Quantifying the Amount of Heterogeneity in Meta-Analysis: A Comparison of Methods

Knapp G
Department of Statistics, University of Dortmund, Germany

In random effects meta-analysis, several confidence intervals on the between-trial variance have been proposed. These confidence intervals can be broadly categorized in two classes. Hardy and Thompson (1996) as well as Biggerstaff and Tweedie (1997) propose two intervals based on likelihood approaches. Biggerstaff and Tweedie (1997) and Knapp et al. (2006) consider appropriate quadratic forms in the treatment effects estimates and propose intervals based on approximate distributions of their quadratic forms. The performance of the confidence intervals with respect to actual confidence coefficient and average length rather differs for different effect size. Knapp et al. (2006) and Viechtbauer (2006) have conducted simulation studies for the effect sizes mean difference and log odds ratio for the above mentioned confidence interval. In the present paper, further results for normal and binary outcome measures will be provided with focus on the effect size standardized mean difference and risk difference.

References

Category I: Chance-related Heterogeneity
Type of presentation: Oral

**A Simple Prediction Interval for Random-Effects Meta-Analysis**

**Higgins JPT, Thompson SG, Spiegelhalter DJ**  
**MRC Biostatistics Unit, Cambridge, UK**

For situations in which heterogeneity cannot be suitably explained by study characteristics, a random-effects meta-analysis is a common approach to synthesizing results from a collection of clinical trials. It is usual to present the estimated mean of the random-effects distribution, along with its confidence interval, and often a test for heterogeneity. Indeed, many meta-analysis software packages produce only these statistics. We argue that this standard presentation of a random-effects meta-analysis is insufficient and potentially misleading, since it fails to illustrate the extent of variability in treatment effects across trials. The among-study standard deviation (often called tau) provides a quantification of the amount of heterogeneity, but can be awkward to interpret. We propose a simple prediction interval for the underlying effect in a future trial. This provides a useful description of the variability of effects across studies, and incorporates consideration of uncertainty in both the mean and the standard deviation of the random-effects distribution.
Non-Parametric Modeling of the Relationship between Baseline Risk and Treatment Benefit

Ghidey W, Stijnen T
Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands

In meta-analysis of clinical trials, often meta-regression analyses are performed to explain the heterogeneity in treatment effects that usually exist between trials. A popular explanatory variable is the risk observed in the control group, the baseline risk, which is usually linearly related. To investigate the relationship, however, fitting an ordinary least squares (OLS) regression of the treatment effect on the baseline risk or fitting a standard random effects meta-regression model can be very misleading (Sharp et al., 1996). The main criticism is that both methods do not consider the fact that the baseline risk is estimated from a finite sample and thus subjected to measurement error. In the literature alternative methods are suggested which accommodate this issue by assuming both the true baseline risks and the measurement errors normally distributed. However, in practice, the normality assumption on the baseline risk might be too strong to adequately describe the underlying true distribution. Among others, Ghidey et al. (2006) relax the normality assumption to more flexible distribution function. Here, we propose an alternative non-parametric method, which does not require any distributional assumption about the true baseline risk. We applied the functional measurement error model approach of Carroll et al. (2006). We illustrate our method on a number of simulated data sets and on a published meta-analysis data set.

References:

Meta Analysis of Randomized Controlled Trials Describing the Effectiveness of Venlafaxine in the Treatment of Major Depressive Disorder in Comparison with Alternative Antidepressant Therapies – Effect on Response, Remission and Relative Tolerability

Freemantle N, Tharmanathan P
University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

Background: A number of different antidepressant types are available, and many randomized trials (most modest in size and statistical power) have evaluated their relative effectiveness. Venlafaxine is a well established antidepressant, and previous work has indicated that it may be superior to SSRIs in treating depression. Various methods for meta analysis of binary data have been developed, including theoretically exact conditional maximum likelihood methods for fixed effects analysis, and numerical simulation methods based upon the Gibbs Sampler, both of which are statistically superior to standard methods.

