Validity and reliability of measurement instruments used in research

CAROLE L. KIMBERLIN AND ALMUT G. WINTERSTEIN

Purpose. Issues related to the validity and reliability of measurement instruments used in research are reviewed.

Summary. Key indicators of the quality of a measuring instrument are the reliability and validity of the measures. The process of developing and validating an instrument is in large part focused on reducing error in the measurement process. Reliability estimates evaluate the stability of measures, internal consistency of measurement instruments, and interrater reliability of instrument scores. Validity is the extent to which the interpretations of the results of a test are warranted, which depends on the particular use the test is intended to serve. The responsiveness of the measure to change is of interest in many of the applications in health care where improvement in outcomes as a result of treatment is a primary goal of research. Several issues may affect the accuracy of data collected, such as those related to self-report and secondary data sources. Self-report of patients or subjects is required for many of the measurements conducted in health care, but self-reports of behavior are particularly subject to problems with social desirability biases. Data that were originally gathered for a different purpose are often used to answer a research question, which can affect the applicability to the study at hand.

Conclusion. In health care and social science research, many of the variables of interest and outcomes that are important are abstract concepts known as theoretical constructs. Using tests or instruments that are valid and reliable to measure such constructs is a crucial component of research quality.

Index terms: Control, quality; Data collection; Errors; Methodology; Research

Am J Health-Syst Pharm. 2008; 65:2276-84

Measurement is the assigning of numbers to observations in order to quantify phenomena. In health care, many of these phenomena, such as quality of life, patient adherence, morbidity, and drug efficacy, are abstract concepts known as theoretical constructs. Measurement involves the operationalization of these constructs in defined variables and the development and application of instruments or tests to quantify these variables. For example, drug efficacy may be operationalized as the prevention or delay in onset of cardiovascular disease, and the related measurement instrument may ascertain data on the occurrence of cardiac events from patient medical records. This article focuses primarily on psychometric issues in the measurement of patient-reported outcomes. However, similar aspects of measurement quality apply to clinical and economic outcomes. Steps to improve the measures used in pharmacy research are also outlined.

Evaluating the quality of measures

Key indicators of the quality of a measuring instrument are the reliability and validity of the measures. In addition, the responsiveness of the measure to change is of interest in many health care applications where improvement in outcomes as a result of treatment is a primary goal of research. Data sources for measures used in pharmacy and medical care research often involve patient questionnaires or interviews. Measures using patient self-report include quality of life, satisfaction with care, adherence to therapeutic regimens, symptom experience, adverse drug effects, and response to therapy (e.g., pain control, sleep disturbance). In addition, measures can be developed

CAROLE L. KIMBERLIN, PH.D., is Professor; and ALMUT WINTERSTEIN, PH.D., is Associate Professor, Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, Gainesville.

Address correspondence to Dr. Kimberlin at the Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, P.O. Box 100496, Gainesville, FL 32610 (kimberlin@cop.ufl.edu).

The authors have declared no potential conflicts of interest.

 1079-2082/08/1201-2276$06.00.
DOI 10.2146/ajhp070364

The Research Fundamentals *section comprises a series of articles on important topics in pharmacy research. These include valid research design, appropriate data collection and analysis, application of research findings in practice, and publication of research results. Articles in this series have been solicited and reviewed by guest editors Lee Vermeulen, M.S., and Almut Winterstein, Ph.D.*

from patient information available in medical records, including ordered tests or medical examinations, and administrative claims. Some of the measures from these data sources are considered more objective (e.g., laboratory tests) because the reliability and validity of the measures are known, with the error margins and reporting of results meeting generally rigorous standards. However, most data sources involve a greater degree of subjectivity in judgment or other potential sources of error in measurement. In such cases, it is incumbent on the researcher to control for known sources of error and to report the reliability and validity of measurements used.

Reliability. According to classical test theory, any score obtained by a measuring instrument (the observed score) is composed of both the "true" score, which is unknown, and "error" in the measurement process.[1] The true score is essentially the score that a person would have received if the measurement were perfectly accurate. The process of developing and validating an instrument is in large part focused on reducing error in the measurement process. There are different means of estimating the reliability of any measure. According to Crocker and Algina,[1] the test developer has a responsibility to "identify the sources of measurement error that would be most detrimental to useful score interpretation and design a reliability study that permits such errors to occur so that their effects can be assessed." Pretesting or pilot testing an instrument allows for the identification of such sources. Refinement of the instrument then focuses on minimizing measurement error.

