

A Large-Scale Assessment of Temporal Trends in Meta-Analyses using Systematic Review Reports from the Cochrane Library

Thomas R. Fanshawe

Luke F. Shaw

Graeme T. Spence

Nuffield Department of Primary Care Health Sciences, University of Oxford

Abstract

Introduction: Previous studies suggest many systematic reviews contain meta-analyses that display temporal trends, such as the first study's result being more extreme than later studies', or a drift in the pooled estimate. We assessed the extent and characteristics of temporal trends using all Cochrane intervention reports published 2008-2012.

Methods: We selected the largest meta-analysis within each report and analysed trends using methods including a Z-test (first versus subsequent estimates); generalised least squares (GLS); and CUSUM charts. Predictors considered include meta-analysis size and review group.

Results: Of 1,288 meta-analyses containing at least four studies, the point estimate from the first study was more extreme and in the same direction as the pooled estimate in 738 (57%), with a statistically significant difference (first versus subsequent) in 165 (13%). GLS indicated trends in 717 (56%); 18% of fixed-effects analyses had at least one violation of CUSUM limits. For some methods, meta-analysis size was associated with temporal patterns and use of a random-effects model, but there was no consistent association with review group.

Conclusions: All results suggest more meta-analyses demonstrate temporal patterns than would be expected by chance. Hence, assuming the standard meta-analysis model without temporal trend is sometimes inappropriate. Factors associated with trends are likely to be context-specific.

Keywords: Generalised least squares; Cumulative meta-analysis; Proteus phenomenon; Quality control charts; Temporal trend

Introduction

Standard meta-analytic methods aim to estimate pooled effect sizes that do not change over time. For this reason, there is an implicit assumption that there is no temporal pattern in the available evidence, as indicated by effect size estimates from the individual contributing studies. There is a growing body of evidence to support the claim that this assumption is unrealistic, as revealed by examining single meta-analyses or groups of meta-analyses within a particular clinical specialty (Bagos and Nikolopoulos, 2009; Jennions and Møller, 2002; Trikalinos et al., 2004).

Ioannidis and Trikalinos made perhaps the most vigorous attempt to demonstrate the existence of temporal trends or patterns in published meta-analyses in a paper of 2005, in which they used the term 'Proteus phenomenon' for the scenario in which the first published study exhibits an effect size showing a more favourable effect of a new intervention than is found in later studies (Ioannidis and Trikalinos, 2005). Many other forms of temporal trend may also arise, although they currently lack a clear taxonomy. Examples include a gradual (linear, or at least monotone) decline in the observed effect size over time, either across the whole meta-analysis or during a certain time-frame within it (Herbeck et al., 2012); and 'outlying' effects from individual studies that do not concord with evidence from other studies, even allowing for a plausible level of heterogeneity (von Dadelszen et al., 2000).

This issue becomes even more relevant in the light of recent developments in the field of cumulative meta-analysis, and the growing popularity of updating meta-analyses sequentially as new study results are published, as an adjunct to performing a periodic full update of the systematic review (Chalmers and Lau, 1993; Lau et al., 1995). The process of undertaking a review update each time a new study is published has also been referred to as 'living systematic review' (Elliott et al., 2014). Cumulative meta-analysis is a particularly effective way to visualise temporal trends in pooled effects.

In spite of the number of papers that have sought to address the issue of temporal trends in published meta-analyses, the prevalence of the phenomenon remains largely unknown, and factors associated with these trends remain unidentified. Noticeably, many of the published non-methodological assessments focus on just one or a small number of meta-analyses and conclude that there is evidence of a temporal trend, to a greater or lesser extent (Elvik, 2011; Tu et al., 2008). It is less clear whether temporal trends exist more widely, or whether they only occur in the areas that have been previously highlighted. In this paper therefore we aim to address this issue using a comprehensive and complete sample of systematic review reports and their meta-analyses published within the Cochrane Library. Our intention is to assess the scope of the phenomenon at large, without restricting our inclusion criteria to a particular clinical area or class of interventions.