Methods: We conducted a meta analysis of all available trials comparing venlafaxine and SSRIs examining the outcomes of response, remission and relative tolerability. Trials were identified through searches of Medline, Embase, Cochrane Library and through accessing unpublished trials held by the manufacturer. Results based on intention to treat analyses, were pooled using theoretically exact methods for fixed effects, and numerical simulation using a Gibbs sampler for random effects. In the random effects analyses the between trial component of variance is parameterized and estimated with uncertainty from the data, rather than taken from the observed heterogeneity as in standard methods. Where significant heterogeneity was observed we aimed to explore the causes using meta regression. In the absence of significant heterogeneity we described the full random effects analyses to incorporate the observed variability between the results of trials, and the uncertainty of that heterogeneity, in the estimate of treatment effect.

Results: We identified 34 trials comparing venlafaxine with an SSRI, including 6374 patients. Venlafaxine was compared with fluoxetine in 18 trials, with paroxetine in 6 trials and with sertraline in 4 trials. Other comparators were citalopram (2 trials), escilatopram (2 trials) and fluvoxamine (2 trials). Response to venlafaxine was superior to that of alternative SSRIs, odds ratio 1.17 (95% CI 1.05 to 1.30; P = 0.0052) for the fixed effects analysis, and 1.19 (95% CI 1.05 to 1.34) for the full random effects analysis. There was no evidence of heterogeneity on this outcome (p 0.29, I² = 12% (95% CI = 0% to 44%).

For remission, venlafaxine was superior to SSRIs, odds ratio 1.24 (95% CI 1.10 to 1.40; P = 0.0004). There was some evidence of heterogeneity (P = 0.08, I² (inconsistency) = 31% (95% CI = 0% to 59%), however this had little effect on the full random effects model, odds ratio 1.27 (95% CI 1.11 to 1.49). Overall drop out was similar for SSRIs and venlafaxine.

Conclusion: Venlafaxine is more effective then SSRIs in achieving response and remission, and appears similarly tolerated. Theoretically exact meta analysis based upon conditional maximum likelihood, and full random effects meta analysis based upon the empirically Bayesian Gibbs sampler, performed well and provided complementary results.
Illustrating Heterogeneity Graphically - the Inclusion of the Estimated Inter-Study Variation into Forest Plots for Random Effects Meta-Analyses

Skipka G
Institute for Quality and Efficiency in Health Care (IQWiG), Cologne, Germany

Background: Meta-analyses are widely used to combine the results of clinical studies by calculating statistics for overall treatment effects. Basically, two different models exist for meta-analyses. The fixed effects model (FEM) assumes each study is measuring the same treatment effect $\theta$. Different estimations for $\theta$ are expected to arise from sampling error only. By contrast, the random effects model (REM) incorporates an inter-study variation $\tau^2$ taking heterogeneous results into account. Usually, it is assumed that the true treatment effects are normally distributed with expectation $\theta$ and variance $\tau^2$. Although these two approaches estimate different parameters (true effect vs. expectation of the distribution of true effects) the results are represented in the same way in practice. Commonly, the point and interval estimation of $\theta$ is drawn in a forest plot as a diamond. But the estimation for the inter-study variation $\tau^2$ is ignored in the representation of the results of REMs.

Objectives: To suggest a graphical approach for including the estimated inter-study variation into forest plots when representing the results of random effects meta-analyses.

Results: We include two rows for the summary statistics in the forest plot in case of REMs: The row “total expectation (95% CI)” represents the point and interval estimation for $\theta$. The row “total heterogeneity (95% CI)” represents the interval $[a-1.96 \times b; a+1.96 \times b]$, where $a$ and $b$ indicate estimators for $\theta$ and $\tau$, respectively. This “heterogeneity interval” delivers an approximated interval, where 95% of the true effects are to be expected.