Reliability estimates are used to evaluate (1) the stability of measures administered at different times to the same individuals or using the same standard (test–retest reliability) or (2) the equivalence of sets of items from the same test (internal consistency) or of different observers scoring a behavior or event using the same instrument (interrater reliability). Reliability coefficients range from 0.00 to 1.00, with higher coefficients indicating higher levels of reliability.

Stability. Stability of measurement, or test–retest reliability, is determined by administering a test at two different points in time to the same individuals and determining the correlation or strength of association of the two sets of scores. The same process may be used when calibrating a medical measurement device, such as a scale. The timing of the second administration is critical when tests are administered repeatedly. Ideally, the interval between administrations should be long enough that values obtained from the second administration will not be affected by the previous measurement (e.g., a subject's memory of responses to the first administration of a knowledge tests, the clinical response to an invasive test procedure) but not so distant that learning or a change in health status could alter the way subjects respond during the second administration.

Internal consistency. Internal consistency gives an estimate of the equivalence of sets of items from the same test (e.g., a set of questions aimed at assessing quality of life or disease severity). The coefficient of internal consistency provides an estimate of the reliability of measurement and is based on the assumption that items measuring the same construct should correlate. Perhaps the most widely used method for estimating internal consistency reliability is Cronbach's alpha.[1-4] Cronbach's alpha is a function of the average intercorrelations of items and the number of items in the scale. It is used for summated scales such as quality-of-life instruments, activities of daily living scales, and the Mini Mental State Examination. All things being equal, the greater the number of items in a summated scale, the higher Cronbach's alpha tends to be, with the major gains being in additional items up to approximately 10, when the increase in reliability for each additional item levels off. This is one reason why the use of a single item to measure a construct is not optimal. Having multiple items to measure a construct aids in the determination of the reliability of measurement and, in general, improves the reliability or precision of the measurement.

Interrater reliability. Interrater reliability (also called interobserver agreement) establishes the equivalence of ratings obtained with an instrument when used by different observers. If a measurement process involves judgments or ratings by observers, a reliable measurement will require consistency between different raters. Interrater reliability requires completely independent ratings of the same event by more than one rater. No discussion or collaboration can occur when reliability is being tested. Reliability is determined by the correlation of the scores from two or more independent raters (for ratings on a continuum) or the coefficient of agreement of the judgments of the raters. For categorical variables, Cohen's[5] kappa is commonly used to determine the coefficient of agreement.[2] Kappa is used when two raters or observers classify events or observations into categories based on rating criteria. Rather than a simple percent agreement, kappa takes into

account the agreement that could be expected by chance alone.

Often, observational instruments or rating scales are developed to evaluate the behaviors of subjects who are being directly observed. However, any measure that relies on the judgments of raters or reviewers requires evidence that any independent, trained expert would come to the same conclusion. Thus, interrater reliability should be established when data are abstracted from medical charts or when diagnoses or assessments are made for research purposes. Interrater reliability in research such as this depends on developing precise operational definitions of variables being measured as well as having observers well trained to use the instrument. Interrater reliability is optimized when criteria are explicit and raters are trained to apply the criteria. Raters must be trained how to make a decision that an event has occurred or how to determine which point on a scale measuring strength or degree of a phenomenon (e.g., a 3-point scale measuring seriousness of a disease) should be applied. The more that individual judgment is involved in a rating, the more crucial it is that independent observers agree when applying the scoring criteria. Before data gathering begins, training should include multiple cases in which raters respond to simulated situations they will encounter and rate, interrater reliability is calculated, disagreements are clarified, and a criterion level of agreement is met. Interrater reliability should again be verified throughout the study. Even when established observational instruments are being used or criteria are explicit, research that relies on observations or judgments should check reliability, and the study protocol should include procedures to determine the level of observer agreement. In most studies, a percentage of observations (e.g., number of charts reviewed) is randomly selected for scoring by two independent raters rather than requiring that two raters judge all observations. In addition, data to establish the consistency with which the primary rater applies the criteria over time are important for establishing the reliability of the instrument. Rater drift can occur when an individual rater alters the way he or she applies the scoring criteria (i.e., becoming more lenient or stringent) over time. Investigators who build in reliability checks throughout the study as data are collected rather than waiting until the end of data collection can identify instances where interrater reliability has begun to deteriorate, perhaps due to rater drift.

