Meta-Analysis: The End of the Beginning, or the Beginning of the End?

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

This issue of KT Update (V2 N5) presents another in a series of brief articles by Dr. Marcel Dijkers. This article discusses the history, present, and possible future of meta-analysis.

Meta-analysis is defined in the Medical Subject Headings (MeSH) as "A quantitative method of combining the results of independent studies (usually drawn from the published literature) and synthesizing summaries and conclusions which may be used to evaluate therapeutic effectiveness, plan new studies, etc., with application chiefly in the areas of research and medicine" (http://www.ncbi.nlm.nih.gov/mesh/?term=meta+analysis). The SEDL-published manual for *Assessing the Quality and Applicability of Systematic Reviews (AQASR)* gives this description: **"**A (statistical) procedure that combines quantitatively the results of several studies that address the same question. This is typically done by identification of a common measure of effect size and other parameters that are more precise and less likely to be in error (due to sampling) than the individual studies being reviewed" (Task Force on Systematic Review and Guidelines, 2013).

It has been claimed that the British statistician Karl Pearson (developer of the Pearson correlation) was the first one to perform a meta-analysis, when the British government asked him to have a look at the results of various clinical trials of inoculation against typhoid fever performed at military bases all over the Empire. According to O'Rourke: "[Pearson] was especially thorough about questioning the consistency of individual trial results and equally keen to discover clues from this for better future research" (O'Rourke, 2007). Ronald Fischer, another name known to those who remember their basic statistics course, made contributions to the science of meta-analysis by combining studies in agriculture, and in his last book, *Statistical methods and scientific inference* (Fisher, 1956) "encouraged scientists to summarize their research in such a way to make the comparison and combination of estimates almost automatic, and the same as if all the data were available" (O'Rourke, 2007).

However, meta-analysis did not really take off until the 1970s, when American psychologists and social scientists started to combine results of the ever-growing numbers of research reports published in the increasingly large numbers of generalist and specialist journals. In fact, one of them, Gene Glass, author of an authoritative textbook in the area (*Integrating findings: The meta-analysis of research* (Glass, 1978), coined the term meta-analysis in a 1976 article in *Educational Researcher* (Glass, 1976).

At the time, the idea of statistical power was not well known, and many small and underpowered studies were published—a practice that still occurs today. Glass and the others realized that if the results of several small studies—each addressing the same research question—could be combined, a potentially less biased and certainly more precise answer to the question could be given. It was appreciated that ideally, the original datasets were to be merged in order to be analyzed again. However, when datasets were decks of punch cards (if that much) and people often did not hang on to datasets after their study was published, it would be a major effort to get authors of primary studies to dig in storage rooms to see if they still had their raw data. It was much easier to harvest the relevant information (a correlation, or a set of means and standard deviations) from the published paper.

Meta-analysis as we know it today consists of a careful scouring of the published (and unpublished) literature, extracting selected information from the published version of applicable studies, and pooling the means or other parameters from these studies without ever touching the individual cases. A number of algorithms have been developed to calculate or estimate the relevant parameters if the primary studies (as they often do) fail to provide a mean or the exact value of F for ANOVA, making inclusion of more studies into the meta-analysis possible (Rosenthal, 1991).

In the 1990's, meta-analysis was 'discovered' in clinical medicine by the new groups that developed the term and concepts of Evidence-Based Medicine (EBM). The application in rehabilitation and disability came even later, when these fields started creating their own version of EBM, generally under the banner of Evidence-Based Practice. Whether combined under the umbrella of the Campbell Collaboration (http://www.campbellcollaboration.org/) or the Cochrane Collaboration

(http://www.cochrane.org/) or operating independently, the clinicians and researchers who want to make evidence available for practitioners and patients/clients to act upon started developing standards for the quality of primary studies—something Glass et al. never had bothered with, or even rejected. Standards for the 'combinability' of study results were also developed, for instance, the various measures of study outcome heterogeneity. They also created new varieties of meta-analysis, such as meta-regression and network meta-analysis.

Even with these advances, meta-analysis of summarized data remains second best. If one had the raw data in a combined file, more could be done—for instance, performing analyses of subgroups, which generally are not analyzed in the primary studies, because even the best of them are not powered for subgroup analysis. There are other reasons that the primary data are needed, e.g. when the authors of a study do not report means but medians, or when other summary data cannot reliably be converted.

In unusual circumstances, an "individual cases meta-analysis" is possible without needing to beg authors for a copy of their dataset, for example, when the primary study reports consist mostly out of case series, and the individual cases can be combined. For instance, Creedon et al. in 1997 could paste together 22 case reports on the effects of intrathecal baclofen on spasticity, and managed to show that over the months after treatment start, the dosage of baclofen rose considerably (Creedon, Dijkers, & Hinderer, 1997). Other findings were based on a combination of results of such case series reports with results of five studies reporting the traditional descriptive statistics. The champions in this area are probably Oh et al., who found 68 reports with 348 patients who did or did not receive adjuvant radiotherapy after subtotal resection for ependymomas, a common glial tumor of the spinal cord (Oh et al., 2013). Their analysis enabled them to conclude that giving radiotherapy after resection improved progression-free survival of the patient.

