Interactive CardioVascular and Thoracic Surgery 27 (2018) 317–321
doi:10.1093/icvts/ivy163 Advance Access publication 2 June 2018

Cite this article as: Barili F, Parolari A, Kappetein PA, Freemantle N. Statistical Primer: heterogeneity, random- or fixed-effects model analyses? Interact CardioVasc Thorac Surg 2018;27:317–21.

Statistical Primer: heterogeneity, random- or fixed-effects model analyses?†

Fabio Barili[a,*], Alessandro Parolari[b], Pieter A. Kappetein[c] and Nick Freemantle[d]

[a] Department of Cardiac Surgery, S. Croce Hospital, Cuneo, Italy
[b] Unit of Cardiac Surgery and Translational Research, IRCCS Policlinico S. Donato, San Donato, Italy
[c] Thoraxcenter, Erasmus MC, Rotterdam, Netherlands
[d] Department of Primary Care and Population Health, University College London, London, UK

* Corresponding author. Department of Cardiac Surgery, S. Croce Hospital, Via M. Coppino 26, 12100 Cuneo, Italy. Tel: +39-017-1642571; fax: +39-017-1642064; e-mail: fabarili@libero.it; barili.f@ospedale.cuneo.it (F. Barili).

Received 30 November 2017; received in revised form 13 April 2018; accepted 17 April 2018

Summary

Heterogeneity in meta-analysis describes differences in treatment effects between trials that exceed those we may expect through chance alone. Accounting for heterogeneity drives different statistical methods for summarizing data and, if heterogeneity is anticipated, a random-effects model will be preferred to the fixed-effects model. Random-effects models assume that there may be different underlying true effects estimated in each trial which are distributed about an overall mean. The confidence intervals (CIs) around the mean include both within-study and between-study components of variance (uncertainty). Summary effects provide an estimation of the average treatment effect, and the CI depicts the uncertainty around this estimate. There are 5 statistics that are computed to identify and quantify heterogeneity. They have different meaning and give complementary information: Q statistic and its *P*-value simply test whether effect sizes depart from homogeneity, T^2 and T quantify the amount of heterogeneity, and I^2 expresses the proportion of dispersion due to heterogeneity. The point estimate and CIs for random-effects models describe the practical implications of the observed heterogeneity and may usefully be contrasted with the fixed-effects estimates.

Keywords: Statistical analysis • Meta-analysis

INTRODUCTION

Meta-analysis is the statistical synthesis of data from related studies, and the results summarize a body of research. Unlike the narrative review, meta-analysis calculates a weighted average treatment effect and its uncertainty [1]. The central unit of meta-analysis is the treatment effect or effect size, which is a measure of the relationship between 2 groups [2]. The effect size can vary across related studies, and the principal goal of the synthesis is the estimation of a summary effect, which is simply a weighted mean of the individual effects. It is also critical to evaluate the robustness of the summary effect, including some expectation on variability among studies and subsequently quantifying it. The observed dispersion of the estimated effect sizes is partly spurious as it always includes a random (or sampling) error inherent in each study, but it may also include a true variation of the effect sizes in each study, namely heterogeneity.

Heterogeneity is the true difference in effect sizes related to intrinsic factors of the studies included in the meta-analysis [2, 3]. Differences in the characteristics of cohorts and in treatment options, together with other reasons, lead to assume that studies will not share a common effect size but will have heterogeneous underlying effects. This assumption on heterogeneity is a critical point when conducting a meta-analysis as it drives different statistical methods for summarizing data and also different interpretation of results. If our understanding is that all studies share the same common effect, we will choose a fixed-effects model; otherwise, if heterogeneity is expected, a random-effects model will be preferred (Fig. 1) [3].

Fixed-effects model

The fixed-effects model assumes that all studies considered in the meta-analysis share the same common true effect size (hence, the term fixed) (Fig. 1A). Differences among observed effects are related to sampling error (ϵ_i; *i* stands for study *i*), and factors influencing the effect size are assumed to be the same in all the studies. There is no heterogeneity ($\zeta_i = 0$) and the variance is completely due to spurious dispersion (within-study variance). The summary effect is the estimate of a common true effect, and the confidence intervals (CIs) depict the uncertainty around this estimate.