Validity. Validity is often defined as the extent to which an instrument measures what it purports to measure. Validity requires that an instrument is reliable, but an instrument can be reliable without being valid. For example, a scale that is incorrectly calibrated may yield exactly the same, albeit inaccurate, weight values. A multiple-choice test intended to evaluate the counseling skills of pharmacy students may yield reliable scores, but it may actually evaluate drug knowledge rather than the ability to communicate effectively with patients in making a recommendation. While we speak of the validity of a test or instrument, validity is not a property of the test itself. Instead, validity is the extent to which the interpretations of the results of a test are warranted, which depend on the test's intended use (i.e., measurement of the underlying construct).

Much of the research conducted in health care involves quantifying attributes that cannot be measured directly. Instead, hypothetical or abstract concepts (constructs), such as severity of disease, drug efficacy, drug safety, burden of illness, patient satisfaction, health literacy, quality of life, quality of provider–patient communication, and adherence to medical regimens, are measured. Hypothetical constructs cannot be measured directly and can only be inferred from observations of specified behaviors or phenomena that are thought to be indicators of the presence of the construct.[1] Measurement of a construct requires that the conceptual definition be translated into an operational definition. An operational definition of a construct links the conceptual or theoretical definition to more concrete indicators that have numbers applied to signify the "amount" of the construct. The ability to operationally define and quantify a construct is the core of measurement.

To understand how a construct might be operationally defined, consider the example of the efficacy of a new drug product. The ability to improve a patient's health may be measured by the decrease of certain symptoms, the delay in onset of a certain disease, length of remission, or the prevention of certain clinical complications. Likewise, the theoretical construct of medication adherence may be operationally defined as a one-month recording of number of missed doses as measured by a medication-event monitoring system (MEMS), which includes microprocessors that record the occurrence and time of each opening of a prescription vial. An operational definition of patient satisfaction with health care might be "patient self-reported responses to items on the 18-item short-form version of the Patient Satisfaction Questionnaire (PSQ)."[6] An even more precise understanding of the operational definition would involve an examination of the specific items on the PSQ-18 instrument. How critical a concise operationalization, including data sources and aggregation of information, is in terms of measurement validity is illustrated with a simple outcome, such as onset of diabetes mellitus. A drug's ability to delay onset could be measured through simple chart review, but diagnosis of diabetes will depend on a patient's

decision to seek health care and the provider's ability to recognize symptoms and make the proper diagnosis. Thus, regularly scheduled follow-up visits and the use of explicit screening protocols will likely increase the accuracy of the estimate and yield a more valid result.

In addition, Crocker and Algina[1] have pointed to the importance of a theoretical foundation by noting that "constructs cannot be defined only in terms of operational definitions but must also have demonstrated relationships to other constructs or observable phenomena." New research that gathers information on the constructs measured by a specific instrument, even one that has been widely used in research, contributes to the evidence regarding the construct validity of that test. In this sense, all of the different studies and validation strategies that provide evidence of a test's validity for making specific inferences about groups of respondents are part of construct validation. Validity evidence is built over time, with validations occurring in a variety of populations. Comprehensive literature reviews on measurement approaches are therefore critical in guiding the selection of measures and measurement instruments.

Construct validity. This type of validity is a judgment based on the accumulation of evidence from numerous studies using a specific measuring instrument. Evaluation of construct validity requires examining the relationship of the measure being evaluated with variables known to be related or theoretically related to the construct measured by the instrument.[1,7] For example, a measure of quality of life would be expected to result in lower scores for chronically ill patients than for healthy college students. Correlations that fit the expected pattern contribute evidence of construct validity. All evidence of validity, including content- and criterion-related validity, contributes to the evidence of construct validity.