Opportunities to harvest individual case data from the literature admittedly are uncommon, and researchers who are not satisfied with statistically combining summary data must start the hard work of acquiring and combining raw data sets. Nowadays, that is much easier than in the 1970s, because most data exist as electronic files in one

3

format or another on a desktop computer, and there is no longer a need to throw them out due to lack of space. Some of the larger and better-curated study datasets are even submitted to research archives or repositories such as the Inter-university Consortium for Social and Political Research (ICSPR: https://www.icpsr.umich.edu/icpsrweb/landing.jsp) which have as their purpose keeping the data intact 'forever' and making them available to qualified researchers.

On the other hand, for data not submitted to such archives, other barriers have been thrown up: informed consent, Institutional Review Boards (IRBs), and for those doing research in the medical and health sciences, HIPAA: the Health Insurance Portability and Accountability Act. For each dataset that is obtained directly from the original investigator, at a minimum the IRB must give approval for such a second use and often the original investigator must scrub her or his dataset clean from personal health information (PHI) before handing it over to a colleague.

And that is only the beginning of another set of problems that must be solved to make datasets combinable. Although software is available to 'translate' one format (e.g. dBase) into another (e.g. SPSS), that only starts a sequence of painstaking work: adjusting variable formats and names, and category codes, so that in the merged dataset each code in each column means exactly the same, wherever the case in question originally came from. One should not start the work of acquiring and streamlining datasets if one is not sure of the value that the analyses of the combined data are going to have. Such was the case for the IMPACT (International Mission for Prognosis and Clinical Trial) database, put together from eight trials and three observational studies by an international group of neurosurgery researchers, that has resulted in a long series of papers on the prognosis of acute care outcomes after moderate and severe traumatic brain injury (TBI), and the better design of future trials to improve such outcomes (Marmarou et al., 2007).

Fortunately, combining databases has become easier and will be even more so in the future, with the developments of defined datasets, common data elements and trialists' collaboratives. All these efforts have the same purpose: to make "the comparison and combination of estimates almost automatic"(O'Rourke, 2007). This is done by selecting key variables that need measuring in a particular domain of basic or

clinical science, and prescribing their data formats, including codes and variable names. One prominent effort in the area of spinal cord injury (SCI) rehabilitation is where the International Spinal Cord Society (ISCoS) has taken the lead in developing the International SCI Data Sets (http://www.iscos.org.uk/international-sci-data-sets). The datasets, available in English and in some instances in Chinese or another language, typically come in two flavors: Basic (suggested to be collected in clinical settings) and Extended (suggested for research use only). As of this writing, there are 19 datasets, that is, groups of variables describing a specific aspect of life after SCI (Biering-Sorensen et al., 2006).

A similar effort underway in the area of acute care and rehabilitation research in TBI has resulted in what are known as the TBI common data elements (CDEs: http://www.commondataelements.ninds.nih.gov/tbi.aspx#tab=Data_Standards). CDEs are recommendations made by prominent researchers for variables and measures to be collected in all TBI research, or at least in all research in a certain subarea (Bell & Kochanek, 2013; Hicks et al., 2013; Whyte, Vasterling, & Manley, 2010). The NIH's National Institute of Neurological Disorders and Stroke (NINDS) and various other federal agencies (including NIDRR) supported the creation of the TBI CDEs, and NINDS is now involved in further structuring the SCI datasets.

Several groups, calling themselves 'trialists' or a similar name, are working to make recommendations for best measures in a specific area and to standardize data collection in future clinical trials, so that merging datasets is a less onerous task—aside from the IRB-related issues. A quick search shows that we now have trialist collaboratives in various areas of health care, including stroke units and acupuncture. (These efforts are not to be confused with multi-site randomized controlled trials, where every site has to follow the exact same protocol developed either by the sponsor, or by the investigators collaboratively. In the 'trialists' groups, research questions and primary outcomes may differ from one investigator to the next, but the databases are designed in such a way that it is easy to merge them.) The aim of all these efforts is to create standards for research, in two meanings of that term: standard (best) information to collect, and standardized ways of formatting the information.

The actual analysis of the pooled individual data generated by two or more studies is not called 'pooled data secondary analysis' or something similar, but, at least in the medical area, is called 'individual patient (data) meta-analysis' (IPDMA). First introduced by Clarke and Stewart (1997) the term has been increasingly used; as often as 58 times in 2013, per PubMed. Without any doubt there were 'individual patient meta-analyses' before 1997 (for instance, it seems Pearson had individual soldier data), and papers are published that do not use the term. A June 2014 PubMed search came up with 286 citations total, while in 2013 Hannink et al. identified 583 IPDMAs in surgery alone (Hannink, Gooszen, van Laarhoven, & Rovers, 2013). The exponential increase in PubMed citations shows that this is a growth industry.