Conclusions: This proposed extension of the forest plot may be helpful to accurately distinguish the results of meta-analyses from FEMs and REMs and to illustrate the amount of heterogeneity graphically.
Meta-Analytic Methods for Diagnostic Test

Kaufmann J
Clinical Statistics EU, Schering AG Berlin, Germany

A diagnostic test is a procedure to increase the probability of a correct diagnosis in diseased and non-diseased patients or diseased and non-diseased observational units, e.g. vessel segments or liver segments. In contrast to therapeutic studies, diagnostic studies are applied to a mixed population of patients with and without the interested diseased.

The most widely used performance measurement is the accuracy, an index of validity for the total population. In demonstrating the efficacy of a new diagnostic agent the true-positive rate (TPR), or sensitivity and the true-negative rate (TNR), or specificity described the outcome of a study.

We consider how to combine independent studies of the same diagnostic test, where each study reports an estimated true-positive rate (TPR) and true-negative rate (TNR). A summary receiver operating characteristic curve SROC has been recommended to present the performance of a diagnostic test, based on data from several independent diagnostic studies (RCT’s).

References

Ann.Inter.Med., 120; 667-676

Benjamin Littenberg, Lincoln E. Moses (1993), Estimating Diagnostic Accuracy from Multiple Conflicting Reports: A New Meta-analytic Method
Med Dec Making, 13 : 313.321

Lincoln E. Moses et al (1993), Combining Independent Studies of a Diagnostic Test into a Summary ROC-Curve: Data Analytic Approaches and some Additional Considerations
Statistics in Medicine, Vol. 12; 1293-1316
Random-Effects Meta-Analysis of Studies Reporting Pairs of Sensitivity and Specificity: A Comparison of Methods

Hamza TH¹, Reitsma JB², Stijnen T¹
¹ Department of Epidemiology and Biostatistics, Erasmus MC - Erasmus University Medical Center, P.O.Box 1738, 3000 DR Rotterdam, The Netherlands
² Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, P.O.Box 22700, 1100 DE Amsterdam, The Netherlands

A Random effects meta-analysis of diagnostic accuracy studies take into account both the within study variability and the between studies variability. The standard method to meta-analyze diagnostic tests when studies report a pair of sensitivity and specificity or a two by two table is the method of Littenberg and Moses [1]. This method has several shortcomings, including the following ones. It does not take account the heterogeneity across studies, assumes the proxy for the threshold is measured without error, does not consider the within study correlation between the estimates and 0.5 is added in each of the two by two table whenever a zero count is encountered. In this paper we considered three different random effects approaches: univariate, approximate bivariate and exact bivariate [2-4], which repair some or all of the shortcomings of Littenberg and Moses [1] approach. Besides, the bivariate approach enables one to derive easily different outcome measures, such as diagnostic odds ratio, likelihood ratios and predictive values. The methods are compared through a simulation study, in terms of bias, mean squared error and coverage probabilities of the confidence intervals. We varied the overall sensitivity and specificity, the between studies variances and covariance, the within study sample size and the number of studies included in the meta-analysis. The methods are illustrated using a published meta-analysis data set.

The exact bivariate method performs better than the univariate and approximate bivariate methods, and gives unbiased estimates of the accuracy, slope and residual parameters. The coverage probabilities are also reasonably acceptable. In contrast the univariate and approximate bivariate methods can produce large biases with poor coverage probabilities especially if sample sizes of individual studies are small or if sensitivities or specificities are large.

Keywords: Meta-analysis, diagnostic test, univariate random effects, bivariate random effects, sensitivity and specificity.

References:
Impact of Methodological Variations Between Randomized Clinical Trials on Results of Meta-Analysis: A Systematic Review and Meta-Analyses of Empirical Evidence

Mukhtar AM, Timm J
Competence Center for Clinical Trials Bremen, University of Bremen, Germany

Background: Combining randomized clinical trials (RCTs) meta-analytically without considering their variations in methodological quality ignores the importance of critical appraisal and may lead to biased results. Meta-epidemiological studies yielded inconsistent results in terms of the existence, direction and quantity of design- and analysis-related bias.