Some previous studies have used information from the Cochrane Library's Database of Systematic Reviews to answer similar questions. For example, one study examined a small subset of 254 reviews, published before 1998 and updated before 2002, to identify factors relating to changes in conclusions when systematic reviews are updated (French et al., 2005). A more recent publication used a subgroup of reviews from the Cochrane Pregnancy and Childbirth Group to compare methods for identifying reviews that were out-of-date and needed updating (Pattanittum et al., 2012). The Cochrane Database therefore provides a rich source of data from which to investigate patterns in meta-analysis on a large scale.

The overall objectives of our study are: (i) to assess the extent of temporal trends in meta-analysis results using a large sample of systematic review results from the Cochrane Library; (ii) to identify factors associated with the presence of temporal trends; and (iii) to review and appraise statistical methods suitable for detecting temporal trends.

Methods

Detecting temporal trends in meta-analysis

A standard meta-analysis models the effect sizes y_i from individual studies i ($= 1, \dots, n$) as being drawn independently and at random from a given distribution, usually a Normal distribution $N(\theta_i, \sigma_i^2)$, with mean and variance to be estimated. In the fixed-effects model, a common value of $\theta_i = \theta$ is assumed for each study, while in the random-effects model θ_i are considered to be independently and identically distributed $N(\theta, \tau^2)$, where θ is the pooled effect size of interest and τ^2 is the heterogeneity parameter. In comparisons between two groups, we consider $\theta = 0$ to represent the point of 'no intervention effect' – corresponding, for example, to a mean difference or a log relative risk of zero. Under this very widely used model, summary estimates are time-invariant, in the sense that a hypothetical change in the order in which the studies occur will not change the pooled parameter estimates.

In this paper we use the term 'temporal trend' to refer to any scenario in which: (i) there is a violation of the assumption that the results of studies contributing to a meta-analysis are independent and distributed as described in the preceding paragraph, and (ii) the deviation away from this assumption can be explained in terms of the order in which the studies were published. We consider time to be represented by the order of publication, rather than the year of publication, and use the term 'meta-analysis size' for the number of studies in a meta-analysis. Further issues relating to time to statistical significance in cumulative meta-analysis have been discussed elsewhere (Berkey et al., 1996).

The patterns of statistical significance that arise within cumulative meta-analyses have been previously categorised into five qualitatively distinct classes (Trikalinos et al., 2004):

- (i) No statistically significant pooled estimate was observed in the whole trajectory (i.e. significance '*never reached*');
- (ii) An earlier statistically significant pooled estimate remained statistically significant up to and including the last-meta analysis ('*retained*');
- (iii) A statistically significant pooled estimate was obtained for the first time at the final meta-analysis ('*only at end*');
- (iv) An earlier statistically significant pooled estimate was no longer statistically significant at the last meta-analysis ('*lost, not regained*');
- (v) The pooled estimate was statistically significant at the final meta-analysis and at some earlier occasion, but at least one non-statistically significant pooled estimate occurred in between ('*lost, regained*').

As more studies are added, the cumulative meta-analysis can move through the categories in a specific order, for example from '*lost, not regained*' to '*lost, regained*' after the addition of a study that results in a return to significance for the pooled estimate.

Beyond classifying the trajectory of cumulative meta-analyses, there are three broad categories of statistical methods for investigating specific types of temporal trend.

The first of these relates to methods suitable for detecting the Proteus phenomenon, which can be expressed as θ_1 arising from a distribution with mean not equal to θ . This might be seen as a 'step change' between the results of the first and subsequent studies, and is easily extended to a group of more than one initial study. This phenomenon has been observed in a variety of settings, both generally (Ioannidis, 2005; Ioannidis, 2006) and for more specific clinical areas and outcomes, including perinatal medicine and myocardial infarction (Ioannidis and Lau, 2001). There is a variety of plausible reasons for this phenomenon, including but not restricted to simple publication bias (early non-statistically significant results are less likely to be published) and the possibility that extreme or statistically significant results of a new treatment are published more urgently than those that show a modest or no effect (Stern and Simes, 1997; Dickersin et al., 2002), either owing to the authors or to editorial processes (Olson et al., 2002).