Content validity. This type of validity addresses how well the items developed to operationalize a construct provide an adequate and representative sample of all the items that might measure the construct of interest. Because there is no statistical test to determine whether a measure adequately covers a content area or adequately represents a construct, content validity usually depends on the judgment of experts in the field.

Criterion-related validity. This type of validity provides evidence about how well scores on the new measure correlate with other measures of the same construct or very similar underlying constructs that theoretically should be related. It is crucial that these criterion measures are valid themselves. With one type of criterion-related validity—predictive validity—the criterion measurement is obtained at some time after the administration of the test, and the ability of the test to accurately predict the criterion is evaluated. For example, surrogate outcomes such as blood pressure and cholesterol levels are based on their predictive validity in projecting the risk of cardiovascular disease, even though some of these associations have been recently questioned. Another type of criterion-related validity is concurrent validity. In establishing concurrent validity, scores on an instrument are correlated with scores on another (criterion) measure of the same construct or a highly related construct that is measured concurrently in the same subjects. Ideally, the criterion measure would be considered to be the gold standard measure of the construct. This strategy of determining the validity of a measure might be seen in a situation in which a new instrument has some advantage over the gold standard measure, such as an increased ease of use or reduced time or expense of administration. These advantages would justify the time and effort involved in the development and validation of a new instrument. An example of such a situation is a researcher developing a self-administered version of an instrument that had been validated for person-to-person interviewer administration. Another example is a clinical researcher wanting to use a brief screening instrument for a condition, such as depression, instead of administering a more extensive measure. Investigators in one study, for example, examined the validity of a single-item question "Do you often feel sad or depressed?" against a more extensive, validated instrument for identifying depression after a stroke.[8] The same approach applies to sources of diagnostic data. For example, researchers may want to determine the validity of using administrative claims data to measure a construct represented by a certain clinical event, such as hospitalization for acute myocardial infarction, rather than using chart reviews, which are time-consuming and costly.

Selecting an appropriate and meaningful criterion measure can be a challenge. Often, the ultimate criterion a researcher would like to be able to predict is too distant in time or too costly to measure. The "criterion problem" exists for many of the ultimate criterion measures investigators would like to predict in health care research. For example, a study that aims to evaluate the effect of pharmaceutical care on the "health" of hypertensive patients will likely not have the necessary follow-up time to establish that the intervention results in reduced morbidity or mortality. Instead, a surrogate outcome, such as reduction in blood pressure, is used. Cost of administration of the "best" criterion measures may also be a barrier. For example, an investigator may want to validate a new self-report measure of medication adherence with concurrent measurement using a MEMS cap. However, because MEMS technology is expensive, a less costly measure, such as pill count or refill records, may instead be used to provide evidence of concurrent validity.

Responsiveness

Responsiveness is the ability of a measure to detect change over time in the construct of interest. For outcome measures intended to evaluate the effects of medical or educational interventions, responsiveness to changes that result from the intervention is required. Reliability is a crucial component of responsiveness. The "noise" that is due to measurement error can mask changes that may, in fact, be attributable to the intervention. For example, using a scale manufactured to weigh trucks will not be helpful when evaluating a new weight-loss drug in humans because the estimates will be too imprecise to identify small changes. The measurement will be valid yet unreliable or imprecise. A new disease-specific quality-of-life instrument that has not demonstrated stability over time when there is no change in health status (which may be an indication of measurement error) may not be able to detect health status changes. Measures that have ceiling effects have a limited ability to assess positive changes that may result from the intervention because there is limited room for subjects to improve their scores. Responsiveness to change can legitimately differ from one population to another, which is why the measure must be appropriate to the subjects being studied. For example, a measure of activities of daily living that includes the ability to dress or wash oneself may be responsive to change among an elderly population of patients undergoing physical therapy or cardiac rehabilitation. However, it would probably not be sensitive to change due to a ceiling effect among a younger group of newly diagnosed hypertensive patients who have not experienced significant disability due to the disease or to the aging process.

Selecting an existing instrument

Before developing a new test or measure, an investigator should identify existing instruments that measure the construct of interest. Using an existing instrument that has substantial evidence of reliability and validity in a variety of populations is more cost-effective than starting from scratch to develop and validate an instrument.