Objectives: The purpose of this study is to systematically review the literature comparing effect sizes of high quality RCTs (HQ-RCTS) and of low quality RCTs (LQ-RCTs). Furthermore, we investigated whether other sources of heterogeneity (chance-, patient- and intervention-related variations) between RCTs were regarded in these comparisons.

Methods: Systematic review and meta-analysis of empirical studies relating the quality of RCTs to their effect sizes. We searched the Cochrane Methodology Register, screened “related articles” of six key meta-epidemiological studies in Medline and checked references of included studies. We included studies which (i) compared the effect sizes of LQ-RCTs vs. HQ-RCTs, (ii) included one or more meta-analysis and (iii) assessed quality of RCTs using scores or the following individual components: generation of allocation sequence, concealment of allocation, blinding, handling of withdrawals. We excluded comparisons that used vote counting. We used the ratio of odds ratios (ROR, i.e. summary effect size of LQ-RCTs to that of HQ-RCTs) as an outcome. For quality scores and each quality component, we combined RORs using a random-effects model. We used regression analysis to investigate whether RORs differ as to medical conditions and interventions.

Results: 109 citations were retrieved for detailed evaluation. 23 studies including 103 empirical comparisons between LQ-RCTs and HQ-RCTs were included in this review. These comparisons were based on 216 meta-analyses and 2273 RCTs. An overestimation of treatment effect by LQ-RCTs in comparison with HQ-RCTs was observed, when Jadad-Score was used for the appraisal of quality [ROR= 0.82 (95% CI: 0.68 - 0.98), 9 comparisons]. RCTs with inadequate allocation concealment and those without blinding exaggerated the effect size by 15% [ROR= 0.85 (0.79 – 0.92), 24 comparisons] and 11% [ROR= 0.89 (0.82 – 0.97), 28 comparisons] respectively, in comparison with HQ-RCTs in these aspects. No relation between the adequacy of allocation sequence and effect estimate was found [ROR= 0.92 (0.82 – 1.02), 12 comparisons]. Similarly, no difference was detected between LQ-RCTs and HQ-RCTs with regard to the appropriateness of reporting or handling of withdrawals (intent-to-treat-analysis) [ROR= 1.01 (0.97 - 1.05), 23 comparisons]. No association was found between RORs and the medical condition in which it was estimated. RORs calculated from one meta-analysis did not differ from those based on more than one meta-analysis. About half of the comparisons investigated other design characteristics and used a random-effects model for synthesis. Patient- and intervention-related variations between RCTs were rarely addressed.

Conclusions: Randomized trials with low quality score, inadequate concealment of allocation and without blinding overestimate the efficacy of interventions. However, most of these empirical studies did not consider clinical causes of heterogeneity. This fact may confound the relation between methodological quality and effect size. Simultaneous investigation of diverse sources of heterogeneity is required.
Impact of Design Characteristics on Outcome in Studies on Tension-Type Headaches

Verhagen AP¹, Stijnen T²
¹ Department of General Practice
² Department of Epidemiology and Biostatistics
ErasmusMC, PO Box 2040, 3000 CA Rotterdam, The Netherlands

Objective: We studied the influence of the methodological quality of individual trials on the outcome of a systematic review on conservative treatments in patients with tension-type headache.

Methods: Included were studies included in five systematic reviews on tension-type headache. From each study we extracted the number of patients in both groups who recovered or improved during follow-up and we extracted data on headache outcome measures such as severity, intensity and frequency. Methodological quality was assessed using the Delphi list. Dichotomous as well as continuous outcomes were used in the analysis. Regression analysis on separate design characteristics and the overall methodological quality on size of treatment effect is performed.