Random-effects model

Random-effects models assume that there are different underlying true effects. These true effect sizes are distributed about some

†Presented at the Annual Meeting of the European Association for Cardio-Thoracic Surgery, Vienna, Austria, 7–11 October 2017.

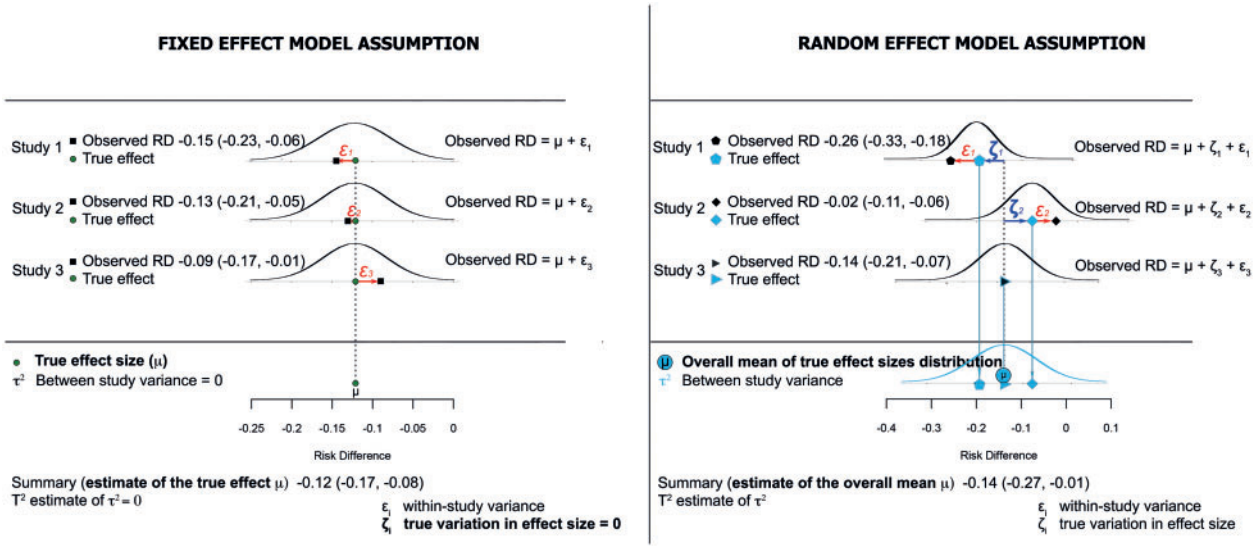

Figure 1: Schematic diagram of the assumption of fixed- and random-effects models. In the fixed-effects model, there is no heterogeneity and the variance is completely due to spurious dispersion. Summary effect is the estimate of the true effect (μ). In the random-effects model, the true effect sizes are different and consequently there is between-studies variance. The summary effect is the estimate of the mean of the distribution of the true effect sizes with an estimated variance of T^2. RD: risk difference.

mean (Fig. 1B) and can be considered as a random sample from a distribution (usually Gaussian)—hence, the term random. Random-effects models are preferred when study cohorts are expected to be different or treatment options are not identical among studies. The variance is accounted by both spurious (within-study variance, ϵ_i) and real dispersion (between-study variance, ζ_i), and a formula is applied to partition it into these 2 components, as the main focus shifts from the summary effect to the identification and quantification of heterogeneity. Summary effects provide an estimation of the average treatment effect, and the CI depicts the uncertainty around this estimate, including the component of heterogeneity [2]. In the presence of heterogeneity, the relative weights are more balanced than those assigned under fixed effects as standard random-effects methods add a common component of variance to each study weight to account for between study variability in treatment effects. Consequently, this double source of variability (within and between study) will lead to wider variance, standard error and CI for the summary effect [2].

For example, we can suppose to conduct a meta-analysis of randomized controlled trials comparing clinical outcomes (30-day mortality and 30-day pacemaker implantation) of adult patients with severe aortic stenosis undergoing either transcatheter aortic valve implantation (TAVI) or surgical aortic valve replacement. Effect sizes can be hypothesized not to be identical across studies as different risk profiles are included and also as different devices were employed. Hence, random-effects model would be preferred.

