REVIEW

THE AMERICAN JOURNAL of MEDICINE®

Commonly Used Data-collection Approaches in Clinical Research

Jane S. Saczynski, PhD,[a,b] David D. McManus, MD,[a,b] Robert J. Goldberg, PhD[b]

[a]Department of Medicine, University of Massachusetts Medical School, Worcester; [b]Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester.

ABSTRACT

We provide an overview of the different data-collection approaches that are commonly used in carrying out clinical, public health, and translational research. We discuss several of the factors that researchers need to consider in using data collected in questionnaire surveys, from proxy informants, through the review of medical records, and in the collection of biologic samples. We hope that the points raised in this overview will lead to the collection of rich and high-quality data in observational studies and randomized controlled trials.

 • The American Journal of Medicine (2013) 126, 946-950

KEYWORDS: Clinical research; Data-collection approaches; Observational studies

In a recent editorial, we described the different types of observational studies and randomized controlled trial designs that investigators often use in carrying out clinical and public health research.[1] Although 2 of the most important steps in successfully carrying out a research project are the clear formulation of key testable hypotheses and careful selection of a cost-efficient, rigorous study design, less information is available for researchers with respect to contemporary methods of high-quality and reliable data collection. With increasing attention being paid to patient-reported outcomes in observational, comparative effectiveness, and clinical trials research, data-collection approaches that combine medical record abstraction, patient interviews, and administrative data will be more commonly used in the future.

Funding: Funding support for this work was provided by the National Institutes of Health (RO1 HL35434). Partial salary support for JSS, DDM, and RJG was provided for by the National Institutes of Health Grant 1U01HL105268-01. JSS was supported in part by funding from the National Institute on Aging (K01 AG33643).

Conflict of Interest: None.

Authorship: All authors had access to the data and played a role in writing this manuscript.

Requests for reprints should be addressed to Robert J. Goldberg, PhD, Department of Quantitative Health Sciences, University of Massachusetts Medical School, 55 Lake Avenue North, Worcester, MA 01655.

E-mail address: Robert.Goldberg@umassmed.edu

0002-9343/$ -see front matter
http://dx.doi.org/10.1016/j.amjmed.2013.04.016

COLLECTING MEANINGFUL DATA IN A CLINICAL RESEARCH STUDY

In the present review, we discuss a number of issues that pertain to the collection of high-quality data in the conduct of clinical, translational, and epidemiologic research projects and ways to enhance the collection of reliable and meaningful data. We also discuss issues related to the accuracy of these data and factors to consider in the possible independent confirmation of information collected from different data sources. The data-collection instruments reviewed include questionnaire surveys and patient self-reported data, use of proxy/informant information, hospital and ambulatory medical records, and analysis of biologic materials.

QUESTIONNAIRE SURVEYS AND PATIENT-REPORTED DATA

Much of the information collected in observational epidemiologic studies is collected in the form of patient/participant self-reports on standardized questionnaires that are self- or interviewer-administered in person, by phone, or via mail or the Internet. The factors on which information is routinely collected in these studies include sociodemographic characteristics, lifestyle practices, medical history, and use of prescribed or over-the-counter medications. Questions also are often asked about participants'

knowledge and attitudes toward various lifestyle and disease predisposing factors. With increasing attention being paid to patient-reported outcomes by funding agencies, such as the National Institutes of Health, the Agency for Healthcare Research and Quality, and the newly formed Patient-Centered Outcomes Research Institute, measures of patient-centered factors, such as quality of life, depression, anxiety, cognitive, and functional status, are increasingly included in these surveys. The Consolidated Standards of Reporting Trials (CONSORT) Statement was updated recently to include standards for reporting patient-reported outcomes in randomized controlled trials, highlighting the increasing awareness of the inclusion of such measures as key outcomes of these rigorous investigations.[2]

CLINICAL SIGNIFICANCE

- This review provides readers with an overview of the issues involved in the collection of high-quality meaningful data in clinical, translational, and epidemiologic research studies.
- The data-collection instruments reviewed include questionnaire surveys, proxy/informant information, hospital and ambulatory care records, and analysis of biologic materials.

Ideally, patient-reported outcomes are measured using standardized, validated instruments to promote the collection of high-quality data and allow for meaningful comparisons across observational studies or randomized trials. Use of standardized assessments also facilitates pooling of data across studies with the goal of establishing clinically relevant cut-points or clinically meaningful change in important patient-related outcomes in response to a lifestyle intervention or medical treatment. Recent federally funded initiatives, such as the National Institutes of Health Toolbox (www.nihtoolbox.org) and Patient Reported Outcomes Measurement Information System (www.nihpromis.org), have highlighted the importance of harmonization of patient-reported outcomes data-collection instruments.