Results: Out of the original dataset of 134 studies 61 studies fulfilled our selection criteria. The number of studies presenting dichotomous was slightly larger than studies presenting continuous data. Our study sample has a higher overall quality compared to the original sample and studies evaluating pharmacological interventions were over represented. All criteria show a non-significant relation with the effect estimate. Whether the outcome is measured dichotomous or continuous appear to have a greater (significant) impact on treatment effect estimates compared to the design characteristics.

Conclusion: In this study sample design characteristics do not show to have an impact on treatment effect estimates, but the way the treatment effect is measured has a significant impact.
During the past twenty years meta-analyses proved to be a valid tool to detect minor to moderate, yet clinically worthwhile improvements in the treatment of malignant diseases. Often (especially when the analysis is based on published data or even conference presentations only) no or limited data on important covariates are available. On the other hand, cancer patients typically exhibit a large amount of disease heterogeneity even within a single trial.

Advanced, inoperable pancreatic cancer is a malignancy with an extremely unfavourable course. After standard chemotherapy, consisting of single drug gemcitabine (GEM) since its introduction in the mid-90s, the median survival time is around six months. Since 2000, a considerable number of randomised trials examined the question, whether the addition of a second cytostatic drug might improve the prognosis. However, while progression-free survival was significantly prolonged in some of the individual studies, only one trial was able to show this for overall survival.

A meta-analysis including 14 trials (reported between 2002 and 2006) comparing GEM vs. GEM+X, comprising a total of 4335 patients and predominantly based on published data (partly in form of conference presentations only), revealed a risk reduction in survival by combination chemotherapy with a pooled hazard ratio of 0.91 (95% conf. interval: 0.85 – 0.97, p=0.004). The overall test for heterogeneity resulted in p=0.78 (I²=0%) and consequently, the application of a random effects model gave virtually the same results as the fixed effects one, although there were some hints that the effect size depended on the class of drug X.

Nevertheless, these presumably clear results proved to be only one side of the coin and probably clinically misleading. The publication on evidence of a strong (possibly qualitative) treatment by performance status interaction from one trial office led to similar analyses in several of the other trials (with 1682 patients in total). A meta-analysis of the trials with information on this patient-level covariate resulted in a much more relevant effect on survival in the good performance patients (hazard ratio: 0.76, 0.67 – 0.87, p<0.0001), while in those with an initial poor performance status combination therapy seems to be inefficient or even harmful to the patient (hazard ratio: 1.08, 0.90 – 1.29).

In conclusion, a minimum of information on the treatment effects in selected subgroups based on established prognostic factors should be recommended in publications of large randomised cancer studies.

References
A Systematic Overview of Economic Analyses of Routine Influenza-Virus Vaccination of Children During Inter-Pandemic Periods

Mau J
Heinrich Heine University Hospital, Institute of Statistics in Medicine, Düsseldorf, Germany

Influenza is a recurrent epidemic with a potential of world-wide "mega-kill". An adequate level of immunization in populations is believed to provide a time window for production and delivery of antiviral drug treatment. Though the issue seems compelling, decision-makers will not engage into a costly program of immunization of a significant part of the population, easily. Previous research found a reduction to one-third of influenza-like illnesses when more than 85% of children were vaccinated, while the Japanese experience amounts to having prevented about 11,000 deaths from pneumonia and influenza per year by routinely vaccinating school-age children [2], and modeling studies would imply containment of annual influenza epidemics by vaccination of about 60% of children in the USA [1].

Therefore, targeted immunization of major spreaders and in pools of high contact rates seems a reasonable strategy under several aspects: (i) reduced incidence in targeted subpopulation as well as at community level ('herd immunity'), (ii) reduced influenza-related morbidity and mortality in vulnerable subpopulations, (iii) reduced health costs, (iv) increased productivity, and also (v) increased readiness for pandemic situations through expanded production capacity and logistics.