In selecting an instrument, the following questions should be addressed:

1. Do instruments already exist that measure a construct the same or very similar to the one you wish to measure? Before you begin searching for instruments, you must have a clearly defined construct or concept that you wish to measure, along with an operational definition and some evidence that the construct can be measured as defined. For example, there is agreement that the efficacy of a new blood-pressure-lowering medication is ultimately defined by a reduction in macrovascular events, but what about the efficacy of a palliative agent for cancer patients? A literature search can help identify how other researchers have defined the construct or a closely related construct. The literature search will ideally result in a list of outcomes and instruments that you can evaluate for possible use in your research.
2. How well do the constructs in the instruments you have identified match the construct you have conceptually defined for your study? In evaluating whether there is congruence, do not rely on the title of the measure or on the operational definition of the construct that appears in a research article or the description of variables in a secondary database, such as a medical record or administrative claims database. Real understanding of the measure usually requires an examination of the actual items or questions and the way data were generated or documented. For example, reviewing the actual items used in a questionnaire to evaluate disease-specific quality of life will provide a better understanding of what aspect or conceptualization of quality of life is addressed. Talking with physicians about their progress notes will aid in deciding whether certain patient information can be expected to be documented in a patient chart or what is often omitted.
3. Is the evidence of reliability and validity well established? Has the measure been evaluated using various types of reliability estimates (e.g., both internal consistency and test–retest) and varied strategies for establishing validity (e.g., content and concurrent validity as well as more extensive evidence of construct validity in varied populations)? Has it been validated in a population similar to the one you will be studying?
4. In previous research, was there variability in scores with no floor or ceiling effects? Did previous studies have a large amount of missing data, either on the measure itself or on items within the measure?
5. If the measure is to be used to evaluate health outcomes, effects of interventions, or changes over time, are there studies that establish the instrument's responsiveness to change in the construct of interest? Obviously, it is important that change in measurement be due to change in the construct rather than to the instability of scores (i.e., lack of reliability of the measure itself). In addition, it would be helpful if there were data on how much change in scores would be required to be considered clinically meaningful.
6. Is the instrument in the public domain? If not, it will be necessary to obtain permission from the author for its use. Even though an instrument is published in the scientific literature, this does not automatically mean that it is in the public domain, and permission from the author and publisher may be required. If it is a copyrighted instrument, you may have to pay a fee to purchase or use the instrument. Some instruments may also require additional fees for scoring.

7. How expensive is it to use the instrument? A mail questionnaire costs less to administer than do telephone or face-to-face interviews. Using electronic data is usually less costly and time-consuming than conducting medical record reviews. However, electronic data may not contain information that is available on patient charts, so a thorough understanding of the limitations of the data available as well as the requirements of measurement for your study is important.
8. If the instrument is administered by an interviewer or if the measure requires use of judges or experts, how much expertise or specific training is required to administer the instrument?
9. Will the instrument be acceptable to subjects? Does the test require invasive procedures? Is the reading level appropriate? Is the respondent's burden, including complexity of questions and time needed to complete the instrument, unlikely to affect response rates or the quality of responses?

Keep in mind that reliability and validity evidence from established instruments is applicable only if you use the instrument in the same form and follow the same administration procedures as used in the validation study. Modifications of validated instruments may require permission from developers and also require validating the modified instrument as if it were a new instrument.

Researchers may be tempted to conclude that available measures do not meet their needs and that they must develop their own instruments. They may view the measures they want to develop as being so straightforward, such as a few questions measuring patient knowledge or a specific item from a medical chart, that they do not need to conduct a pilot test to determine reliability and validity. Researchers may then go to considerable effort collecting data only to find at the end of the study that subjects do not vary much in their responses to the instrument or that documentation in the chart was inadequate, so the measure was not able to correlate with any other variable of interest. Subjects may misinterpret questions. Responses may be highly skewed. Internal consistency may be so low that item responses cannot reasonably be combined into a single summated score. In other types of studies, a researcher may obtain biased results by incorrectly assuming that diagnostic codes are valid without determining their relationship to other measures that should indicate the presence of the disease. Assuming medical records adequately capture the information needed to construct a measure and that chart reviewers will interpret information uniformly can also threaten the validity of findings. Careful attention to the development of instruments, regardless of how straightforward the measures may seem, along with pilot testing to determine their reliability and validity, is crucial to the conduct of quality research.