Formal and informal methods can be used to assess Proteus phenomenon-type effects. Most simply, the point estimate of the first study result ($\hat{\theta}_1$) can be compared to the final pooled effect size ($\hat{\theta}$) from all studies; under the standard meta-analytic assumption, $P(\hat{\theta}_1 > \hat{\theta}) = 0.5$. More formally, a hypothesis test can be conducted using the Z-test given by equation (5) of Bago and Nikolopoulos (2009), in which the estimate from the first study is compared to the pooled estimate from all subsequent studies; under the null hypothesis, these two estimates are independent realisations from distributions with the same means.

The second category of methods attempts to detect a constant temporal trend across the trajectory of a cumulative meta-analysis, and one such approach is described by Bago and Nikolopoulos (2009). This regression method uses generalised least squares (GLS) and adjusts for artefactual correlation between pooled cumulative estimate induced by including similar sets of study results at each time point. The parameter of interest represents the gradient of a fitted trend line, which can be tested against a null value of zero. In this study we classify non-significant slopes as '*flat*', and significant slopes as '*diminishing*' (if the fitted line moves towards the point of no effect, $\theta = 0$), '*diverging*' (away from $\theta = 0$) or '*crossing*' (crosses $\theta = 0$). In contrast with the Proteus phenomenon, this classification attempts to identify gradual or systematic drift in point estimates. Possible explanations include delays in publication time of non-statistically significant results and a tendency for later studies to test successful interventions in subpopulations where their effectiveness is likely to be reduced. Other regression-based methods to detect temporal trends include regression of the point estimates from individual studies (Gehr et al., 2006) and Bayesian shrinkage (Baker and Jackson, 2010).

The third category is grounded in a substantial body of work relating to quality control and related cumulative-sum (CUSUM) and \bar{X} (also known as Shewhart) chart methodology (Kang, 2011). Its application to cumulative meta-analysis has previously been described in detail (Kulinskaya and Koricheva, 2010). The main objective of the \bar{X} chart is to identify 'outlying' or aberrant study results, which might occur either in an individual study or in a run of consecutive studies, while the CUSUM chart is aimed at detecting a shift in the mean level. The basis of both methods is to construct a series of control limits around a target level of effect size. These limits depend on the precision of the results of each individual study (as determined by the sample size), and a violation occurs when either a single study result or a run of study results falls outside the limits (Lucas, 1985).

Unlike the first two categories of methods, quality control-type methods have an element of subjectivity, in that they require the user to define a 'target level' for the pooled estimate and/or to impose a measure of stringency on the width of the control limits, commonly expressed as a fixed multiple of the standard error. In these methods, it can be conceptually and methodologically difficult to disentangle aberrant study results from those that may be regarded as contributing to high, but plausible, between-study heterogeneity. As the heterogeneity variance parameter (τ^2) is rarely known in advance, it must be estimated and re-estimated as part of the model fitting procedure, which is especially difficult during the first part of a cumulative meta-analysis. For these reasons, quality control methods are generally used only for fixed-effects meta-analyses (Kulinskaya and Koricheva, 2010).

Acquisition and preparation of the data files

We identified intervention reviews that were published or substantially updated (given a new citation record) between January 2008 and December 2012 from the relevant issues

within the Cochrane Library (e.g. www.cochranelibrary.com/cochrane-database-of-systematic-reviews/table-of-contents/2008/issue1). The latest version of a review was selected if it appeared more than once between 2008 and 2012, and reviews that had no associated data file were excluded. The Cochrane review group responsible for each review was also obtained (e.g. from www.cochrane.org/CD000038), and this entire process was done in an automated fashion ('web scraping') using the R package *rvest* (Wickham, 2016). Data from the reviews that had associated data files were exported from their '.rm5' files (obtained from Cochrane) into '.csv' format using Review Manager (version 5.3).