With regard to these considerations, immunization of children, of day care, preschool and school ages, has come into focus. The discussion has mainly been led in the USA, and was taken up in Spring 2006 in the EU. The monetary commitment that routine immunization of children during inter-pandemic periods would entail has motivated economic studies of cost-benefit, too.

Economic analyses differ in methodologies, age groups of children, health systems, assumptions about cost factors, and in including influenza-related mortality and vaccination-related adverse events or not, and results are clearly not uniform. Studies published in 2005 and 2006 are less optimistic, due to more conservative assumptions and a wider set of cost factors. Consequently, routine vaccination of all children is not seen as cost saving in general anymore, and routine vaccination of high-risk children of all ages plus routine vaccination of all children aged 6 to 23 months seems to emerge as currently best founded from a theoretical and empirical viewpoint.

The systematic overview is based on a common decision-tree model in order to identify differences in modelling assumptions and lacking data. The dependence of most influential cost factors on differences in socio-cultural setting between the US, Asia, and Europe is the most consequential, though not unexpected finding.

References

Meta-Analysis Based on Individual Versus Aggregate Patient Data: A Systematic Review of Empirical Comparisons

Mukhtar AM, Timm J
Competence Center for Clinical Trials Bremen, University of Bremen, Germany

Background: Meta-analyses based on individual patient data (MA-IPD) from randomized clinical trials (RCTs) are considered as the “gold standard” for performing systematic reviews by many authors 1, 2. However, published mathematical derivations emphasize the equivalence of summary effect sizes estimated by MA-IPD and by meta-analysis using aggregate patient data (MA-APD) 3, 4. The main advantage of MA-IPD over MA-APD is suggested to be related to the investigation of differential treatment effects in subsets of patients and interventions.

Objectives: The purpose of this study is to systematically review the literature dealing with empirical comparisons between MA-IPD and MA-APD with regard to: (1) the incorporation of chance-related heterogeneity when combining RCTs (synthetic function of MA) and (2) the investigation of methodological and clinical causes of variations between RCTs (analytic function of MA).

Methods: We searched the Cochrane Methodology Register and Medline for citations related to MA-IPD. Moreover, we screened fifty-one trialists' collaborations, the Cochrane Database of IPD Reviews and references of included studies for relevant comparisons. We included comparisons which (i) were based on RCTs and (ii) presented quantitative results for MA-IPD and MA-APD or from which these can be calculated. We excluded citations comparing only the structures, costs or organization of both types. We calculated the ratio of odds ratios (ROR, i.e. summary effect size of MA-APD to that of MA-IPD) for each comparison and combined them using a random-effects model. Furthermore, we studied the relation between summary effect sizes of MA-IPD and MA-APD using a weighted least squares regression.

Results: 53 citations were retrieved for detailed evaluation. 18 studies were included in this review. Almost one third of them was in oncology. 23 empirical comparisons were extracted. Half of them used mortality or survival as outcome measure. Half of them used mortality or survival as outcome measure. Only one comparison used a positive outcome. All MA-IPD considered stratification by RCT and all but two of them combined the data after the intention-to-treat principle. Most of the comparisons do not report on test of heterogeneity and combine the results in fixed effects model. Weighted regression showed no difference between summary effect sizes of the studied MA-IPD and MA-APD [OR= 0.99 (95%-CI 0.93 – 1.15)]. Meta-analysis of the RORs yielded similar results [summary ROR= 0.99 (95% CI 0.55-1.78)]. Methodological variations of RCTs were considered in 13 comparisons, two of them found a relation, however inconsistent, between the quality of RCTs and the effect size. Clinical heterogeneity was investigated by both types of meta-analysis in only 4 comparisons, with no consistent differences in this regard.

Conclusions: (1) In accordance with the mathematical derivations 3, 4, no empirical evidence of difference in the synthetic function of MA-IPD and MA-APD was found. (2) Most of the empirical studies did not compare MA-IPD and MA-APD with regard to their analytic function. Therefore, this function requires more investigation.