METHODOLOGY

Under the random-effects model, attention is focused on quantifying heterogeneity and understanding its implications [2]. Specific methodologies are employed to partition the total dispersion, isolate the true variance and give an array of statistics for abstracting the interpretation of results (Fig. 2).

Q statistic (also known as Cochrane's Q)

Q statistic (also known as Cochrane's *Q*) is the weighted sum of squares; more easily, a measure of the total observed dispersion of the estimated effect sizes. It is a standardized value, and it is not affected by the metric of the effect size; hence, it is not a measure of dispersion on the same scale of the effect size (not comparable).

Q — degrees of freedom

Q - df is the part of dispersion related to differences in the true effects (heterogeneity or excess variation). It is calculated by subtracting to *Q* the degrees of freedom (df), which represents the within-study error. It is also a standardized measure.

Test for assumption of homogeneity

Test for assumption of homogeneity is based on *Q* statistics and tests the null hypothesis that all studies share a common effect size. The test performs badly in the small sample setting, and the results are sensitive to the excess of dispersion and the number of studies included as increase of dispersion moves towards significance and an increased number of studies strengthen the evidence of the test. To be noted, a significant *P*-value confirms that the true effects vary while a non-significant *P*-value should be discussed as it depends not only on the robustness of effect sizes but it can also account for low power (small number of studies, wide within study variance). Moreover, the homogeneity test, as well as the *Q* statistic, cannot be employed as an estimate of the amount of heterogeneity, and it simply tests the null hypothesis that all effect sizes are consistent.

T^2 and T-estimates of the variance and standard deviation of the true effect sizes

T^2 is the estimate of the variance of the true effect sizes (τ^2), derived from the observed effects. Different from *Q*, it is expressed