We imported each of the CSV files into R using the *read.rm5* function from the *meta* package (Schwarzer, 2015). Reviews that contained no numerical data, those that identified just one study for inclusion in all proposed meta-analyses, and those for which the authors decided not to pool any study results using meta-analysis were excluded.

Most reviews include more than one meta-analysis as they investigate more than one outcome variable and/or different subgroups. For each review, we identified the meta-analysis that contained the largest number of primary studies. We considered only meta-analyses in which the authors pooled study results: for example, if the authors pooled results within separate subgroups but did not pool across the subgroups, we selected the largest subgroup. Meta-analyses in which authors had presented results from different subgroups within a single study but subsequently (incorrectly) pooled the subgroup results as if they came from different studies were ignored.

In some cases, the data files contained ambiguous or missing data for the year in which a primary study was conducted; when this occurred, we corrected this information using the study label, which customarily includes the study date (e.g. 'Smith 2010'). If there were inconsistencies between the study label and the study year field in the data file, we used the year given in the study label, usually the date of the primary publication for the study as identified by the authors of the Cochrane review. If year information was unobtainable, even after manually checking the reference list of original review, the review was excluded. For reviews whose largest meta-analysis size occurred for two or more analyses, we chose the first to appear in the data-file on the basis that this was more likely to be the primary outcome. If a meta-analysis contained two or more studies from the same year, these were assumed to have occurred in the same order as they appeared in the data file, consistent with a sequential approach to analysis in which new study results might become available one at a time. Our final processed data file therefore contained a single meta-analysis from each eligible review, with the individual studies within each meta-analysis ordered by publication date.

The selected largest meta-analyses were coded as a triplet of numbers, 'comparison-outcome-subgroup', as recorded in the corresponding columns of the data files. Files include the summary measure used by the review authors, usually one of Risk Ratio/Relative Risk (RR), Odds Ratio (OR), Hazard Ratio (HR), Rate Ratio, Risk Difference (RD), Mean

Difference (MD), or Standardised Mean Difference (SMD). When other summary measures were reported and could not be recoded directly (for example, converting the free-text string 'Peto Odds Ratio' to 'OR'), two authors independently checked the original review and, by consensus, either recoded as one of the main seven summary measures or classified it into a single separate category ('Other'). We also retained information about the first and last year of the studies in each meta-analysis (the difference between these being termed the 'year range'), the review group, the total number of participants in the analysis, and the method of pooling adopted (Mantel-Haenszel, Inverse Variance or Peto).

Statistical methods

First, we summarised the complete set of largest meta-analyses descriptively. For implementation of methods relating to temporal trend, only meta-analyses containing at least four studies with estimable treatment effects were analysed as we judged that there would be insufficient scope to assess temporal trends if the number of studies was very small. We performed standard and cumulative meta-analysis, with the same model specification (summary measure and choice of fixed or random effects) adopted by the authors of each review, using the *metacr* and *metacum* functions in the *meta* package of R, respectively (Schwarzer, 2015).

We implemented the following methods:

- Binary comparison of point estimates of first study and final meta-analysis and Z-test to compare the first study with the subsequent pooled estimate;
- Classification of cumulative meta-analysis results into the five categories described above, according to their patterns of statistical significance;
- The GLS method (Bagos and Nikolopoulos, 2009), using the *xtgls* function of Stata/SE 14.1 for Windows;
- For fixed-effects meta-analyses only, the CUSUM and \bar{X} methods, with the control limits for the CUSUM chart set at \pm five standard deviations, and the control limits for the \bar{X} chart set at the target level \pm three standard deviations. The control limits are those commonly applied (Kulinskaya and Koricheva, 2010). The target level for the \bar{X} chart was set at the point estimate of the cumulative estimate after four studies and applied both retrospectively to the first four studies and to all subsequent studies. The choice of target level is pragmatic, as it would be inappropriate to use the same target across studies of different types that use different outcome measures, and may not correspond to the level chosen in a fully prospective sequential meta-analysis. For this reason, all meta-analyses of size four or more are analysed. The choice of target level is addressed further in the Discussion.

