The fundamentals of quantitative measurement

The main purpose of the EBN Notebook is to equip readers with the necessary skills to critically appraise primary research studies and to provide a more detailed description of some of the methodological issues that arise in the papers we abstract. In the July 1999 issue of *Evidence-Based Nursing*, the EBN Notebook explored the concept of sampling.[1] In this issue we will provide a basic introduction to quantitative measurement of health outcomes, which may be assessed in studies of treatment, causation, prognosis, diagnosis, and in economic evaluations. Examples of health related outcomes are blood pressure, quality of life, patient satisfaction, and costs.

Health can be measured in many different ways; the various aspects of health that can be measured are referred to as *variables*.[2] For example, in the treatment study by Dunn *et al* in this issue of *Evidence-Based Nursing* (p 117), the interventions (known as the *independent* variables) were lifestyle and structured exercise programmes and the outcomes (known as the *dependent* variables) were physical activity and cardiorespiratory fitness. In a treatment study, the independent variables are those that are under the control of the investigator, and the dependent variables are the outcomes that may be influenced by the independent variable. In a causation study, the investigator relies on natural variation between both variables and looks for a relation between the 2 variables. For example, when determining whether smoking causes lung cancer, smoking is the independent variable and lung cancer is the dependent variable. In the abstracts included in *Evidence-Based Nursing*, the independent variables are identified under the "intervention" section for treatment studies and under the "assessment of risk factors" section for causation studies. The dependent variables are identified under the "main outcome measures" section.

Types of variables

Variables can be classified as nominal, ordinal, interval, or ratio variables. *Nominal (categorical) variables* are simply names of categories. Some nominal variables (referred to as *dichotomous* variables) have only 2 possible values, such as sex (men or women), survival (dead or alive), or whether a specific feature is present or absent (eg, diabetes or no diabetes); others may have several possible values, such as race (white, black, Hispanic, and others). The actual number of categories can be determined by the researcher; for example, race can be defined as 2 options (black or non-black) or by several possible options. No hierarchy is presumed with nominal data—that is, being alive is not twice as good as being dead (although most patients would argue with us about that one). In contrast, *ordinal variables* are sets of "ordered" categories.[2] For example, patients are often asked to rate the severity of their pain on a scale of 0–10, where 0 is no pain and 10 is unbearable, excruciating pain. Although we can safely say that a pain rating of 8 is worse than a pain rating of 5, we do not really know how much these 2 ratings differ because we do not know the size of the intervals between each rating.[2] Ordinal scales have also been used to grade pressure sore severity and to classify the staging of various cancers (eg, stage I, II, or III). *Interval variables* consist of an ordered set of categories, with the additional requirement that the categories form a series of intervals that are all exactly the same size. Thus, the difference between a temperature of 37°C and 38°C is 1 degree, and between 38°C and 39°C is 1 degree, and so on. However, an interval scale does not have an absolute zero point that indicates complete absence of the variable being measured. Because there is no absolute zero point on an interval scale, ratios of values are not meaningful—that is, 2 values cannot be compared by claiming that one is "twice as large" as another. A *ratio variable* has all the features of an interval variable but adds an absolute zero point, which indicates none (complete absence) of the variable being measured. The advantage of an absolute zero is that ratios of numbers on the scale reflect ratios of magnitude for the variable being measured.[3] To illustrate, 100°C is not twice as hot as 50°C (interval data) but 100 cm is twice as long as 50 cm, and a pulse of 80 beats per minute is twice a pulse rate of 40 beats per minute (ratio data).

The reason for stressing the differences between types of variables is that the type of variable dictates, to a large extent, the method of statistical analysis used by the researcher. It is meaningless to discuss the average or mean value of nominal or ordinal data because they are categories. Thus, the notion of a "mean" or "average" sex or race makes no sense; counts or frequencies of the number of individuals in each category, however, are useful. Conversely, the mean blood pressure (ratio variable) in a sample of patients is more meaningful than a count of the number of patients with each blood pressure measurement. In a subsequent EBN Notebook, we will explore different ways to describe and analyse data.

Issues in measurement

It is important to remember that most measurements in healthcare research encapsulate several things: the "real" or true value of the variable that is being measured; the variability of the measure; the accuracy of the instrument with which we are measuring; and perhaps the position of the patient or the skill and expectations of the person doing the measurement. Some of these elements are within the control of the measurer (eg, ensuring that a scale is at 0 before we weigh someone), whereas other elements are not (eg, a patient's blood pressure varies by time of day; therefore researchers try to assess blood pressure at the same time each day).

Some measures are more *objective* than others and are less likely to be influenced by human error or bias. Examples of objective measures include all cause mortality (ie, whether one is "dead" or "alive") and serum cholesterol concentrations. In contrast, *subjective measures* may be influenced by the perception of the individual doing the measurement (eg, patient self reported pain ratings). Most paper and pencil type questionnaires are subjective measures. The Beck Depression Inventory for Primary Care described in the

diagnosis study by Steer *et al* in this issue (p 126) is an example of a subjective paper and pencil questionnaire.

Frequency counts, such as incidence or prevalence, are often used when we want to know the extent of a disease or condition in a population. Others may be more interested in the beneficial and harmful effects of an intervention, such as differences in the rates of sexually transmitted diseases after a behavioural intervention provided to minority women (see the treatment study by Shain *et al* p 121).