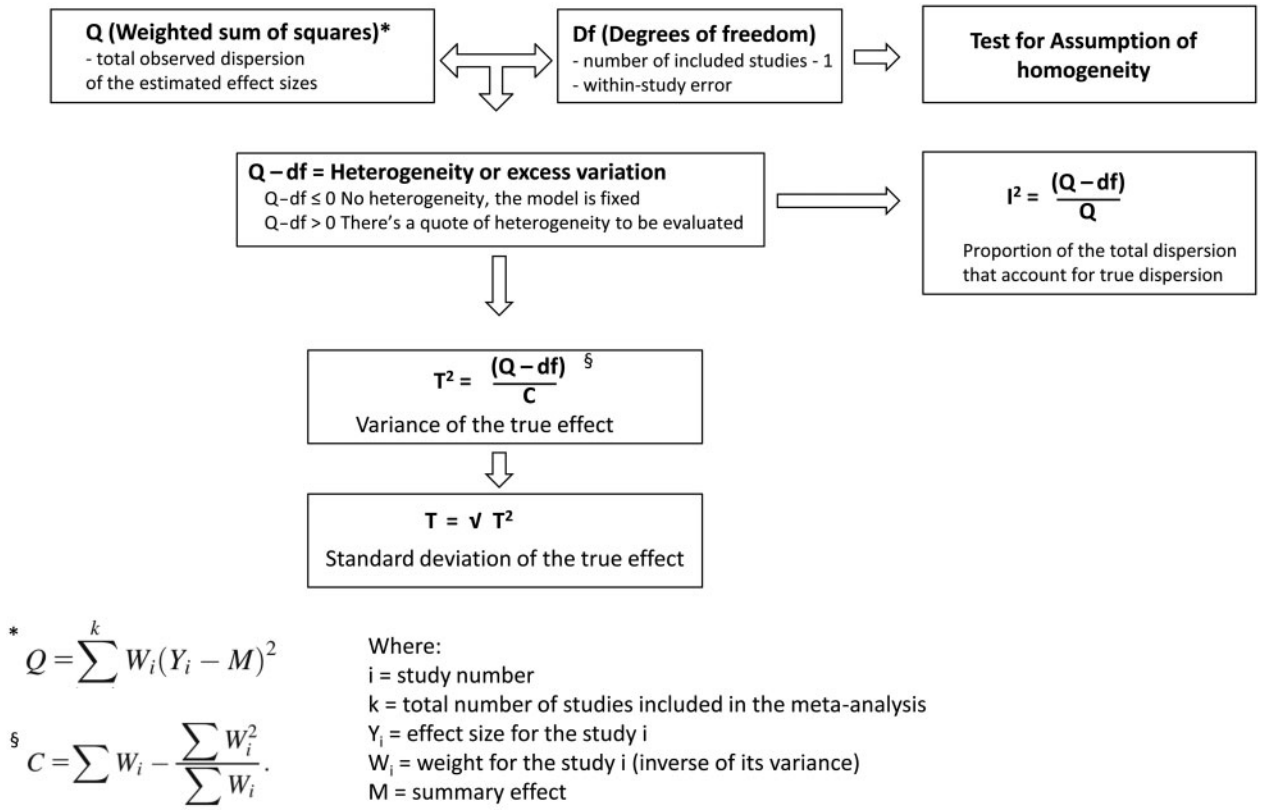

Figure 2: A flowchart of the array of statistics for abstracting the interpretation of results.

in the same metric of the summary effect, and it represents the amount of true dispersion of the effect sizes. The most common method for estimating the between-studies variance in a meta-analysis is the DerSimonian–Laird estimator [4], which is based on the method of moments and may be biased in some settings. T is the square root of T^2 and represents the estimate of standard deviation of the normal distribution (τ) of the true effect sizes. It has the same metric of the summary effect. Assuming a normal distribution of the true effect sizes, it can be used to describe the distribution of the effects around their mean, calculating the 95% CI of the summary effect. Increasing T-values reflect the increased true variance around the mean in the summary estimate.

The I^2 statistic expresses the proportion of the total dispersion that accounts for true dispersion being the ratio between the excess of dispersion and total dispersion. It is calculated on Q, and hence it is not the estimate of an underlying amount but only a descriptive statistic. It is a measure of inconsistency among the findings of the studies, and it is not affected by the number of studies included in the meta-analysis. It was suggested that 25%, 50% and 75% could be considered low, intermediate and high inconsistency, respectively [5]; nonetheless, these cut-offs are simply thresholds of crude guidelines and the evaluation of I^2 statistic should overcome them.

In summary, there are 5 statistics that are computed to identify and quantify heterogeneity. They have different meanings and give complementary information: Q statistic and its P-value simply test whether effect sizes are homogeneous, T^2 and T quantify the amount of heterogeneity, and I^2 expresses the proportion of dispersion due to heterogeneity. A sixth, and potentially much more useful, statistic describing the effects of heterogeneity is the random-effects estimator of the pooled treatment effects.

Common statistical software and languages have functions to estimate heterogeneity. Fixed- and random-effects meta-analyses can be implemented in the R packages 'Meta', 'metafor', 'rmeta' and 'epiR'. A tutorial for conducting meta-analysis with R with the package 'metaphor' is described by Viechtbauer [6]. RevMan 5 is the software developed for preparing and maintaining Cochrane Reviews, and it is possible to choose random- or fixed-effects models while conducting meta-analysis. Macros for conducting meta-analysis in SPSS can be found in the web (e.g. http://mason.gmu.edu/~dwilsonb/ma.html). In Stata, Meta and Metan commands have been developed to generate fixed- and random-effects meta-analysis. The %METAANAL macro is an SAS version 9 macro that produces the DerSimonian–Laird estimators for random- or fixed-effects model.

REPORTING

Meta-analysis should be reported following published guidelines, such as PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and MOOSE (Meta-analysis of Observational Studies in Epidemiology) [7, 8].

Authors should explicitly state the rationale for the choice of the model, underscoring potential sources of variability of the studies included in the meta-analysis. In the results and/or in the forest plot, the evaluation of heterogeneity should be reported, including the Q statistics, the test for assumption of homogeneity, the I^2 statistic and the estimate of the variance of the true effect sizes T^2. The random-effects estimator and CIs describe the importance of heterogeneity in the practical setting. In the discussion, authors should make inference not only on the summary effect but also on the dispersion.