Item-response theory

In recent years, Rasch models and item-response theory (IRT) or latent-trait models have provided an alternative framework for understanding measurement and alternative strategies for judging the quality of a measuring instrument. Readers are referred to other resources for more information on Rasch and IRT models.[1,9-11]

The National Institutes of Health, along with research teams throughout the United States, initiated the development of the Patient-Reported Outcomes Measurement Information System, which will create item banks of patient-reported outcomes validated using modern measurement theory.[12] This initiative is building item pools and developing questionnaires that measure key health outcomes related to many chronic diseases, including measures such as fatigue and pain. These items will be available to investigators, and the repository will become a resource for "accurate and efficient measurement of patient-reported symptoms and other health outcomes in clinical practice."[12]

Measurements using self-report

For many of the measurements used in health care, researchers rely on the self-report of patients or subjects. With surveys, researchers rely on responses to questions to provide measurements of the constructs of interest. While self-reports of behavior, beliefs, and attitudes are prone to known biases, there are no acceptable alternative means of measurement for many constructs (e.g., level of pain, depression, patient satisfaction with care, quality of life).

Self-reports of behavior such as dietary intake, adherence to medication regimens, and exercise frequency and intensity are particularly subject to problems with social desirability biases. Subjects may provide responses that are socially acceptable or that are in line with the impression they want to create. In addition, self-report questions may elicit an estimation of behavioral frequency rather than the recall and count response desired by the researcher. The use of estimation rather than recall is a function of how information is retrieved from memory, how frequency-response scales are formulated, and other specific aspects of the instrument.[13-15] For example, behaviors that occur with high frequency, such as dietary intake or taking a scheduled medication for a chronic condition, are not likely to be specific in memory for a very long period of time. If it is desired that specific events be recalled rather than estimated, the time frame must be of very short duration and in the immediate past. Therefore, asking patients how many doses of a medication they missed in the past month or past year will likely result in an estimate or educated guess, whereas a question about the past 24 hours or past three days may reflect actual recall. Asking

subjects about stressors they encountered in the past 24 hours is likely to lead to recall of minor, daily hassles, whereas a question about stressors in the past year is likely to lead the subjects to interpret the question as being about major life events and answer accordingly. When a list of alternative responses is provided, the response options themselves determine the way subjects interpret the question and the way they respond.

Often, the response choices require subjects to provide their own judgment about frequency using undefined response alternatives (e.g., on an ordinal scale from "seldom" to "frequently"). Such terms can mean very different things to different subjects. One person who reports ingesting a "moderate" amount of alcohol may be referring to two to three alcoholic drinks a day, while someone else may define moderate consumption as two to three drinks a month. When asking questions about frequency of behavior, it is usually best to let the subject fill in the blank on an item with a clearly defined reference period. An example of such a question is "How many doses of (specific medication) have you missed taking completely in the past three days?" The open format requires a specific description of the behavior of interest as well as a specific time frame.

The International Epidemiological Association's European Questionnaire Group issued a report on problems arising from the questionnaires used to collect information on exposure, outcomes, and confounders.[16] The report noted that

> Published results often fail to reproduce the exact wording of key questions used to define exposures or outcomes, nor do they always provide adequate information on how the data collection instruments were developed, or if procedures such as pre-testing, validity checks, or pilot studies were used to ensure accuracy.

Use of self-report or poorly designed measures can result in misclassification bias (error in classifying either exposure status or effect [e.g., disease] in patients or subjects). Patient recall of previous drug exposure, for example, has been shown to be subject to error.[17-19]

In case–control studies, recall bias is of concern when there are no objective markers of exposure. Individuals with the disease or outcome of interest are more likely to remember relevant exposures than are healthy controls.[20] One approach that is recommended to address this recall bias is to have a control group affected by a disease different from that of cases to introduce a similar bias toward recall of exposure.