What measurement issues should I look for when reading an article?

ARE THE MEASURES RELIABLE AND VALID?

These are 2 critically important properties of measurement. *Reliability* refers to the degree to which a measure gives the same result twice (or more) under similar circumstances, and may relate to the measure being used or the people using it. For example, if a patient's blood pressure is measured every 4 minutes on the same arm, by the same nurse, and the patient is not subject to any intervention such as activity or medication, you would expect to get similar sphygmomanometer readings. The extent to which repeated readings are similar is called reliability. Assessment of the similarity of repeated readings taken by the same nurse provides a measure of *intra-rater* or *within-rater reliability*. You would also hope that 2 different nurses measuring the same patient's blood pressure under the same circumstances would get similar readings. The extent to which the readings from 2 different nurses are similar is known as *inter-rater* or *between-rater reliability*.

Validity is the ability of a measurement tool to accurately measure what it is intended to measure. There are many different types of validity, but one of the most important is *criterion related validity*, which requires comparison of a given measure with a *gold standard*, or the best existing measure of the variable.[4] In the study by Steer *et al* in this issue (p 126), the results obtained from the Beck Depression Inventory for Primary Care were compared with the results of a standardised interview based on *DSM-IV* criteria and conducted by a physician. The interview results were considered to be the gold standard. Other examples of gold standards are direct central venous pressure readings for sphygmomanometer measures of blood pressure and serum hormone concentrations for the results of a urine test for pregnancy.

IS THE MEASURE SUBJECT TO BIAS?

There are several potential sources of bias. It is not important to remember what they are called, but you should be able to recognise sources of bias in a study. One way that bias can occur in a study is when the healthcare providers, patients, and data collectors participating in an intervention study are not *masked* or *blinded* to the treatment allocation. In an ideal world, studies would be "triple blinded"—that is, the healthcare provider delivering the intervention, the patient, and the research staff measuring the outcomes would not know which treatment the patient was receiving. Although triple blinding is possible in randomised trials evaluating new drugs, it is far more difficult to achieve in evaluations of most nursing interventions. Often, neither the nurses delivering the intervention nor the patients receiving the intervention can be masked (eg, nurses know that they are providing a patient education intervention and patients know that they are receiving it). In such studies, it is often possible, however, to mask the person measuring the outcome. By ensuring that the person measuring the outcome is masked to a patient's group allocation, researchers try to minimise the bias that could be introduced by unconscious adjustments assessors might make if they were aware of a patient's group allocation. For example, in the study by Dunn *et al* (p 117), which compared 2 interventions to increase physical activity, the people who assessed blood pressure, pulse rate, and body fat did not know which intervention participants had received. If they had known, this might have influenced their perceptions when they were doing the measurements, particularly if they had a clear opinion about which intervention was most effective. Similarly, participants reporting their own level of activity might alter their reporting of actual behaviour depending on whether they enjoyed or if they wished they had been allocated to a different group. Beginning with this issue of *Evidence-Based Nursing*, we will specify in the description of the design, whether the study was unblinded, single, or double blinded and who was blinded.

Another common type of bias is *social desirability bias*, in which people's responses to questions may reflect their desire to under report their socially unfavourable habits, such as the number of cigarettes smoked, illicit drug use, or unsafe sexual practices. Conversely, people may overestimate what they perceive to be socially desirable practices, such as exercise participation or daily intake of fruits and vegetables.

A third type of bias is *recall bias*, which acknowledges that human memory is fallible. Reports of seat belt use 5 years ago or fibre intake last month, for example, are not as accurate as concurrent or prospective measurements, where seat belt use or diet diaries are recorded on a daily basis.

Investigators often use strategies to try to overcome these potential biases. These strategies include having outcome assessors who do not know the purpose of a study nor which intervention the patient received; having study participants complete self report questionnaires in a private area, ensuring that their responses to sensitive or potentially embarrassing questions are confidential; and collecting information on a prospective basis (ie, as it happens), rather than on a retrospective basis (historically).

Conclusion

In summary, readers of research reports need to consider the type of measures that are used, the reliability and validity of the measures, and methods used to minimise bias in the measurement of outcomes. These are some of the elements considered when selecting studies for abstraction in *Evidence-Based Nursing*. In the next issue of the journal, the EBN Notebook will address how study outcomes are analysed and the appropriateness of the statistical test for the type of data collected.

DONNA CILISKA, RN, PhD*

**School of Nursing,*
Faculty of Health Sciences,
McMaster University,
Hamilton, Ontario, Canada

NICKY CULLUM, RN, PhD

Centre for Evidence Based Nursing,
Department of Health Studies,
University of York,
York, UK

ALBA DiCENSO, RN, PhD*

1 Thomson C. If you could just provide me with a sample: examining sampling in qualitative and quantitative research papers [editorial]. *Evidence-Based Nursing* 1999 Jul;**2**:68–70.
2 Norman GT, Streiner DL. *PDQ statistics*. Toronto: BC Decker, 1986.
3 Gravetter FJ, Wallnau LB. *Essentials of statistics for the behavioral sciences*. California: Brooks/Cole, 1998.
4 Anthony D. *Understanding advanced statistics. A guide for nurses and health care researchers*. Volume 4. Edinburgh: Churchill Livingstone, 1999.