EDITORIAL

Does response shift bias invalidate asking patients if they are better or worse?

When assessing the success of treatments, clinicians ask patients whether and by how much the patient is better or worse than before therapy was started (ie, “Since xx, has your condition changed, and if so, by how much?”). Many trials do include such transition questions, but there is a debate as to their accuracy. The debate centers around the extent to which there is a ‘response shift’ caused by the patient changing their recollection and interpretation of the ‘before’ assessment by the time of the ‘after’ assessment. This has led the FDA to rarely accept ‘transition change’ scores (eg, how much has your pain/quality of life improved since the beginning of the trial?) in pivotal trials, preferring to subtract the ‘before’ absolute status (eg, how is your pain/quality of life today?) assessed at the beginning of the study from the same absolute (eg, how is your health/quality of life today?).

Response shift is defined as “a change in the meaning of one’s self-evaluation of a target construct as a result of 1) a change in the respondent’s internal standards of measurement (ie, scale recalibration); 2) a change in the respondent’s values (ie, the importance of component domains constituting the target construct; ie, reprioritization), or 3) a redefinition of the target construct (ie, reconceptualization).” Barclay and Tate examine the first 2 of these (recalibration of the target construct; ie, reprioritization), or 3) a redefinition of the target construct (ie, reconceptualization)."

Barclay and Tate examine the first 2 of these (recalibration and reprioritization) in a cohort of older men with and without stroke. Recalibration of physical function occurred in both groups. Reprioritization of role limitations because of physical health occurred in the stroke-free groups. In another article, Grovle et al. show that a transition index consisting of patient ratings of global perceived change over the past 2 years were strongly influenced by the current health status. In contrast to the above concerns, there is anecdotal evidence from arthritis and multiple sclerosis trials that patients can detect and report change before the absolute measure demonstrates this. The JCE has invited an article series from some thought leaders in this area to investigate whether and when such ‘transition change questions’ are valid.

Two other articles address methodological issues of patient-reported outcomes. Patient-reported outcomes are the focus of the PROMIS program, which is now producing different item banks. In this issue, Rose et al. present details of the development and some psychometric properties of the Physical Function item bank. This item bank consists of 124 PROMIS items covering upper, central, and lower extremity functions and instrumental activities of daily living; we still await a direct comparison with the existing legacy instruments in trials of interventions to establish their ability to reduce the sample size requirement in new clinical trials. Patient satisfaction is an important component of patient-reported outcomes and quality assurance. Hawthorne et al. challenge the traditional models of patient satisfaction developed in the 1980s and instead have adopted a new framework of 7 components (Appropriate access, provision of health information, empathy with the patient, participation in making choices, satisfaction with the treatment provided, effectiveness of treatment [including the extent to which treatment meets patient expectations of care and helps the patient in their daily life], and general satisfaction). They have developed and begun to establish the psychometric properties of a new generic, short, valid, and reliable measure of patient satisfaction, the Short Assessment of Patient Satisfaction (SAPS) scale.

Four articles in this issue address systematic review methods. Conventional diagnostic test accuracy (DTA) meta-analyses have the potential to provide summary estimates that are highly improbable for a particular target setting. Willis and Hyde propose a tailoring approach to address this by use of routine data from practice to define an “applicable region” for studies in receiver operating characteristic space. After qualitative appraisal, studies are selected based on the probability that their study accuracy estimates arose from parameters lying in this applicable region. This is applied to the example of the Pap test in the UK NHS Cervical Screening Programme.

Systematic reviews often exclude non-English language articles from their analyses, and there is evidence to suggest that English-language bias may reduce the pooled treatment effect size and produce less precise estimates of effect. However, study teams conducting systematic reviews may find it challenging to locate sufficient teams of foreign-language reviewers to include non-English language. Busse et al. propose use of a simple 3-step rule (excluding languages with less than 3 articles, reviewing titles and abstracts for clear indications of eligibility, and noting the lack of a clearly reported statistical analysis [unless the word “random” appears]); this led to accurate classification of 51 of 53 articles.

As attested to in past JCE articles, consensus is hard to achieve on searching strategies in general and in
disadvantaged populations where the terms are often vague and enormous numbers of citations are found in MEDLINE because of vague searching terms [1,2]; Cooper et al. report on a new population search filter for hard-to-reach populations. This filter increased search efficiency by reducing the number of references by 65% for a systematic review in order to capture all relevant populations (eg, homeless people, immigrants, substance misusers) in a public health systematic review. These authors tested the filter in a Medline case study to inform guidance in the UK on identifying and managing tuberculosis among hard-to-reach groups.

The impact of missing participant data in trials and meta-analyses of trials is always of concern if there is the possibility that in positive trials (ie, those with a claimed significant treatment effect) in the intervention group, the outcomes of participants with missing data are worse than the outcomes of those with available data. The traditional way of handling this is to assume that all dropouts in the experimental group had bad outcomes and those in the control group had good outcomes; this is often not plausible. Therefore, Ebrahim et al., in a previous paper in JCE, have proposed a process of a complete-case analysis as the primary meta-analysis and applying 4 progressively more stringent imputation strategies as sensitivity analyses to assess the risk of bias associated with missing participant data. The previous paper showed how to apply this to meta-analyses using continuous data with the same instrument [3]. The paper in this issue does the same with examples for continuous data measured with different instruments.

The remaining articles cover a number of different topics. The genetic epidemiology community has agreed on the GRIPS criteria for scientific reporting of Genetic topics. The genetic epidemiology community has agreed with different instruments. Peters et al. demonstrate this using the example of the likely falsely optimistic increased estimate of the change in life expectancy after the unification of East and West Germany. This example is used to demonstrate that this effect can be as large as 10 years.

A myriad of methods are used in the most common software packages to calculate sample size requirements for the difference between two proportions; these are dependent on the choice of sample size formula and software. Bell et al. used 4 sample size formulae to calculate sample size for 9 scenarios. Software documentation for SAS, Stata, G*Power, PASS, StatXact, and several R libraries were searched for default assumptions. Each package was used to calculate sample size for 2 scenarios. Sample size varied as much as 60% depending on the formula used. This was due to variation in the sample size formulae and the default settings implemented in multiple statistical software. Better documentation of the sample size formulae and default settings is required in many statistical software packages.

Peter Tugwell
J. André Knottnerus
Editors

E-mail address: anneke.germeraad@maastrichtuniversity.nl
(J.A. Knottnerus)

References