Use of secondary data

Data originally gathered for a different purpose are often used to answer a research question. These data may have addressed a different research question or may have been gathered for clinical, billing, or legal purposes. Secondary data include pharmacy records, electronic or paper medical records, patient registries, and insurance claims data. The first consideration when deciding whether secondary data can be used is to verify that the data set appropriately measures the variables required to answer the research questions. If the data elements are not present, consideration can be given to whether appropriate proxy measures of variables of interest are available. The use of proxy measures requires careful conceptual analysis of how closely the variables of interest and proxy measures are associated. For example, it seems intuitive that a claims database could be used to identify all patients who suffered a stroke during a certain time period as long as they were eligible for benefits. However, strokes may have been silent and required no medical intervention, patients may have died before medical care could be sought, stroke may have been misdiagnosed, or certain medical services may not have been covered by the insurance company and thus may not appear in the billing database. Understanding how the information represented in the data set was generated, whether and how it was coded, who coded it and for what purpose, and how consistent coding was across sites and at different times in longitudinal data sets or among different coders is important in evaluating the reliability of data. Utilization of diagnostic codes in charge data of clinical encounters has frequently been criticized because the selection of codes is often driven by reimbursement rather than clinical accuracy. Examining prior research that has applied these data sets can help determine what is known about the reliability and validity of data.

Even when original medical charts are used, it must be recognized that this information was not collected for research purposes and that documentation was guided by institutional policy, provider training, and provider preference. Moreover, while retrospective chart review is often used as the gold standard for validation of other measures, chart review is itself vulnerable to problems of unreliability, even though evidence of the reliability of data abstracted from charts is frequently not reported in research articles. A review of research in emergency medicine journals found that, of 244 articles utilizing chart review for data abstraction, interrater reliability was mentioned in 5% and statistically tested in only 0.4% of the articles.[21] The authors of the review also reported that additional steps to ensure the reliability and validity of chart review data (e.g., use of a standardized abstraction form, abstractor training, abstractor monitoring, blinding of abstractors to study hypotheses) were not mentioned in the study methods. Blinding to study hypotheses was mentioned in only 3% of studies, even though observer

bias is a recognized source of inaccurate study results. Other research has found that certain types of data elements abstracted from charts do not have adequate levels of interrater reliability.[22] Researchers interested in extracting data from medical charts are referred to articles describing procedures that can help ensure the quality of data abstracted.[23,24]

Use of surrogate measures

The Food and Drug Administration defines a surrogate endpoint of a clinical trial as "a laboratory measurement or a physical sign used as a substitute for a clinically meaningful endpoint that directly measures how a patient feels, functions, or survives. Changes induced by a therapy on a surrogate outcome are expected to reflect changes in a clinically meaningful endpoint."[25] The use of surrogate outcomes to operationally define a construct, such as drug efficacy, has become increasingly popular, as application of these measures is typically faster and less costly. Results are obtained after shorter follow-up periods, and the number of patients and length of time patients have to participate in experiments are reduced. For a surrogate outcome to be valid, it should be in the direct pathophysiologic pathway of a disease, and it should be reasonable to expect that the pharmacologic action of the new drug is mediated through this pathway. If these two conditions are true, the drug effect on the surrogate outcome can be extrapolated toward "true" measures of morbidity or mortality. However, even well-established surrogate outcomes have recently been questioned.[26] For example, the Heart and Estrogen/Progestin Replacement Study found that the demonstrated improvement in low-density-lipoprotein (LDL) and high-density-lipoprotein cholesterol levels did not result in an expected improvement on cardiac events.[27] Most recently, the Effect of Combination Ezetimibe and High-Dose Simvastatin vs Simvastatin Alone on the Atherosclerotic Process in Subjects with Heterozygous Familial Hypercholesterolemia (ENHANCE) trial found negative results for effects on intima–media thickness, even though the combination of ezetimibe and simvastatin demonstrated improved effects on LDL cholesterol levels as well as C-reactive protein.[28] These examples have alerted the research community that surrogate outcomes remain nothing more than substitutes and can only approximate the truth. There are few surrogate outcomes with superior scientific acceptance of validity than LDL cholesterol, which should caution us about the use and interpretation of research findings. Following this train of thought and recalling the previous discussion of the operationalization of theoretical constructs, it could be argued that all measures only approximate the truth. Invalid or unreliable measures can harm a study to the same extent as a poor study design or inadequate sample size.