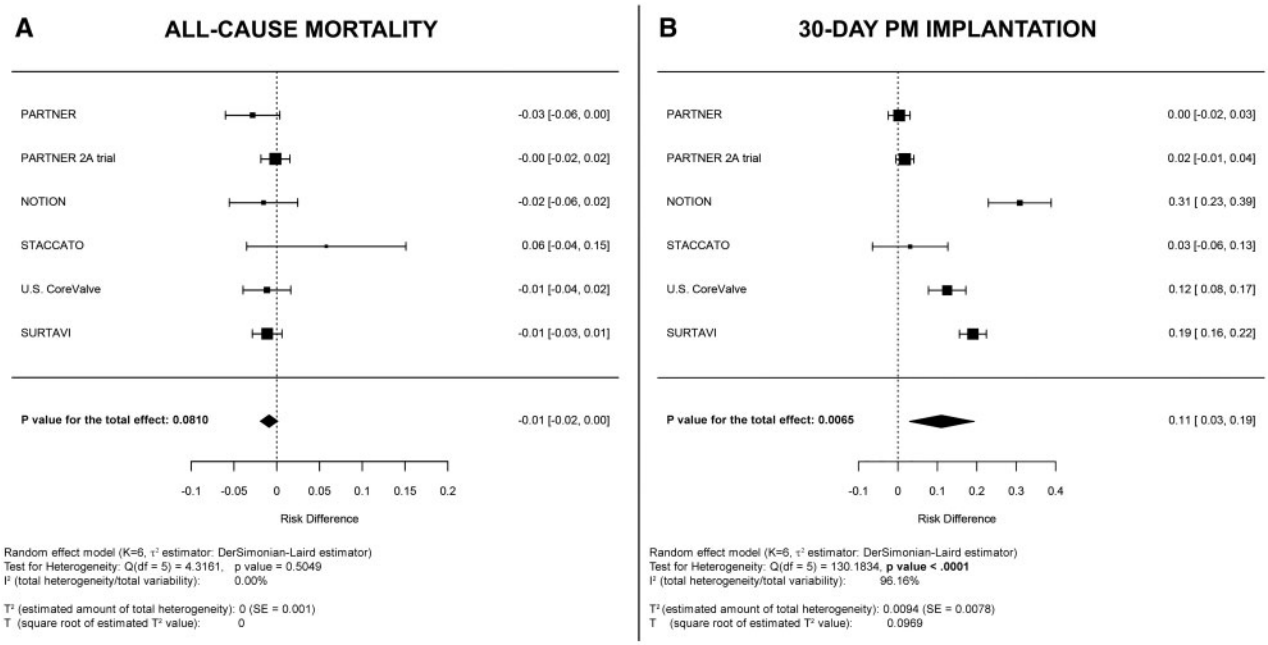

Figure 3: Random-effects meta-analysis of 6 trials that examine the effect of TAVR versus surgical aortic valve replacement on 30-day incidence of mortality (**A**) and pacemaker implantation (**B**). In the forest plot for 30-day mortality, there is no heterogeneity and the random-effects analysis reduces to fixed-effects analysis. (**B**) Heterogeneity is significant and the summary effect is an estimate of the true effect sizes. df: degrees of freedom.

There are some notes to keep in mind. First, a very small number of studies can lead to a poor estimate of heterogeneity. Hence, the random-effects model has been correctly chosen but there is insufficient information for applying it. In this case, one possible option could be to avoid reporting a summary effect as conclusions on effect size and its CI cannot be drawn, or an alternative could be represented by a different approach, such as a Bayesian one, where the extent of heterogeneity maybe inferred through an informative prior. Moreover, the practice of performing a fixed-effects model and subsequently moving to the random-effects model if the test of homogeneity is significant should be discouraged as the choice should be based on hypothesis on common effect sizes and not on a statistical test that often suffers of low power. Differences in the characteristics of cohorts (e.g. different preoperative risk profiles) and in treatments options (such as different devices with potential implementation of interventions), together with other reasons (different ethnicity, geographical variation, etc.), lead to assume that studies will not share a common effect size and should be analysed with the random-effects model. Further, the standard methods for random effects (DerSimonian and Llaird) include a component of variance to describe the between-study variability adaptively, diverging from the fixed-effects model when the *P*-value for heterogeneity is significant. If the random-effects model is chosen and T^2 was demonstrated to be 0, it reduces directly to the fixed effect, while a significant homogeneity test in a fixed-effect model leads to reconsider the motivations at its basis. However, the contrast of the fixed- and random-effects results provides a useful description of the importance of heterogeneity in the results. Finally, the interpretation of random-effects meta-analysis can be implemented by a prediction interval, which is a measure that provides a predicted range for the true treatment effect in an individual study [3]. It resembles reference ranges usually employed in other areas of medicine, such as those for blood pressure or birth weight across the population [3].

