Journal of Clinical Epidemiology 79 (2016) 1–2

Journal of Clinical Epidemiology

EDITORIAL

Are triallists guilty of 'imbalanced salami slicing'- by favouring positive results in secondary publications?

In their nice paper in this issue, Ebrahim et al. add further evidence of the perverse incentives that contribute to publication bias; they show that the results of many RCTs are spread over several publications and that in a sample of 197 subsequent publications they test different usually not pre-specified hypotheses and almost always [over 90%] report at least one statistically significant result. The authors call for a more systematic approach of increasing public availability of protocols and providing more complete and complementary information from the trial in systematic reviews.

In a second article on RCTs, discontinuation of randomised trials due to inability to recruit sufficient patients is all too common and given the time, energy, resources and bureaucracy that are needed get a study started, let alone the opportunity costs from the patients that were entered this is a further example of research waste. Successful acquisition the outcome data is another cause of studies not being completed—there continues to be conflicting evidence for whether text message reminders improve questionnaire response rates; in this issue Kedinga et al. found no or little effects from text message notifications sent before [1% better] or after receipt [8%] of a follow-up questionnaire within an RCT for over 500 patients with depression; only little or no effect was seen.

In a third article on RCTs, Saragiottoa et al. provide yet another example of the dangers of subgroup claims, this time in low back pain trials; a total of 91 claims of a subgroup effect were reported in the 38 included trials, of which 28 were considered strong claims and 63 were cautious. None of the subgroup claims met all 10 credibility criteria, and only 24% satisfied at least five criteria. The authors call for the provision of a completed checklist as a table or electronic supplement by trials undertaking subgroup analyses; this should probably be done through a consensus process for reporting guidelines such as is recommended by the Equator initiative [1].

As Johnston et al. point out rare diseases pose a unique challenge to clinicians and researchers. Because of their low prevalence, establishing the impact of potential treatments is difficult. When sample sizes are necessarily limited, high instrument responsiveness (i.e., the ability to detect all important effects, even if small) is particularly important. They reviewed the state of the art by systematically reviewing the endpoints used in observational and experimental studies in rare lysosomal storage diseases. They found 62 interventional studies addressing patients with Fabry (55%), Gaucher (19%), Pompe (16%), and mucopolysaccharidoses (11%). Generic patient reported outcome measures were often used that are insufficiently responsive to detect small but important meaningful effects of new therapies; such type 2 errors of missing small but really important benefit will lead to injustice since the public rightly wants decisions on these therapies to be evidence—based.

In a second article on intervention study endpoints Juul et al. provide an example of how a questionnaire for assessing neck pain was developed based on a classic test theory and the International Classification of Functioning; they show how the floor and ceiling effects seen in the SF 36 instrument can be minimised. A third article on outcomes by Gorter et al. tackles the ongoing active debate as to the merits versus the complexity of using an item response theory (IRT) measurement model versus the more commonly used classical test theory measurement model (sum scores). They provide two examples, one real life and one simulated where they claim the IRT was clearly better. These are interesting but a systematic approach to comparing and contrasting the benefits versus the disadvantages, is needed to make progress in this debate.

JCE has published a variety of articles on setting priorities. A format that is attracting much attention at international meetings is the Evidence & Gap Maps format developed by Snilsveit et al. This provides an attractive visual display of what evidence is available and equally importantly what is missing [the Gap] for a full range of important health, wellbeing and economic outcomes for different intervention options. These are being used not only by researchers but also by commissioners and funders of research. Examples include approaches to HIV, maternal and child interventions, and water and sanitation.

Quality assurance is becoming a challenge to the GRADE methodology given its wide adoption. As with any successful approach training is essential to ensure quality, so given the impressive global spread of the GRADE approach with its adoption by multiple organisations, it is good to see the article by Norris et al. describing an instrument to provide and assess a set of minimum

http://dx.doi.org/10.1016/j.jclinepi.2016.11.005

skills and experience required to perform specific tasks related to guideline development using GRADE. This has three components, self report with a standardized ''GRADE curriculum vitae (CV)'', demonstration of skills using worked examples, and a formal evaluation using a written or oral test.

