Outcomes Part II

What Makes a Bad Outcome Measure?

The editorial in the November/December 2019 issue of *Research in Gerontological Nursing* discussed qualities of good outcome measures for randomized controlled trials (RCTs) (Kovach, 2019). In this editorial, I'll point out eight qualities to avoid in outcome measures, for there are many ways to botch a significant and well-designed RCT by making a few bad measurement choices.

CHOOSING A NON-VALIDATED SURROGATE OR CORRELATE MEASURE

Surrogate markers are often used as proxies in gerontology for an outcome that is either more distal in time, more difficult, or more expensive to directly measure. Problems can arise if the surrogate is correlated with the true outcome of interest, but the relationship is either not in the direct causal pathway or the pathway is multi-causal. There is a risk of a false negative conclusion if the surrogate outcome is not in the causal path. There is a risk of a false positive if the surrogate effects one causal path, but the intervention actually had an effect on another causal mechanism (Fleming & Powers, 2012). There are many examples in the literature of biomarkers used as surrogates that led to false conclusions (Fleming & Powers, 2012).

The dose of intervention needed to create a positive effect is important and is often difficult to establish in many nonpharmacological clinical trial studies. The dose of intervention needed to establish positive effects in the true outcome is unanswered when a surrogate measure is used. Although there is a lot of guidance on validating biomarker surrogate endpoints and criteria for judging the validity of a biomarker surrogate (Lassere et al., 2007), there is a need for more research on validating the use of psychometric tools and clinical assessment findings as surrogate measures.

INCONGRUOUS MEASURES

A measure must represent the intended construct. Just because a tool says it measures a phenomenon, does not mean the items are actually representing the intended construct. It is extremely important that researchers understand what they want to measure and if the tool they are considering using will actually measure that construct. Match the items in the tool with your theory and intended use as an outcome measure for a particular intervention. In addition, if the instrument you're using has not been validated for the population, situation, and purpose for which it is applied, measurement error and the validity of the entire study can be seriously compromised. Data from the electronic medical record (EMR) and secondary analyses studies are particularly prone to incongruous measurement.

POORLY SPECIFIED MEASURES

Failing to provide precise definitions for outcomes may introduce considerable measurement error. For example, if the development of comorbid events are outcomes in your study, what is the exact definition you'll use to include or exclude diagnoses such as urinary tract infection, pneumonia, or functional decline? If you obtain outcomes measures from the EMR, there is a need for precise definitions and procedures to guide extraction of information. If falls are an outcome, for example, is lowering a patient to the ground during transfer considered the same as a fall associated with a fracture?

THE WRONG SCALED APPROACH

There are three main types of measurement tools that we create or use with some sort of scaling approach: *indicator* (i.e., psychometric), *clinical* (i.e., causal or clinimetric),

Copyright © SLACK Incorporated

and *predictor variables*. For this discussion of outcomes used in RCTs, I am contrasting indicator and clinical variables. Indicator variables measure a latent construct. The latent construct is indicated by the items in the tool, but the items do not influence the underlying concept. The items are more or less measuring one attribute. Hence, summing the items and calculating means makes sense most of the time. The items in clinical health measures are not as homogeneous and may be causes of an underlying construct. For example, if a tool measures health-related quality of life (QOL), items might include pain, side effects of treatment, and social support. Pain is not a result or an indicator of QOL, but rather a cause of decreased QOL. For one research participant, the presence of just one item, such as severe pain, may be enough for low QOL. However, another participant without pain but with seven other items marked as moderately severe could also have poor QOL. In this situation, summing and calculating means does not make as much sense. Know the type of scaling approach you're using, the ability of that tool to detect clinically significant differences that are hypothesized to result from your intervention, and the appropriate analytic approach.

MEASURE TOO DISTAL OR LOW OCCURRENCE OF EVENT

This is a tricky one because if an important outcome is something that occurs infrequently, it may be clinically more meaningful than a more proximal outcome. For example, improved forced expiratory volume of the lungs following an intervention may be measured in a short time frame, but the more distal measure of lower respiratory infection may be more meaningful. If outcomes are distal to the intervention, however, many confounding variables can be introduced between the intervention and distal outcome. In addition, if the event occurs infrequently, the sample size needed to prevent Type II error may compromise feasibility. Specifying the mechanisms of action by which the intervention is hypothesized to achieve outcomes and thoughtfully choosing mediators and proximal and distal outcomes that are supported by theory is a reasonable approach to the conundrum of measuring infrequent and distal measures.