Typically, surveyed individuals are asked to respond to these questions in a yes/no manner, on a Likert-type scale (eg, very often to not at all often), or with open-ended responses. The choice of responses is dictated by the investigator and the standardized instrument (if one is used). The selection of the type of response desired is often made on the basis of the difficulty of the question asked and the depth of knowledge and level of precision the investigator would like to have about a particular factor.

Standardized instruments often have different forms that vary in length, so an investigator can decide whether a "long" (eg, Short Form-36)[3,4] or "short" (Short Form-12) version is best suited for a study. Tests with multiple length versions typically have published psychometric properties (eg, sensitivity and specificity of screening tests) that guide investigators in choosing a test version. For example, a consenting study participant might be asked a series of questions about their level of physical activity in the present or during a recent period of pertinent exposure. The number and depth of these questions would be determined in part by how this variable would be used in subsequent analyses and presented in peer-reviewed publications. If the factor of physical activity was to be simply used as a controlling variable in stratified or multivariable-adjusted regression analyses, then a briefer assessment of physical activity might be more acceptable with the added benefit of reduced respondent burden. On the other hand, if an investigator is particularly interested in the role of type of aerobic activity, level of exercise intensity, or duration of physical activity, then a more extensive battery of questions might be asked about this factor with objective validation of self-reported activity carried out or a standardized instrument used.

Although the use of validated, standardized instruments is preferred, these data-collection tools are not always available. If standardized instruments do not exist for a specific construct to be measured, investigators will often create "home-grown" scales. It is important to carefully design these home-grown instruments, ideally with the input of a psychometrician, and to pilot test all measures before using them in a formal research study. These pilot efforts ideally would involve validation of the instrument against a gold standard (eg, clinical diagnosis) or important study outcome. One needs to carefully balance the need for independent validation of participant responses, and the attendant costs and logistic issues associated with such, versus simply discussing the lack of validation of certain variables as a study limitation. These decisions should be discussed with a senior, experienced mentor who has been involved in observational clinical research studies or randomized trials for many years. The advantages and disadvantages of questionnaire data are summarized in **Table 1**.

PROXY/INFORMANT DATA

The collection of information about study participants through the use of proxy respondents can be one of the more challenging tasks for an investigator. Moreover, the accuracy/validity of the proxy's responses and his/her extent of knowledge about various health-related aspects of the study participant need to be thoughtfully considered in determining the type and quantity of information to be elicited from the proxy respondent. On the other hand, especially in observational studies in which the cases or controls in a retrospective study may have died or may not be capable of/competent to provide their own responses, information from proxies may be the only source of data available. In some situations, informant perspectives are important data elements, even if different from that of the patient. For instance, family member reports of the type and amount of assistance a patient requires with activities of daily living may be qualitatively different, but equally important, as that reported by the patient.

Table 1 Advantages and Disadvantages of Questionnaire Survey Data

Advantages	Disadvantages
Can collect personal or risk factor data not typically contained in hospital/ambulatory care records	Validating individual survey responses can be difficult, burdensome, costly, and of questionable utility
Can elicit information in an analytically desirable and standardized manner	If response rates are less than desirable, one may question the representativeness of the study sample and its generalizability
Can maintain high survey response rates through various financial or other incentives	Responses might differ if questions are asked in-person vs by phone vs by mail/Internet

Increasingly, informal caregivers are being recognized as "stakeholders" in many research studies, particularly those that focus on patient-reported outcomes, such as quality of life. In cases of questionable mental status or noncommunicative state of a patient, informants can be helpful and important in providing information to help establish a "baseline" for a patient. In these situations, informants can report on the patient's level of cognitive and physical function and of independence, which are important outcomes in many contemporary clinical research studies. For some domains, validated informant questionnaires exist. For instance, the Informant Questionnaire on Cognitive Decline in the Elderly[5] is an informant measure of cognitive function and informant responses on the Short Form-36 and Activities of Daily Living scales have been used as assessments of health-related quality of life and functional status with varying results.[6,7]

REVIEW OF AMBULATORY OR HOSPITAL MEDICAL RECORDS

Because of the ubiquity and abundance of high-quality data embedded within medical records, they are a commonly used source of information in clinical research studies. Information contained in hospital or ambulatory care records may be used as the sole source of data or complementary to other instruments used to elicit information. Decisions about the adequacy of using the medical record as the sole or main source of data for a given study hinges on the investigator's hypotheses, study sample size, budget and timeline, and extent and type of data available in a given record system. Medical records can be important sources of information that can reliably document participants' medical history, clinical, laboratory, or physiologic profile at varying time points in a cost-efficient manner. On the other hand, the data contained in medical records can be frustrating to use and, in some cases, conflicting or of questionable accuracy because of the nonstandardized manner in which this information is collected, recorded, or abstracted by various health care professionals and members of research teams. The increasing use of electronic medical records and their merger with administrative data have eased data abstraction efforts and, with increasing use of standardized data entry sets, reduced data heterogeneity.