EXAMPLE

We can aim to meta-analyse randomized controlled trials comparing 30-day mortality and 30-day pacemaker implantation of adult patients with severe aortic stenosis undergoing either TAVI or surgical aortic valve replacement. We choose to evaluate the risk difference of outcomes between treatment and control groups. The 7 included trials differ in the perioperative risk profiles, as [9, 10] are performed in intermediate-risk, whereas [11–14] have been performed in high-risk patients. Moreover, treatment options are also different because different TAVI devices have been employed across studies [9–14]. These considerations can lead to assume that heterogeneity (between study differences in treatment effects) is anticipated and the random-effects model is preferred.

The analysis of heterogeneity for 30-day mortality demonstrates that trials are homogeneous (Fig. 3A), being the test for assumption of homogeneity (see Methodology section) *P*-value = 0.50 and the percentage of heterogeneity on total variability (I^2) of 0%, suggesting that the variability in study estimates is entirely due to chance. The estimate of the variance of the true effect sizes (T^2) is 0. In this case with no source of heterogeneity and only within-study variance, the random-effects model coincides with the fixed-effects model, as shown in Fig. 4A, and the summary risk difference (-0.009; 95% CI -0.0191 and 0.0011) is the estimate of a common true effect size. The point estimate thus suggests that average mortality under TAVI is 0.9% lower than under surgical aortic valve replacement, but the 95% CIs include a reduction of 1.9% or an increase of 0.1%.

The analysis of heterogeneity for 30-day pacemaker implantation shows significant heterogeneity across studies with the test for assumption of homogeneity with *P*-value <0.0001 (Fig. 3B) and high inconsistency (I^2 96.16%). The estimate of the variance of the true effect sizes (T^2) is 0.0094. The summary risk difference (0.11; 95% CI 0.03–0.19) is the estimation of the mean of the distribution of the effects. As the CI does not contain zero, there is

