Journal of Clinical Epidemiology 111 (2019) v–vi

Journal of Clinical Epidemiology

EDITORIAL

Should the fragility index be routinely reported for systematic reviews?

Statistical significance can be very fragile. Given the strength of the comprehensiveness of systematic reviews, the robustness of the statistical meta-analyses is of substantial importance to those using them as the evidence basis for practice and policy. The clinical importance of the magnitude of the effect size and its statistical significance are most commonly used to assess this. In this issue Atal et al. add another- the fragility index. This is defined as the extent to which the statistical significance of meta-analyses can be changed (from statistically significant to non-significant, or vice versa) after modifying the event status of patients in specific arms of specific trials. In reviewing more than nine hundred meta-analyses it is salutary that the median number of events that would need to be misclassified to lose the statistical significance was only 12 and in a quarter of cases only five events need to be misclassified. In addition to the clinical importance and statistical significance of the effect size, maybe this fragility index should be also be routinely reported.

Visual presentation of complex data

More opportunities for visual presentation of complex data are appearing, since the number needed to treat has become one of the standard ways of communicating the size of benefits and harms from pairwise meta-analyses. Patients and clinicians will welcome the paper by Veroniki et al demonstrating six easy-to- understand approaches to graphically presenting the more complex comparative effectiveness estimates from network meta-analyses of more than one intervention: bar plot, Cates plot, or forest plot for a single outcome, and a bubble plot, scatterplot, or rank-heat plot for more than two outcomes.

Best practice in conducting systematic reviews is increasingly sophisticated as evidenced in the recent update of the Cochrane Handbook for systematic reviews of interventions. Publishing a protocol ahead of any systematic review is a core component of this best practice but for many reasons as Runjic et al. report a number of these Cochrane protocols never get published as full reviews. These authors make some suggestions for addressing this.

Systematic approaches to establishing core outcome sets for intervention studies are increasingly popular to ensure inclusion of the most important outcomes and to facilitate pooling of results in systematic reviews [1,2]. A common approach used by many organisations such as GRADE [3] is to use a nine point scale and include any outcomes rated as critical which is defined as 7,8, 9 on this 9 point scale. The experience in several core outcome set group such as OMERACT [1] and SONG [4] is that impractically large numbers on outcomes are ranked as critical [i.e. 7,8,9 on the 9 point scale]. De Meyer et al. report on their experience with core outcome set development for incontinence–associated dermatitis which included a controlled trial in using the Delphi process comparing a nine point scale to a three point scale; they show that restricting the categories to three substantially reduces the number of critical outcomes [from 24 to 13]. Further experience with other alternatives including ranking is needed.

Studies of diagnostic accuracy need a reference standard (often called a ‘gold standard’) against which the discriminatory ability of the test of interest (index test) is evaluated by comparing its results with those of the reference standard in a group of individuals suspected of the target disease. This reference standard is often established by expert panels where specific cases are definitively classified dichotomously as having/not having the target disease. Jenniskens et al. provide some convincing arguments backed by a simulation to show that this dichotomous approach risks developing biased estimates of the sensitivity and specificity of diagnostic tests used to make the diagnosis of interest. They make some suggestions on how this bias can be minimised.

Cashin et al. raise some cautionary concerns about the current enthusiasm reflected in the recommendations by funders in the UK and US to include the study of mechanisms in controlled trials of interventions using approaches such as mediation analysis. The fundamental goal is to decompose the “total effect” of an intervention (or exposure) on an outcome into an “indirect effect” that is channeled through a selected mediator and a “direct effect” that is not channeled through the selected mediator. There are accepted methods for this but there are no reporting standards and Cashin et al. report major deficiencies in between a third and a half of their sample of 54 published systematic reviews including 2008 primary studies, across 11 health care fields and 26 health conditions over the past 10 years, that may seriously limit the usability of this evidence. This needs addressing.

The increasing popularity of mixed methods research including both qualitative and quantitative designs reflects the acceptance that one is not superior to the other but that both

https://doi.org/10.1016/j.jclinepi.2019.05.008

are needed in health research. Thus it is good to see the paper by Hong et al. describing development and evolution of the mixed methods appraisal tool by a transdisciplinary group.

Two commentaries extend the fascinating controversy and debate covered by four recent articles [5-8] in the December 2018 issue on the merits or flaws in memory based methods for assessing dietary intake. Barnes tackles this by partitioning each argument into two principal components: data and claim. and then reviewing reasons for the disagreements, followed by suggesting potential opportunities for resolution. Porta and Vandenberg complete the series by calling for more integrative research between the biomedical and environmental research communities.

This issue also contains four more articles in the popular GRADE series covering topics including assessing bias in non randomised studies of interventions, assessing values and preferences for patient outcomes, and strategies from moving from diagnostic test accuracy to patient important outcomes and recommendations.

Peter Tugwell
Andre Knottnerus
E-mail address: ltugwell@uottawa.ca (P. Tugwell)

References

[1] Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino MA, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67:745−53.
[2] Turnbull AE, Dinglas VD, Friedman LA, Chessare CM, Sepúlveda KA, Bingham CO, et al. A survey of Delphi panelists after core outcome set development revealed positive feedback and methods to facilitate panel member participation. J Clin Epidemiol 2018;102:99−106.
[3] Guyatt GH, Oxman AD, Ksunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;64:395−400.
[4] Tong A, Craig JC, Nagler EV, Van Biesen W, SONG Executive Committee and the European Renal Best Practice Advisory Board; SONG Executive Committee and the European Renal Best Practice Advisory Board. Composing a new song for trials: the Standardized Outcomes in Nephrology (SONG) initiative. Nephrol Dial Transpl 2017;32:1963−6.
[5] Archer E, Marlow ML, Lavie CJ. Controversy and debate: Memory-Based Methods Paper 1: the fatal flaws of food frequency questionnaires and other memory-based dietary assessment methods. J Clin Epidemiol 2018;104:113−24.
[6] Martín-Calvo N, Martínez-González MÁ. Controversy and debate: Memory-Based Dietary Assessment Methods Paper 2. J Clin Epidemiol 2018;104:125−9.
[7] Archer E, Marlow ML, Lavie CJ. Controversy and Debate: Memory Based Methods Paper 3: Nutrition's 'Black Swans': Our reply. J Clin Epidemiol 2018;104:130−5.
[8] Martín-Calvo N, Martínez-González MÁ. Controversy and Debate: Memory-Based Methods Paper 4. J Clin Epidemiol 2018;104:136−9.