and health policy research. *Journal of Professional Nursing, 9*(4), 204–211.

Mainous, A.G., & Hueston, W.J. (1997). Using other people's data: The ins and outs of secondary data analysis. *Family Medicine, 29*(8), 568–571.

McArt, E.W., & McDougal, L.W. (1985). Secondary data analysis — A new approach to nursing research. *Image: Journal of Nursing Scholarship, 17*(2), 54–57.

McCarthy, D.B., Shatin, D., Drinkard, C.R., Kleinman, J.H., & Gardner, J.S. (1999). Medical records and privacy: Empirical effects of legislation. *Health Services Research, 34*(1 Pt 2), 417–425.

McNair, D.M., Lorr, M., & Droppleman, L.F. (1981). *Profile of Mood States (POMS) manual.* San Diego: Educational and Industrial Testing Service.

Mills, J.L. (1993). Data torturing. *New England Journal of Medicine, 329*(16), 1196–1199.

Norman, G.R., & Streiner, D.L. (1994). *Biostatistics: The bare essentials.* St. Louis: Mosby.

Oxman, A.D., & Guyatt, G.H. (1992). A consumer's guide to subgroup analyses. *Annals of Internal Medicine, 116,* 78–84.

Pedhazur, E.J., & Schmelkin, L.P. (1991). *Measurement, design and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.

Polit, D.F., & Hungler, B.P. (1995). *Nursing research: Principles and methods (5th ed.).* Philadelphia: Lippincott.

Rogers, W.J., Johnstone, D.E., Yusuf, S., Weiner, D.H., Gallagher, P., Bittner, V.A., Ahn, S., Schron, E., Shumaker, S.A., & Sheffield, L.T. (1994). Quality of life among 5,025 patients with left ventricular dysfunction randomized between placebo and enalapril: The Studies of Left Ventricular Dysfunction. *Journal of the American College of Cardiology, 23,* 393–400.

Rubin, D.B. (1987). *Multiple imputation for non-response in surveys.* New York: Wiley.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*(434), 473–489.

Shaver, J.F. (1985). A biopsychosocial view of human health. *Nursing Outlook, 33,* 186–191.

Sieber, J.E. (1991). Social scientists' concerns about sharing data. In J.E. Sieber (Ed.), *Sharing social science data: Advantages and challenges* (pp. 141–150). Newbury Park, CA: Sage.

SOLVD Investigators. (1990). Studies of Left Ventricular Dysfunction (SOLVD) — Rationale, design, and methods: Two trials that evaluate the effect of enalapril on patients with reduced ejection fraction. *American Journal of Cardiology, 66,* 315–322.

SOLVD Investigators. (1991). Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *New England Journal of Medicine, 325,* 293–302.

Szabo, V., & Strang, V.R. (1997). Secondary analysis of qualitative data. *Advances in Nursing Science, 20*(2), 66–74.

Williams, R.B., Barefoot, J.C., Califf, R.M., Haney, T.L., Saunders, W.B., Pryor, D.B., Hlatky, M.A., Siegler, I.C., & Mark, D.B. (1992). Prognostic importance of social and economic resources among medically treated patients with angiographically documented coronary artery disease. *Journal of the American Medical Association, 267,* 520–524.

Wilson, H.S. (1989). *Research in nursing.* Redwood City, CA: Addison-Wesley.

Authors' Note

This work was funded in part by a doctoral training fellowship from the National Health Research and Development Program, Health Canada, at the School of Nursing, McGill University, from 1992 to 1995, and an institutional postdoctoral fellowship from the National Institute for Nursing Research, National Institutes of Health (T32 NR07104), at the University of Pennsylvania, both held by Sean Clarke.

The authors gratefully acknowledge the feedback of Beth McNutt and Pat Patrician on earlier drafts of this paper.

Correspondence may be directed to: Sean Clarke, Center for Health Outcomes and Policy Research, University of Pennsylvania School of Nursing, 420 Guardian Drive, Philadelphia, PA 19104-6096 USA. Telephone: 215-898-9669. E-mail: <sclarke@nursing.upenn.edu>.