A logical extrapolation of all this may be to create a database into which all investigators in a particular domain can 'pour' their data prospectively. Any investigator could contribute part or all of the data from an investigation, once a subject in her research has given permission for his information to be shared with other qualified investigators, who may not even be known, for studies not yet identified. If necessary an investigator may create new variables that are unique to her project, until a time when they might prove useful in another investigation.

This is currently taking place in FITBIR, the Federal Interagency Traumatic Brain Injury Research informatics system (https://fitbir.nih.gov/). FITBIR is a depository for phenotypic, genomic and imaging data from TBI studies, that can be used to define and manage information, and contribute, upload and store data associated with studies. It incorporates and extends the TBI CDEs, and allows for uploading and sharing not just the data, but the protocols and data collection forms that were used to collect those data. A unique aspect of FITBIR is that each individual is given a Globally Unique IDentifier (GUID), which makes it possible to combine the same person's data, even if they were generated by multiple studies conducted by one or multiple investigators.

Pearson, and certainly Fisher, saw the potential of the combination and joint analysis of the individual case data of multiple studies, even though the technology that was available in their time prohibited turning the idea into reality. Even when Glass and his colleagues in the 1970s started doing one meta-analysis after another, the technology had not yet come far enough that individual subject meta-analysis could

become a reality. Their meta-analysis, as we know it from school or later self-study, was developed with a focus on the analysis of aggregated data as published in the literature. Will that approach be replaced entirely by individual subject data meta-analysis, whether the data come from studies combined after the fact, or designed to prospectively follow standards in measurement and recording? Are we observing the beginning of the end of meta-analysis as we knew it? Or is it the end of the beginning of an entirely new paradigm, where collaboration between scientists goes so far that they put their data into one joint vessel?

References

Bell, M. J., & Kochanek, P. M. (2013). Pediatric traumatic brain injury in 2012: the year with new guidelines and common data elements. *Critical Care Clinics, 29*(2), 223-238. doi:10.1016/j.ccc.2012.11.004 ER

Biering-Sorensen, F., Charlifue, S., DeVivo, M., Noonan, V., Post, M., Stripling, T., & Wing, P. (2006). International spinal cord injury data sets. *Spinal Cord: The Official Journal of the International Medical Society of Paraplegia, 44*(9), 530-534. doi:10.1038/sj.sc.3101930

Clarke, M. J., & Stewart, L. A. (1997). Meta-analyses using individual patient data. *Journal of Evaluation in Clinical Practice, 3*(3), 207-212.

Creedon, S. D., Dijkers, M. P., & Hinderer, S. R. (1997). Intrathecal Baclofen for severe spasticity: A meta-analysis. *International Journal of Rehabilitation and Health, 3*, 171-185.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Glass, G. V. (1978). Integrating findings: The meta-analysis of research. *Review of Research in Education, 5*, 351-379.

Hannink, G., Gooszen, H. G., van Laarhoven, C. J., & Rovers, M. M. (2013). A systematic review of individual patient data meta-analyses on surgical interventions. *Systematic Reviews, 2*, 52-4053-2-52. doi:10.1186/2046-4053-2-52

Hicks, R., Giacino, J., Harrison-Felix, C., Manley, G., Valadka, A., & Wilde, E. A. (2013). Progress in developing common data elements for traumatic brain injury research: version two--the end of the beginning. *Journal of Neurotrauma, 30*(22), 1852-1861. doi:10.1089/neu.2013.2938

Marmarou, A., Lu, J., Butcher, I., McHugh, G. S., Mushkudiani, N. A., Murray, G. D., . . . Maas, A. I. (2007). IMPACT database of traumatic brain injury: design and description. *Journal of Neurotrauma, 24*(2), 239-250. doi:10.1089/neu.2006.0036

Oh, M. C., Ivan, M. E., Sun, M. Z., Kaur, G., Safaee, M., Kim, J. M., . . . Parsa, A. T. (2013). Adjuvant radiotherapy delays recurrence following subtotal resection of spinal cord ependymomas. *Neuro-Oncology, 15*(2), 208-215. doi:10.1093/neuonc/nos286

O'Rourke, K. (2007). An historical perspective on meta-analysis: dealing quantitatively with varying study results. *Journal of the Royal Society of Medicine, 100*(12), 579-582. doi:100/12/579 [pii]

Rosenthal, R. (1991). *Meta-Analytic Procedures for Social Research* Sage Publications.

Task Force on Systematic Review and Guidelines. (2013). *Assessing the quality and applicability of systematic reviews (AQASR)*. Austin, TX: SEDL, Center on Knowledge Translation for Disability and Rehabilitation Research.

Whyte, J., Vasterling, J., & Manley, G. T. (2010). Common data elements for research on traumatic brain injury and psychological health: current status and future development. *Archives of Physical Medicine and Rehabilitation, 91*(11), 1692-1696. doi:10.1016/j.apmr.2010.06.031