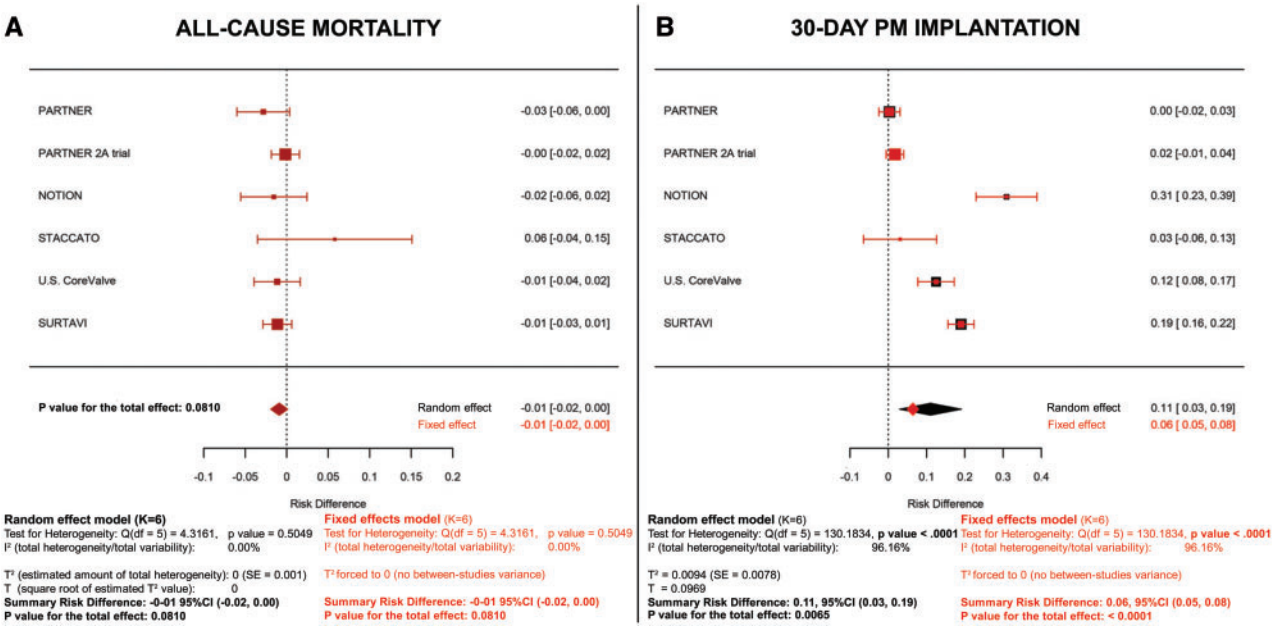

Figure 4: Comparison between random- and fixed-effects models in the example. Both fixed- and random-effects models were applied to the example in order to underscore the differences on estimation. The fixed-effects model is reported in red, and the random-effects model is depicted in black. (**A**) There is coincidence between the 2 models, as heterogeneity is null and the random-effects model is reduced to the fixed-effects model. In the second outcome (**B**), there is a significant heterogeneity and hence different estimates are obtained by applying fixed- or random-effects model, as fixed-effects model does not consider the between-studies variance, and summary estimate is performed by forcing $T^2 = 0$, although it is significant (red). The appropriate choice of random-effects model (black) leads to more balanced relative weights and to wider variance, standard error and confidence interval for the summary risk difference.

good evidence that on average TAVI is related to the increased incidence of 30-day pacemaker implantation. Figure 4B shows the implication of model choice; in random effect, the relative weights are more balanced and the double source of variability led to wider variance, standard error and CI for the summary effect.

CONCLUSIONS

In summary, heterogeneity assessment is an important step in meta-analysis as in many cases the assumption of the same true effect across studies is implausible. Thus random-effects meta-analysis, which accounts for unexplained heterogeneity, will continue to be prominent in the medical literature [3].

Conflict of interest: none declared.

REFERENCES

[1] Fleiss JL. The statistical basis of meta-analysis. Stat Methods Med Res 1993;2:121–45.
[2] Borenstein M, Hedges LV, Higghins JPT, Rothstein HR. Introduction to meta-analysis. Chichester, West Sussex, UK: Wiley, 2009.
[3] Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011;342:d549.
[4] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88.
[5] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557–60.
[6] Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Soft 2010;36:1–48.
[7] Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 2010;8:336–41.
[8] Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D *et al.* Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis of Observational Studies in Epidemiology (MOOSE) group. JAMA 2000;283:2008–12.
[9] Leon MB, Smith CR, Mack MJ, Makkar RR, Svensson LG, Kodali SK *et al.* Transcatheter or surgical aortic-valve replacement in intermediate-risk patients. N Engl J Med 2016;374:1609–20.
[10] Reardon MJ, Van Mieghem NM, Popma JJ, Kleiman NS, Søndergaard L, Mumtaz M *et al.* Surgical or transcatheter aortic-valve replacement in intermediate-risk patients. N Engl J Med 2017;376:1321–31.
[11] Smith CR, Leon MB, Mack MJ, Miller DC, Moses JW, Svensson LG *et al.* Transcatheter versus surgical aortic-valve replacement in high-risk patients. N Engl J Med 2011;364:2187–98.
[12] Thyregod HG, Steinbrüchel DA, Ihlemann N, Nissen H, Kjeldsen BJ, Petursson P *et al.* Transcatheter versus surgical aortic valve replacement in patients with severe aortic valve stenosis: 1-year results from the all-comers NOTION randomized clinical trial. J Am Coll Cardiol 2015;65:2184–94.
[13] Adams DH, Popma JJ, Reardon MJ, Yakubov SJ, Coselli JS, Deeb GM *et al.* Transcatheter aortic-valve replacement with a self-expanding prosthesis. N Engl J Med 2014;370:1790–8.
[14] Nielsen HH, Klaaborg KE, Nissen H, Terp K, Mortensen PE, Kjeldsen BJ *et al.* A prospective, randomised trial of transapical transcatheter aortic valve implantation vs. surgical aortic valve replacement in operable elderly patients with aortic stenosis: the STACCATO trial. EuroIntervention 2012;8:383–9.