One major limitation of using the medical record as a primary data source is that potentially important patient-reported information is often lacking, which is typically limited to the reporting of a "chief symptom" or symptoms directly related to the present symptom. If clinical information is stigmatized (eg, sexual history, alcohol or drug use) or difficult to assess systematically in primary care settings (eg, cognitive status, depression), it is often underreported in the medical record. It also is important to note that factors (eg, medication use) are defined by clinicians, not by trained study staff or study participants, and certain variables may not be accurately coded. Moreover, the extent of documentation about key medical history or clinical variables can vary widely between providers (including conflicting data) and health care systems. Heterogeneity can create considerable difficulties in the construction of key study variables or in their use.

For example, in studying a purported association between macular degeneration and a number of different dietary components, it would be important to document the presence of various medical history conditions that may affect an individual's dietary practices and the development of macular degeneration. In this example, we would be particularly interested in ascertaining the presence of a history of type 2 diabetes mellitus on the basis of information contained in medical records. Inasmuch, one needs to consider how this condition and related chronic medical conditions would be classified on the basis of information contained in medical records. For example, is diabetes considered present if there is a simple notation of this condition in the patient's medical history by a sole provider? On the other hand, might there be a need for the documentation of various key elements of each condition to be noted in the medical records (eg, multiple elevated serum glucose levels obtained under fasting conditions) before a diagnosis of diabetes can be accepted? For several relatively common conditions, such as heart failure and stroke, independently and extensively validated algorithms have been developed to ascertain the presence of these important chronic diseases.[8-10]

Depending on the major research questions under study, resources available, and amount of variability/precision willing to be accepted in documenting the presence (or equally important, the absence) of each of these comorbid conditions, rules of acceptance and rejection can be applied in the consideration of these factors. Likewise, the investigator also might decide to simply ask the survey participant whether or not diabetes had been ever diagnosed in his/her

Table 2 Advantages and Disadvantages of Hospital/Ambulatory Care Records

Advantages	Disadvantages
Readily available and contain much useful demographic and clinical information	Oftentimes data contained in medical records are nonstandardized and inconsistently collected and recorded
Can be linked to other follow-up information sources	Information is often incomplete or missing
Can be used to characterize the medical history and clinical course of hospitalized and outpatient individuals	Independent checks on validity or reliability are atypically performed
Can provide data on medication intensity and duration	Information on etiologic or prognostic factors of importance is often not obtained or asked about or recorded in a standardized manner

past. This should be a simple thing to do, but the investigator needs to have considered beforehand how he/she will analyze the data if personal responses are not consistent with their medical record findings. **Table 2** summarizes the advantages and disadvantages of using medical records.

COLLECTION OF BIOLOGIC MATERIAL

An increasing number and array of contemporary clinical and translational research investigations involve the collection of biologic samples from study participants. These include personal factors, such as hair, saliva, urine, and serum. Biologic samples are increasingly being used to profile participants' metabolic, proteomic, or genomic status and thereby better understand their underlying pathophysiology or their response to a treatment or disease. Although it is beyond the scope of the present article, the ethical implications of genetic research warrant special thought and consideration. Furthermore, various imaging modalities (eg, computed tomography or magnetic resonance imaging, nuclear scans, ultrasonography) are being used to obtain deeper insights into underlying anatomic, pathologic, and biologic mechanisms involved in the development of disease, its prognosis, or response to treatment and suggest areas of future research endeavor.

Despite the important information these biologic samples provide on disease, its various causes, and its natural history, there are a number of factors to consider in the collection of biologic materials (**Table 3**). One important factor to consider when obtaining biologic samples is the frequency of collection (often a balance between participant burden and pathophysiologic insights gained from the ability to assess change in a factor over time), timing of specimen collection (especially when this biologic variable has been shown to exhibit circadian variation), cost (both to the participant and to the investigator in terms of invasiveness and complexity, respectively), variability in test measurement (often presented as a coefficient of variation), and careful need for standardization of test methods and their interpretation (eg, referencing vs a gold standard).

For example, an investigator may be contemplating carrying out a prospective study of racial differences in serum biomarkers and echocardiographic determinants of atrial fibrillation. In addition to the collection of clinical and demographic historical information, a baseline echocardiogram and serum levels of various biomarkers, such as B-type natriuretic peptide, are to be assessed. Investigators need to balance the need for further information with regard to changes in each of these parameters leading to atrial fibrillation with participant burden.