Turning to systematic reviews, Stanley and Massey argue that Cochrane Reviews need to not only describe potential biases but act to use accepted methods for adjusting for such biases to at least present alternative conclusions. They provide the worked example of how the use of meta-regression results in a dramatically different result when the sources of biases of concern are accommodated by meta-regression; the 50% to 70% increase in smoking cessation reported by the Cochrane Collaboration systematic review virtually disappears. Two letters (Mac Kenzie and Rogers and Stanley) comment on this. A second article on systematic reviews addresses selective outcome reporting; this is now felt to be of sufficient concern that the new Cochrane Risk of Bias (http://www.bristol.ac.uk/media-library/sites/social-community-medicine/robis/robisguidancedocument.pdf) instrument specifies this as specific bias that should be assessed in all systematic reviews. The paper by Tricco et al. adds to the evidence base of this problem. They found that a third of 96 systematic reviews registered in the PROSPERO database changed or did not specify the primary outcome.

Diagnosis and Prognosis: Korevaar et al. make a strong case that diagnostic accuracy studies should be prospectively registered in publicly accessible registries. In a review of 399 abstracts presented over 4 years at a leading Vision and Ophthalmology conference that reported estimates of sensitivity, specificity, area under the receiver operating characteristic curve, or diagnostic odds ratio, only 57% of these were subsequently published as full length articles in a peer reviewed journal. Biswas et al. suggest that useful partitioning by baseline risk can be clinically useful in targeting those at worst prognosis. They show that the benefit as assessed by Number Needed to Treat [NNT] for patients starting a cardiac rehabilitation program varies dramatically by the baseline risk of the patient [e.g., increasing age, lower baseline fitness, history of diabetes]. Defining these different levels of NNT will be useful to target patients for whom the expected survival yields and programmatic attentiveness needs are greatest.

Four articles address large database work: Toson et al. report on their new ICD-10 version of the Multipurpose Australian Comorbidity Scoring System to assess comorbidities in a cohort of 25,000 over 65 years old individuals with a hip fracture. They demonstrate that this performed better than the Charlson and Elixhauser comorbidity scales for accuracy in predicting in-hospital mortality, and 30-day mortality. Regarding geographic and temporal validity of prediction models, different approaches can be useful to examine model performance. Austin et al. using the example of a prediction model for chronic heart failure in nearly 15,000 patients admitted to 90 Canadian hospitals over 2 tine periods, show how the robustness of prediction models can be checked for geographically and temporal differences. In another ICES study Van Walraven et al. show that [in contrast to continuous variables], for categorical variables such as diagnostic codes, using the examples of renal disease and primary subarachnoid hemorrhage, administrative database mathematical analyses of the sensitivity and specificity of diagnostic code accuracy did not vary notably with changes in disease prevalence. Martens et al. carried out a series of simulations to see how often adding new risk factors to prediction models made a meaningful difference.

In another study relevant to avoiding research waste, Hemkens et al. document the parlous situation of poor reporting of studies using routinely collected health data. In a review of 124 studies in PubMed from 2012, less than half clearly described the design in title or abstract, framed a focused research question, provided essential information on codes or classification algorithms, or reported basic details required for replication.

Adding to the debate on sample size: Hanley discusses the background principles and provides an alternative set of sample size estimates for simple and multiple linear regression to those proposed by Austin and Steyerberg in this Journal last year.

Mehta et al. report on the scoring system of the Charlson Comorbidity Index (CCI) [2]. One of the points they raise is that in the calculation of the CCI score an additive approach is being applied instead of a multiplicative one, which is earlier also criticized by Harrell [3] and Moons et al [4]. This paper by Mehta is accompanied by a response from Charlson and Wells and an editorial comment.

Peter Tugwell
J. Andŕe Knottnerus
E-mail address: *Laura.Tugwell@uottawa.ca* (P. Tugwell)

References

[1] Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. BMC Med 2010;8:24.
[2] Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40:373−83.
[3] Harrell F. Regression coefficients and scoring rules. J Clin Epidemiol 1996;49:819.
[4] Moons KG, Harrell FE, Steyerberg EW. Should scoring rules be based on odds ratios or regression coefficients? J Clin Epidemiol 2002;55:1054−5.