ELSEVIER

Journal of Clinical Epidemiology 65 (2012) 503−510

Journal of Clinical Epidemiology

Robust meta-analytic conclusions mandate the provision of prediction intervals in meta-analysis summaries

Petra L. Graham[a,*], John L. Moran[b]

[a]*Department of Statistics, Macquarie University, Sydney, NSW 2109, Australia*

[b]*Department of Intensive Care Medicine, The Queen Elizabeth Hospital, Woodville, South Australia 5011, Australia*

Accepted 12 September 2011; Published online 21 January 2012

Abstract

Objectives: Results of meta-analyses typically conclude that future large studies may be mandated. However, the predictive ability of these estimates is deficient. We explored meta-analytic prediction intervals as means for providing a clear and appropriate future treatment summary reflecting current estimates.

Study Design: A meta-epidemiological study of binary outcome critical care meta-analyses published between 2002 and 2010. Computation of 95% DerSimonian-Laird and Bayesian random-effects meta-analytic confidence intervals (CI) and 95% credible intervals (CrI), respectively, and frequentist (PI) and Bayesian (PrI) prediction intervals for odds ratio (OR) and risk ratio (RR) were undertaken. Bayesian calculations included the probability that the OR and RR point estimates ≥ 1.

Results: Seventy-two meta-analyses from 70 articles were identified, containing between three and 80 studies each, with median nine studies. For both frequentist and Bayesian settings, 49−69% of the meta-analyses excluded the null. All significant CrI had high probabilities of efficacy/harm. The number of PI vs. PrI excluding 1 was 25% vs. 3% (OR), 26% vs. 3% (RR) of the total meta-analyses. Unsurprisingly, PI/PrI width was greater than CI/CrI width and increased with increasing heterogeneity and combination of fewer studies.

Conclusion: Robust meta-analytic conclusions and determination of studies warranting new large trials may be more appropriately signaled by consideration of initial interval estimates with prediction intervals. Substantial heterogeneity results in exceedingly wide PIs. More caution should be exercised regarding the conclusions of a meta-analysis.

Keywords: Meta-analysis; Random-effects; Prediction intervals; Frequentist analysis; Bayesian analysis; Posterior probability; Predictive distribution

1. Introduction

Meta-analysis has become an established paradigm in medicine [1]. However, a number of questions have continued to present themselves to this endeavor [2]; in particular, the predictive ability of meta-analyses, which has usually been appraised in terms of the discordance, or otherwise, between the results of meta-analyses and ''large'' trials [3−6]. The requirement for the most current evidence or ''updating'' of meta-analyses has also been recognized [7], and recent empirical investigations have addressed this requirement [8,9].

Thus, the need for both systematic overviews of randomized trials [10] and large simple randomized trials [11] has been expressed, although some degree of impatience with the former has recently materialized [12]. The question then remains: how best to approach the predictive ability of a meta-analysis? The discordance metric requires either a retrospective review of published large trials and meta-analyses or a prospective assessment with an obligatory publication time lag. That is, one must wait for either the next large trial or a ''signal'' for the updating of a previous meta-analysis. Depending on the particular topic, such a signal for updating may not manifest itself for a number of years [9]. We may thus ask what would be an appropriate future treatment summary that would reflect current meta-analytic estimates? This would appear to be the predictive distribution, from a Bayesian perspective, or a meta-analytic prediction interval from a frequentist perspective [13,14].

The purpose of the present study was to compare estimated Bayesian and frequentist prediction interval(s) from a series of binary outcome meta-analyses and answer the following questions: what is the relationship between these prediction interval(s) and current estimates with respect to interval estimates and interval width?; and what are the

* Corresponding author. Tel.: 2-9850 6138; fax: 2-9850 7669.
E-mail address: Petra.Graham@mq.edu.au (P.L. Graham).

0895-4356/$ - see front matter
doi: 10.1016/j.jclinepi.2011.09.012

What is new?