Literature:
A growing interest in sharing research data can be observed in many disciplines, especially in genetics and molecular biology [Noor et al., 2006]. Thereby we recognize a shift towards decentralization of data, integration of various databases and sophistication as well as simplification of data linkage, manipulation and analysis. However this interest is confined mainly to data from observational studies and routinely collected data.

The superiority of meta-analysis based on individual patient data (MA-IPD) in contrast to meta-analysis using aggregated patient data, is suggested to be mainly related to enabling: (1) the investigation of differential treatment effects in subgroups of patients and interventions - enabling more reliable investigation of clinical heterogeneity and tailoring of health care (2) the checking of data within trials for completeness – combating inadequate reporting and reducing reporting bias - and for plausibility - reducing errors and detecting fraud - (3) the standardization of variables between trials - enabling more synthesis – (4) the execution of time to event analysis according to the intention-to-treat principle as well as (5) the collection of more data by including unpublished randomized clinical trials – reducing publication bias - and more patients and follow up data - reducing patient exclusion bias -. However, collaborative meta-analyses of randomized clinical trials based on individual patient data are still uncommon [Koopman et al., 2005; Mukhtar, 2006]. It is well recognized, that MA-IPD is a more time and resource consuming endeavour. Moreover, sharing individual patient data is facing limitations related to research cultures/ policies, privacy/ confidentiality, intellectual proprietary/ copy rights, patents and data infrastructure /deposition /management.

The main focus of our study is confidentiality and privacy issues related to sharing individual patient data for research purposes. We reviewed the legal situation in Germany with regard to anonymisation, pseudonymisation and data security. Accordingly, we developed a model structure for data sharing. We addressed controversial issues such as balancing omission of identifiers and extending data analysis as well as the question whether patient consent for data sharing is required. Furthermore, we appraised relevant policy and legal frameworks from UK and USA.

Literature

Koopman L, van der Heijden G, et al. (2005) Methodology of subgroup analyses used in individual patient data meta-analyses versus meta-analyses on published data. XIII Cochrane Colloquium; Oct 22-26; Melbourne, Australia
Sample Size Distribution - An Overlooked Source of Spurious Bias in Funnel Plots

Rücker G, Schwarzer G, Carpenter JR
Institute of Medical Biometry and Medical Informatics, University Medical Centre Freiburg, Germany

Background: Comparing the recently proposed arcsine test with established tests for publication bias in meta-analyses with binary outcomes (1), we found that power and size depended on the variability of the sample sizes of the included trials. With little variation in sample size, and no small (n < 20) trials, established tests (2; 3) showed greatly inflated type 1 error, which was corrected by recent proposals (4; 5). By contrast, when the included trials had a range of sample sizes from 6 to 1000 or more, established tests performed not worse. This talk explains these observations, and outlines the advantages of our newly proposed arcsine test.

Methods: Tests on publication bias are based on a funnel plot, which plots standard error versus treatment effect. If the outcome is binary, treatment is often measured by the log odds ratio (logOR). For this measure, the standard error depends on the estimated treatment effect (4). This induces a relationship between the performance of tests for publication bias and the sizes of the trials in the meta-analysis. We discuss the effects and consequences of this.

We further propose using the variance stabilizing property of the arcsine transform, thus replacing logOR by the arcsine risk difference in tests for publication bias. We show how this simple proposal avoids many problems with existing tests.

Results: For fixed sample size and population treatment effects, the log odds ratio is approximately related to its standard error through the hyperbolic cosine. Thus, if a meta-analysis contains a range of study sizes, the final look of the funnel plot arises from a mix of these curves.

We found that excluding small sample sizes, regardless of the treatment effect, systematically removes some of these curves from the mixture and creates an artificial appearance of publication bias. Further, if trial sizes are small, we may get trials with no events in one or possibly both arms. Excluding these also artificially creates the appearance of publication bias and additionally leads to an overestimate of the treatment effect.