Conclusion

In health care and social science research, many of the variables of interest and outcomes that are important are abstract concepts known as theoretical constructs. Using tests or instruments that are valid and reliable to measure such constructs is a crucial component of research quality.

References

1. Crocker L, Algina J. Introduction to classical and modern test theory. Orlando, FL: Harcourt Brace Jovanovich; 1986:1-527.
2. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994:251.
3. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika.* 1951; 16:297-334.
4. DeVellis RF. Classical test theory. *Med Care.* 2006; 44(11, suppl 3):S50-9.
5. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960; 20:37-46.
6. Marshall GN, Hays RD. The Patient Satisfaction Questionnaire Short-Form (PSQ-18). Santa Monica, CA: RAND; 1994.
7. Cambell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull.* 1959; 56:81-105.
8. Watkins C, Daniels L, Jack C et al. Accuracy of a single question in screening for depression in a cohort of patients after stroke: comparative study. *BMJ.* 2001; 17:1159.
9. Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. Mahwah, NJ: Lawrence Erlbaum; 2001:1-288.
10. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park, CA: Sage; 1991:1-153.
11. Smith EV, Smith RM. Introduction to Rasch measurement. Maple Grove, MN: JAM; 2004.
12. National Institutes of Health. PROMIS: Patient-Reported Outcomes Measurement Information System. www.nihpromis.org/default.aspx (accessed 2008 Jun 2).
13. Schwarz N. Self-reports: how the questions shape the answers. *Am Psychol.* 1999; 54:93-105.
14. Schwarz N, Oyserman D. Asking questions about behavior: cognition, communication and questionnaire construction. *Am J Eval.* 2001; 22:127-60.
15. Sudman S, Bradburn N, Schwarz N. Thinking about answers: the application of cognitive processes to survey methodology. San Francisco: Jossey-Bass; 1996:1-304.
16. Olsen J. Epidemiology deserves better questionnaires. *Int J Epidemiol.* 1998; 27: 935.
17. Beiderbeck AB, Sturkenboom MC, Coebergh JW et al. Misclassification of exposure is high when interview data on drug use are used as a proxy measure of chronic drug use during follow-up. *J Clin Epidemiol.* 2004; 57:973-7.
18. Ray WA, Thapa PB, Gideon P. Misclassification of current benzodiazepine exposure by use of a single baseline measurement and its effects upon studies of injuries. *Pharmacoepidemiol Drug Saf.* 2002; 11:663-9.
19. Korthuis PT, Asch S, Mancewicz M et al. Measuring medication: do interviews agree with medical record and pharmacy data? *Med Care.* 2002; 40:1270-82.
20. Tripepi G, Jager KJ, Dekker FW et al. Bias in clinical research. *Kidney Int.* 2008; 73:148-53.
21. Gilbert EH, Lowenstein SR, Koziol-McLain J et al. Chart reviews in emergency medicine research: where are the methods? *Ann Emerg Med.* 1996; 27: 305-8.
22. Yawn BP, Wollan P. Interrater reliability: completing the methods description in medical records review studies. *Am J Epidemiol.* 2005; 161:974-7.
23. Reisch LM, Fosse JS, Beverly K et al. Training, quality assurance, and assess-

ment of medical record abstraction in a multisite study. *Am J Epidemiol.* 2003; 157:546-51.
24. Gearing RE, Mian IA, Barber J et al. A methodology for conducting retrospective chart review research in child and adolescent psychiatry. *J Can Acad Child Adolesc Psychiatry.* 2006; 15:126-34.
25. Temple RJ. A regulatory authority's opinion about surrogate endpoints. In: Nimmo WS, Tucker GT, eds. Clinical measurement in drug evaluation. New York: Wiley; 1995:1-22.
26. D'Agostino RB Jr. The slippery slope of surrogate outcomes. *Curr Control Trials Cardiovasc Med.* 2000; 1:76-8.
27. Hulley S, Grady D, Bush T et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA.* 1998; 280:605-13.
28. Katelein JJ, Akdim F, Stroes ES et al. Simvastatin with or without ezetimibe in familial hypercholesterolemia. *N Engl J Med.* 2008; 358:1431-43. [Erratum, *N Engl J Med.* 2008; 358:1977.]