On the basis of the current literature and existing clinical knowledge, the investigators in this study would need to know how much echocardiographic atrial size and B-type natriuretic peptides change over key periods of time in patients with or at risk for atrial fibrillation. These concerns need to be built into data-collection efforts and need for systematic assessment of serial changes in these factors. Depending on the degree of change in these parameters, this might entail the collection of serial echocardiograms every 2 years, every 4 years, or more often, such as every 3 months, depending on the extent of change in left atrial size that might predispose an individual to the development of atrial

Table 3 Advantages and Disadvantages of Biologic Data

Advantages	Disadvantages
May provide novel insights into underlying disease pathophysiologic processes	Need to be collected under standardized conditions with considerable attention to detail
Can serve as an important end point of relevance	Ongoing quality-control procedures needed
Can be linked to other sociodemographic, medical history, and clinical data to obtain insights into disease occurrence and prognosis	Need to consider impact of possible biologic circadian variation for purposes of timing and frequency of data-collection efforts
	May need collection of multiple measures at baseline to adequately profile subsequent changes

fibrillation. On the other hand, because there may be more volatility or change in the serum biomarkers being examined, more frequent blood assays may be required and balanced with a participant's willingness to return to the clinic and associated discomfort/burden. Inasmuch, compromises in the intensity of data-collection efforts need to be balanced with patient-related concerns and the importance of keeping high rates of retention in a long-term longitudinal study.

Another major consideration with respect to the use of biologic data is when such samples are obtained relative to the definition of key study variables and outcomes (eg, are they concurrent or separated by considerable time). The importance of timing of the collection of various descriptive or risk factors is illustrated by the following example. An investigator wants to perform metabolomic profiling to examine differences between hepatic and circulating levels of a certain factor. To obtain in vivo hepatic tissue samples, he/she performs the investigation using patients undergoing a hepatic biopsy and obtains blood samples in the presurgical holding area at the time of intravenous line placement to minimize participant inconvenience. However, this study could be undermined should the metabolomic profile of the liver be influenced by medications administered for procedural sedation, thereby confounding any comparisons between hepatic and circulating levels of factors of primary interest.

Storage of biologic samples and technical factors relating to their measurement also warrant special consideration when interpreting or performing studies involving biologic specimens.

CONCLUSIONS

There are a number of factors to consider in deciding which and how much data to collect in any clinical research investigation. Investigators often believe that "more is better" and that it is important to collect information on as many scientifically "interesting" factors as possible. This premise may be misguided and place an unnecessary burden on study participants, as well as lead to the collection of considerable data that would never be used, analyzed, or presented in a scientific publication.

It is often useful and time well spent to identify those data elements that are essential and those that are academically "interesting" but may not be considered central to the key study hypothesis; this will greatly assist in narrowing down one's study questions and collecting data in as timely and rigorous a manner as possible. Moreover, it helps to create a list of the 5 to 10 major articles that might result from one's proposed research study and create an analysis plan for each article. By doing so, you will be able to separate the "data wheat" from the "data chaff" and hone in on those questions of key relevance and data elements that comprise these variables.

One also needs to carefully think about the independent validation of any self-reported responses and how intrusive, costly, and potentially burdensome this process may be. Validation of one's data, although important, can be a tricky and cumbersome route to follow with its attendant logistic and staffing complexities.

References

1. Goldberg RJ, McManus DD, Allison J. Greater knowledge and appreciation of commonly used research study designs. *Am J Med*. 2013;126:169.e1-169.e8.
2. Calvert M, Blazeby J, Altman DG, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO Extension. *JAMA*. 2013;309:814-822.
3. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care*. 1992;30:473-483.
4. Ware JE Jr, Gandek B. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. *J Clin Epidemiol*. 1998;51:903-912.
5. Jorm AF, Korten AE. Assessment of cognitive decline in the elderly by informant interview. *Br J Psychiatry*. 1988;152:209-213.
6. Ostbye T, Tyas S, McDowell I, Koval J. Reported activities of daily living: agreement between elderly subjects with and without dementia and their caregivers. *Age Ageing*. 1997;26:99-106.
7. Andresen EM, Vahle VJ, Lollar D. Proxy reliability: health-related quality of life (HRQoL) measures for people with disability. *Qual Life Res*. 2001;10:609-619.
8. Carnahan RM. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative data: summary of findings and suggestions for future research. *Pharmacoepidemiol Drug Saf*. 2012;21(Suppl 1):90-99.
9. Saczynski JS, Andrade SE, Harrold LR, et al. A systematic review of validated methods for identifying heart failure using administrative data. *Pharmacoepidemiol Drug Saf*. 2012;21(Suppl 1):129-140.
10. Andrade SE, Harrold LR, Tjia J, et al. A systematic review of validated methods for identifying cerebrovascular accident or transient ischemic attack using administrative data. *Pharmacoepidemiol Drug Saf*. 2012;21(Suppl 1):100-128.