Conclusion: Using an arcsine risk difference alleviates the key problems with the log odds ratio and log risk ratio. Specifically, it

- reduces error inflation in tests for publication bias,
- reduces dependence of tests on sample size distribution,
- does not require to exclude trials with no events in one or both groups,
- is not affected by reversion of the outcome direction (in contrast to the risk ratio),
- has a geometric interpretation as an arc length corresponding to the risk difference,
- is easy to calculate.

References
In fractal epidemiology causal analysis is carried out by the use of conventional statistical tools and fractal methods. The latter are necessary for investigating the causal complexity/heterogeneity of studies and their setting. Fractals, ubiquitous in the living world, are statistically self-similar quantities whose dimension is non-Euclidean.

Fractal epidemiology contains an uncertainty-relation in which uncertainty related to the reproducibility of a study \( u_R \) is inversely associated to the simultaneous uncertainty related to the accuracy of the study outcome \( u_O \), that is, the more one reduces \( u_O \) the more \( u_R \) will increase and vice versa: \( u_R \cdot u_O \geq k > 0 \). Both \( u_R \) and \( u_O \), defined in \((0,1]\), are functions of the frequency of fractal sequences in the sequence set that describes the study. Linear and logistic (non-fractal) functions that mimic \( u_R \) and \( u_O \) have been developed. The surrogate functions make it easy to compute \( u_R \) and \( u_O \).

The uncertainty-relation has consequences of some interest: The minimum of \( u_O \) is easy to determine for every study and the corresponding, simultaneous, maximal \( u_{RM} \) follows directly from the uncertainty-relation. \( \Re = 1 - u_{RM} \) becomes a stochastic measure of the reproducibility of the study. For \( \Re \)-values close to one, the study and its outcome may be regarded solid; for small \( \Re \)-values, on the other hand, the study may be considered futile, a solitary phenomenon on the basis of which no firm conclusions can be drawn as to the tested hypothesis. \( \Re \) is a quantity of great value in fractal meta-analysis and prediction. The uncertainty-relation may even be applied to traditional studies. Thus, the outcome of a prospective study presented as easy to reproduce, may well be considered worthless if there is good reason to believe that the study setting is ‘highly contaminated’ with fractals.
Calibrating, Interpreting and Combining the Evidence in Experiments

Kulinskaya E, Morgenthaler S, Staudte RG
Imperial College London, UK

We propose a new approach to defining evidence in experiments. If adopted, this approach makes meta-analysis a much easier task.

Many simple applications of statistics can be transformed onto the probit scale by means of variance stabilizing transformations (vst). Denote a statistic of an appropriate test by $T$ and its vst by $h(T)$. Then under both the null hypothesis and the alternatives, the evidence measure $h(T)$ has a normal distribution with standard deviation one. Thus evidence as defined here is a random quantity with a well known distribution, and it has a standard error of one unit when estimating its expected value. The above statements are only approximately true, but true enough where it matters, for effect sizes and small to moderate sample sizes arising in applications.

Each application requires its own special transformation, and the mathematical level required for applying them is minimal. Examples in the talk include standard 1-sample t-test, Welch t-test for two samples with unequal variances (Kulinskaya and Staudte, 2006, Statistics in Medicine (in press)), and 1- and 2-sample binomial applications.

The standard tasks such as sample size calculation, or finding confidence intervals for a parameter of interest become very easy to perform.

The quality of achieved variance stabilisation and coverage probabilities of resulting confidence intervals are demonstrated by simulation.

It is also very easy to combine the evidence from several studies. Transform observed values of $k$ statistics $T_1,T_2,\ldots,T_k$ and calculate the combined evidence

$$T=\frac{k-1}{2}(h(T_1)+h(T_2)+\ldots+h(T_k))$$

The book is to be published by Wiley in 2007.