Systematic Reviews of Incidence and Prevalence Studies: Tools for Assessing the Quality of the Primary Research

Marcel Dijkers, PhD, FACRM
Icahn School of Medicine at Mount Sinai
Department of Rehabilitation Medicine

This issue of KT Update *presents another in a series of brief articles by Dr. Marcel Dijkers. Tools are identified to evaluate the quality of the primary studies of incidence and prevalence as part of a systematic review of epidemiological reports.*

Evidence-based practice (EBP) needs up-to-date, reliable information relevant to treatment, diagnosis, prognosis, and other aspects of care and management. Teams of practitioners and researchers have obliged by producing high-quality primary studies and systematic reviews of such primary studies. Systematic reviewers are faced with the problem that primary studies may differ from one another in their basic design and in the quality of development and implementation of the blueprint for the research. It makes sense to rely more—or entirely—on the better studies, but what constitutes "better" (Verhagen, de Vet, de Bie, Boers, & van den Brandt, 2001)?

Methodologists have developed many study designs (what one might call big D decisions—e.g., randomized controlled trial [RCT] versus pre-post study) and added crucial details (small d decisions—e.g., blinding) that we generally can rank from weak to better to best. For instance, in treatment studies an RCT is better than a case series, and triple blinding is better than double blinding. But what if two studies differ in multiple aspects? Investigation A is better in aspects 1, 3, and 7, but study B better in aspects 2, 4, and 5—which of the two, if any, is better overall? We know that in principle better studies (that is, research that is designed to be strong and implemented in adherence to its protocol) come closer to the truth. In the parlance of systematic reviewers, they are less liable to generate biased results—that is, these investigations are more likely to produce a quantitative finding that (except for the vagaries of sampling or randomization) is an adequate reflection of reality. In practice, however, we do not know enough about how each specific aspect of research design and implementation affects study outcomes. Or even whether design (with a small d or capital D) and implementation affect various types of study outcomes differentially.

Two schools of thought have developed among those who are most directly confronted with this issue: systematic reviewers. One school suggests we use quantitative rating scales or checklists to create an overall quality score for each study being considered for a systematic review and use a cutoff score to either exclude low-quality studies or to identify quality classes that get a different weight in the review. There are a number of such rating scales (Moher et al., 1995; Olivo et al., 2008; Sanderson, Tatt, & Higgins, 2007); well-known for evaluating intervention research are the Jadad (Jadad et al., 1996) and PEDro (Maher, Sherrington, Herbert, Moseley, & Elkins, 2003) scales.

The second school essentially holds that by throwing multiple disparate items into the mixing bowl of such rating scales or checklists, we potentially penalize studies for design aspects that in actual fact do not affect the veracity of their findings (Higgins & Green, 2011, chapter 8). Proponents of this thinking state that in a systematic review we have to investigate the potential impact of each research design element individually and make decisions on the basis of the findings. Do randomized clinical trials (in intervention research) on average report a smaller effect size than observational studies? Then we should not give much weight to the observational studies. Does blinding of the clinician seem to make no difference for the effect size reported? Then for the outcomes of relevance in *this* particular review, the issue of clinician blinding can presumably be disregarded.

Both schools of thought have pros and cons. To evaluate the distortive effect of individual design elements, one needs a fairly large number of studies and has to assume (or even better, demonstrate) that there is no association between the presence of one design element and that of the next. The association between blinding the clinician and blinding the patient has to be essentially zero, or else any effect-size difference associated with clinician blinding might be partly if not entirely due to patient blinding. The only way around that is meta-regression, which requires an even larger number of primary studies. The rating-scale approach is limited by the assumption (made in all scales I am aware of) that all elements (random allocation, concealment, blinding, etc.) are equally important and that addition (the mainstay of traditional psychometric instruments) is the appropriate model for combining them (Nunnally & Bernstein, 1994; Streiner & Norman, 2003).

These thoughts came to mind when recently colleagues and I started a systematic review of the incidence and prevalence of enduring medical conditions in the chronic phase after the onset of moderate-to-severe traumatic brain injury (TBI). Systematic reviews of incidence and prevalence studies are not very common; systematic reviewers have focused on intervention research, with diagnostic and prognostic studies in a distant second and third place. In fact, a quick PubMed search performed in mid-November 2016 showed that there were 60,184 papers with "systematic review" in the title, only 1,702 of which (2.8%) also had "prevalence" or "incidence" in the title. (Among the latter, 117, or 6.9%, concerned disability or rehabilitation.) A more careful search in multiple databases might shift that percentage somewhat, but likely not by much.