This study is the first to look at the effect of prediction intervals in a large number of published meta-analyses. Among the meta-analyses examined, approximately half suggested clinical benefit or harm, whereas only up to a quarter of studies had prediction intervals supporting this finding. Prediction intervals should be mandated as providing a more appropriate future treatment summary while accounting for the heterogeneity between the studies.

meta-analytic metrics (odds ratio [OR] or risk ratio [RR]), number of studies in the meta-analysis, and estimation technique (Bayesian vs. frequentist) determinate in any such relationship. Within the Bayesian paradigm, we also estimated the posterior probability (*P*) that the meta-analytic estimate and predicted treatment effect were equal to null or greater (where null = 1) for both OR and RR [15]. We determined that this overall approach would quantify meta-analytic predictive uncertainty in a more immediate and clinically accessible manner than previous analyses based on a discordance metric.

2. Methods

Meta-analyses with binary outcomes were identified and selected by electronic search over the period 2002 to June 2010. The search strategy was (1) restricted to a dominant medical paradigm; that of the critically ill, (2) an electronic search with key words ''meta-analysis,'' ''critically-ill,'' using the National Library of Medicine MEDLINE via OVID, and (3) and a focused electronic search of major critical care (*American Journal of Respiratory and Critical Care Medicine*, *Chest*, *Critical Care Medicine*, *European Respiratory Journal*, *Intensive Care Medicine*, *Journal of Critical Care*, *American Journal of critical care*, and *Thorax*), specialist, and general medical journals using the above key words. We reviewed the abstracts of trials generated by the electronic search, and the full-text of the meta-analyses was retrieved for detailed evaluation. Primary outcome data were separately extracted, entered, reviewed, and verified by the two investigators (J.L.M. and P.L.G.) before analysis. Where more than one *primary* outcome was listed, a statistically significant outcome was chosen if available. If the primary outcome contained subgroupings, overall summaries were obtained if that was what the authors intended. For two articles, results were obtained for each of two subgroups because a combination of these subgroups was not justified from a pathophysiological/ clinical perspective. Where duplicate meta-analyses addressed the same question, the latest calendar year publication was preferred.

For each meta-analysis, 95% confidence intervals (CIs) and 95% credible intervals (CrIs) for the overall OR and RR of treatment vs. control were respectively calculated using (1) a DerSimonian and Laird frequentist random-effects meta-analysis model [16] and (2) a Bayesian meta-analytic model of the form described by Warn et al. [17] except that vague normal [0, 100] priors were placed on the log odds of the control (or log of proportion of events in the control arm for the RR model) rather than on the proportion of events in the control arm; on the log scale, these priors are locally noninformative. Like Warn et al., we chose a uniform distribution on the interval 0–2 as a prior for the between studies standard deviation (τ). This represented a reasonable selection of plausible values for this parameter without being too vague.

Frequentist (PI) and Bayesian (PrI) prediction intervals were then calculated for each meta-analysis using a standard methodology [13,14]. Note that the PI uses the *t* distribution that is standard practice when the underlying variability of the true overall mean effect is unknown and must be estimated from the data. By contrast, the standard meta-analytic CI uses the normal distribution. In the Bayesian analysis, the posterior probability that the estimated overall treatment effect (*P*) or predicted treatment effect (*P*p) was $\geq$1 (i.e., null = 1) was calculated for both OR and RR [14,15]. This was easily achieved in the Bayesian context by counting the proportion of Markov chain Monte Carlo iterates that are the value of interest or greater; in this case, 1. A threshold $P < 0.1$ was considered as providing strong evidence for beneficial treatment effect (i.e., the estimated OR or RR was not often at least 1), and $P > 0.9$ was considered to provide strong evidence for treatment harm (i.e., the estimated OR or RR was often at least 1), after Aitkin et al. [18]. Although not calculated here, the posterior probability of more clinically relevant treatment effects may be calculated in the same way using an appropriate threshold for a given therapeutic area. Heterogeneity was presented as the variance, τ^2, between studies [19]; τ^2 close to 0 indicates little heterogeneity, τ^2 around 0.25 indicates moderate, and $\tau^2 > 1$ reflects substantial heterogeneity [20].

Computation used WinBUGS software [21], for Bayesian analyses, using three simultaneous runs of the program with disparate starting values, the first 10,000 iterations being discarded and results reported on the basis of a further 100,000 iterations; the R package ''meta'' [22]; and user-written routines.

3. Results

A total of 70 articles comprising 72 meta-analyses were identified, containing between three and 80 studies each, with median nine and interquartile range (IQR) eight studies; 16 meta-analyses had five or lesser combined studies. The meta-analysis articles were classified as pharmaceutical therapeutic (57%), nonpharmaceutical therapeutic