MEASUREMENT OF ABSOLUTE OR PERCENTAGE CHANGE AND OTHER UNIT OF MEASUREMENT ISSUES

An issue of considerable debate is whether outcomes in clinical trials should be expressed as absolute changes or percentage changes from baseline. One approach is to choose the method with higher statistical power. Another approach is to choose the method that communicates efficacy most clearly to people likely to be interested in understanding the results. A reduction in the standard deviation from baseline to posttest in the intervention group can increase confidence in the stability of the intervention group and the effectiveness of the intervention.

Presenting health outcomes in absolute or relative terms can influence interpretation and may lead to confusion about the efficacy of interventions. An absolute difference or change from baseline to outcome is a subtraction and a relative difference is a ratio. For example, if the risk for pneumonia is 2 (2%) in 100 in a group of older adults and 1 (1%) in 100 in older adults who are performing diaphragmatic breathing exercises, the absolute difference is 2% – 1% = 1%. If this is expressed as a relative risk, 1% is divided by 2% and the researcher can state that the diaphragmatic breathing intervention reduced the risk of pneumonia by 50%. Hence, it is important to know both the numerator and denominator when a relative risk is used.

MULTIPLICITY

Be careful not to use too many outcome measures or too many time points, as both can increase the probability of making a false claim regarding effectiveness. The timing of measures is critical. Optimally, we have good theory and pilot data to support the timing of our measures. Often, however, it is not that straightforward. Various factors impact timing of measures. A study may have multiple outcome measures, multiple components in the intervention, and different study participants who respond at different times to the treatment. So, repeated measures are common. Statistical corrections may be needed to account for the increased likelihood of Type I error on your models. Multiplicity also affects the degrees of freedom used in a power analysis and may make the sample size needed for your study very large.

SELF-REPORT AS A BAD IDEA

In the November/December editorial (Kovach 2019), I described a few situations in which patient perceptions can be useful in intervention studies. However, subjective reports can be highly biased. Humans tend to want to appear in a good light and may inaccurately judge aspects of themselves, their health, and their well-being. A person may not be aware that their gait has deteriorated, but measurement may reveal a decrease in step height. Outcomes may be exaggerated in trials that rely on subjective outcomes. Research funders may perceive a reliance on subjective measurement as methodologically simplistic and flawed if less biased and more precise objective measures are available.

CONCLUSION

Although feasibility is a necessary consideration, in this highly competitive funding market choose the most precise, valid, and reliable measures you can find. If that measurement approach requires you to work with technology, economic variables, or biomarkers that are out of your range of expertise, seek highly seasoned scientists who are known for and have a track record of funded research using those measurements. If you need outside experts, it is probably a more sophisticated study. Based on my experience, I can say that partnering with others who have expertise in measuring sophisticated outcomes has kept me growing and interested in my work, while helping me conduct more meaningful clinical trial research.

Conducting clinical trial research involves a huge investment of effort and resources. Using measures that fall short of detecting meaningful outcomes threatens our science and the health and well-being of older adults. Thoughtfully addressing how outcomes are chosen, collected, reported, and interpreted is fundamental to rigorous research pursuits.

REFERENCES

Fleming, T. R., & Powers, J. H. (2012). Biomarkers and surrogate endpoints in clinical trials. *Statistics in Medicine, 31*(25), 2973–2984. https://doi.org/10.1002/sim.5403 PMID:22711298

Kovach, C. R. (2019). Outcomes part I: What makes a good outcome measure? *Research in Gerontological Nursing, 12*(6), 271–273. https://doi.org/10.3928/19404921-20191024-01 PMID:31755963

Lassere, M. N., Johnson, K. R., Boers, M., Tugwell, P., Brooks, P., Simon, L.,...Wells, G. (2007). Definitions and validation criteria for biomarkers and surrogate endpoints: Development and testing of a quantitative hierarchical levels of evidence schema. *The Journal of Rheumatology, 34*(3), 607–615. PMID:17343307

Christine R. Kovach, PhD, RN, FAAN, FGSA
Editor

The author has disclosed no potential conflicts of interest, financial or otherwise.

doi:10.3928/19404921-20191206-01

Copyright © SLACK Incorporated