Our issue was, What criteria do we use to evaluate the quality of the primary studies of incidence and prevalence? A fairly extensive search of the literature did not produce much in the nature of quality measures for the type of epidemiological research we planned to summarize. Most leads that emerged were dead ends: "critical appraisal" tools published or posted by various EBP or evidence based medicine (EBM) organizations. The problem is that these were developed for intervention research and have nothing specific to say about epidemiological research. Even the ones for cohort studies were not of use, because they looked at the study designs as observational studies of interventions and worried mostly about confounding. The same can be said about, for example, the well-known Downs and Black (1998) scale: a good instrument to dissect the quality of studies focused on intervention or causality in general, but offering limited applicability in our case.

We did find a few articles that discussed the issue of the quality of prevalence studies from an EBP perspective (e.g. Boyle, 1998; Fiest, Pringsheim, Patten, Svenson, & Jette, 2014; and Harder et al., 2014). A review article with a promising title (Sanderson et al., 2007) led us back to the critical appraisal tools. A short paper by Harder (2014) emphasized the issue that critical appraisal of prevalence studies should distinguish between methodological quality and reporting quality—good advice, but one that cannot always be followed in the thickets of study quality assessment. Harder's note was written as a comment on a paper by Munn, Moola, Riitano, & Lisy (2014) that reported on the development of a critical appraisal tool of prevalence studies. That article led us to two other papers, by Loney, Chambers, Bennett, Roberts, & Stratford (1998) and by Hoy et al. (2012).

Loney's critical appraisal tool really was a checklist with eight items, scored 0–1 points each. Its shortness and simplicity were attractive, but the questions asked appeared to us not to be incisive enough. Hoy's tool, in itself based on an earlier one by Leboeuf-Yde & Lauritsen (1995) recommended itself by providing criteria for scoring its 10 items as low-risk versus high-risk, and in addition, offering notes and examples for each. Unfortunately, the tool was created to assess prevalence studies of disorders that used door-to-door surveys in third-world countries. To make it fit the reports of chronic disorders after TBI on the basis of clinical series we were addressing required us to make some phrasing changes. In addition, these small series had a high risk of seriously underestimating or overestimating the true incidence, but the Hoy tool did not include an item addressing power. We created one that we added to the existing ten.

As always seems to be the case, after we had modified the Hoy tool and had started using it we came across two additional candidate instruments, one by Giannakopoulos, Rammelsberg, Eberhard, & Schmitter (2012) and one by Shamliyan et al. (2011)—the latter based on an apparently exhaustive review of earlier attempts (Shamliyan, Kane, & Dickinson, 2010).

It would seem that most teams creating systematic reviews of incidence and prevalence studies never search for a tool for quality assessment, or else give up the search. Shamliyan, Kane, & Jansen (2012) reported that of 145 reviews of this type, only 37% planned a quality assessment of the primary studies. Disability and rehabilitation researchers should, however, realize that, just as in the case of systematic reviews of primary studies of interventions or of diagnostic tools, a careful look at the quality of those studies is always warranted. And tools for doing so exist (Shamliyan et al., 2010). Whether one uses the information for omitting certain studies, or weighting all of them, or simply for informing the reader of the quality of the research underlying the evidence, a careful look at big D and small d design issues is always warranted.

References

Boyle, M. (1998). Guidelines for evaluating prevalence studies. *Evidence Based Mental Health, 1*(2), 37–39.

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health, 52*(6), 377–384.

Fiest, K. M., Pringsheim, T., Patten, S. B., Svenson, L. W., & Jette, N. (2014). The role of systematic reviews and meta-analyses of incidence and prevalence studies in neuroepidemiology. *Neuroepidemiology, 42*(1), 16–24. doi:10.1159/000355533

Giannakopoulos, N. N., Rammelsberg, P., Eberhard, L., & Schmitter, M. (2012). A new instrument for assessing the quality of studies on prevalence. *Clinical Oral Investigations, 16*(3), 781–788. doi:10.1007/s00784-011-0557-4

Harder, T. (2014). Some notes on critical appraisal of prevalence studies: Comment on: "The development of a critical appraisal tool for use in systematic reviews addressing questions of prevalence." *International Journal of Health Policy and Management, 3*(5), 289–290. doi:10.15171/ijhpm.2014.99

Harder, T., Takla, A., Rehfuess, E., Sanchez-Vivar, A., Matysiak-Klose, D., Eckmanns, T., . . . Wichmann, O. (2014). Evidence-based decision-making in infectious diseases epidemiology, prevention and control: Matching research questions to study designs and quality appraisal tools. *BMC Medical Research Methodology, 14*, 69-2288-14-69. doi:10.1186/1471-2288-14-69

Higgins J. P. T., & Green S. (Eds.). (2011). *Cochrane Handbook for Systematic Reviews of Interventions* (Version 5.1.0 ed.) Baltimore, MD: Cochrane Collaboration.

Hoy, D., Brooks, P., Woolf, A., Blyth, F., March, L., Bain, C., . . . Buchbinder, R. (2012). Assessing risk of bias in prevalence studies: Modification of an existing tool and evidence of interrater agreement. *Journal of Clinical Epidemiology, 65*(9), 934–939. doi:10.1016/j.jclinepi.2011.11.014

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials, 17*(1), 1–12.

Leboeuf-Yde, C., & Lauritsen, J. M. (1995). The prevalence of low back pain in the literature. A structured review of 26 Nordic studies from 1954 to 1993. *Spine, 20*(19), 2112–2118.

Loney, P. L., Chambers, L. W., Bennett, K. J., Roberts, J. G., & Stratford, P. W. (1998). Critical appraisal of the health research literature: Prevalence or incidence of a health problem. *Chronic Diseases in Canada, 19*(4), 170–176.

Maher, C. G., Sherrington, C., Herbert, R. D., Moseley, A. M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical Therapy, 83*(8), 713–721.

Moher, D., Jadad, A. R., Nichol, G., Penman, M., Tugwell, P., & Walsh, S. (1995). Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials, 16*(1), 62–73. doi:019724569400031W

Munn, Z., Moola, S., Riitano, D., & Lisy, K. (2014). The development of a critical appraisal tool for use in systematic reviews addressing questions of prevalence. *International Journal of Health Policy and Management, 3*(3), 123–128. doi:10.15171/ijhpm.2014.71

Nunnally J. C., & Bernstein I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw Hill.

Olivo, S. A., Macedo, L. G., Gadotti, I. C., Fuentes, J., Stanton, T., & Magee, D. J. (2008). Scales to assess the quality of randomized controlled trials: A systematic review. *Physical Therapy, 88*(2), 156–175. doi:ptj.20070147

Sanderson, S., Tatt, I. D., & Higgins, J. P. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. *International Journal of Epidemiology, 36*(3), 666–676. doi:dym018

Shamliyan, T., Kane, R. L., & Dickinson, S. (2010). A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for

diseases. *Journal of Clinical Epidemiology, 63*(10), 1061–1070.
doi:10.1016/j.jclinepi.2010.04.014

Shamliyan, T., Kane, R. L., & Jansen, S. (2012). Systematic reviews synthesized evidence
without consistent quality assessment of primary studies examining epidemiology of chronic
diseases. *Journal of Clinical Epidemiology, 65*(6), 610–618.
doi:10.1016/j.jclinepi.2011.10.017

Shamliyan, T. A., Kane, R. L., Ansari, M. T., Raman, G., Berkman, N. D., Grant, M., . . .
Tsouros, S. (2011). Development quality criteria to evaluate nontherapeutic studies of
incidence, prevalence, or risk factors of chronic diseases: Pilot study of new checklists.
Journal of Clinical Epidemiology, 64(6), 637–657. doi:10.1016/j.jclinepi.2010.08.006

Streiner, D. L., & Norman G. R. (2003). *Health measurement scales. A practical guide to their
development and use* (3rd ed.). Oxford, UK: Oxford University Press.

Verhagen, A. P., de Vet, H. C., de Bie, R. A., Boers, M., & van den Brandt, P. A. (2001). The art
of quality assessment of RCTs included in systematic reviews. *Journal of Clinical
Epidemiology, 54*(7), 651–654. doi:S0895435600003607