The results of statistical tests were recorded in a summary data file, which is included as Supplementary Material.

We carried out additional tests to assess the association between the existence of a temporal pattern within a meta-analysis (the classification into which the cumulative meta-analysis fell, for the methods listed above) and key characteristics of the meta-analysis. We investigated the relationship between the classification of cumulative meta-analysis patterns and review group using the chi-squared test, and adjusted this association for meta-analysis size, number of participants, year range, model type (fixed or random effects) and summary measure using logistic regression.

A similar approach was used for the classifications of the GLS model, and with the existence of violations using the CUSUM and \bar{X} methods, as the outcomes in separate logistic regression models. As many review groups had very few eligible meta-analyses, we primarily investigated review groups that contributed at least 30 meta-analyses, with those from other review groups combined into a single reference category for the purposes of model fitting. Similarly, for the control chart methods we investigated review groups that contributed at least 30 fixed effects meta-analyses.

Finally, we estimated the magnitude of change in effect size by converting the estimates of the slopes from the GLS model to estimates of effect size per year, and used these as measures of change in clinical effectiveness.

Results

Of 3,195 eligible systematic reviews identified in the Cochrane Library, 2,008 had a meta-analysis meeting our inclusion criteria, and 1,288 of these (40% of the total) had a largest eligible meta-analysis containing at least four studies (Figure 1). Within this group, which is the basis of the analysis below, the authors used a fixed effects model in 705 cases (55%) and a random effects model otherwise.

Table 1 shows the characteristics of the included meta-analyses. In the analysed sample, a typical meta-analysis included seven primary studies containing a total of almost 1000 participants, spanning fourteen years. The majority of analyses contained ten or fewer studies (Figure 2). Relative summary measures, such as the relative risk (691 analyses, 54%), were more frequently used than absolute measures, and the choice of measure (Supplementary Figure 1) and meta-analysis model (i.e. fixed or random, Supplementary Figure 2) varied widely between review groups, of which the Pregnancy and Childbirth (117 analyses, 9%) formed the largest subgroup.

In 916 (71%) of the 1,288 meta-analyses, the point estimate of the first study was more extreme (further from the null value, corresponding to no effect) than the final pooled estimate, and both were on the same side of the null in 1028 cases (80%, including 47 cases in which the first point estimate equalled the null value). These numbers can be further broken down as '*More extreme/Same side*' 738 (57%), '*More extreme/Different side*' 178

(14%), '*Less extreme/Same side*' 290 (23%) and '*Less extreme/Different side*' 82 (6%). In the absence of a temporal trend the percentage in the '*More extreme/Same side*' category should not exceed 50%; in our dataset it was 57.3%, 95% confidence interval: [54.5, 60.0]. Furthermore, 165 meta-analyses (13% of the total) showed a statistically significant difference between the first and subsequent estimates, as measured by the p -value from the Z-test, well above the 5% that would be expected by chance alone.

In the classification of studies into the five patterns of statistical significance, 501 (39%) were classified as '*Retained*', 431 (33%) as '*Never reached*', 170 (13%) as '*Lost, regained*', 154 (12%) as '*Lost, not regained*' and 32 (2.5%) as '*Only at end*' (see Figure 3 for examples of each pattern). The tendency for studies to fall into these categories was strongly related to the number of studies in the meta-analysis (Figure 4). For example, the '*Lost, regained*' category requires the most 'events' to occur and has the largest average size, although there was considerable overlap between categories (Supplementary Figure 3).

The distribution of the five categories varied by review group (Figure 5; $p=0.007$, chi-squared test). For the '*Lost, not regained*' category, this effect remained even after adjustment for meta-analysis size and model type (Supplementary Table 1), but this was not substantially the case for the '*Lost, regained*' category, for which the associations with meta-analysis size and use of a random-effects model were stronger (Supplementary Table 2). Additional adjustment for number of participants, time range and summary measure did not affect the overall conclusions. The occurrence of the '*Lost, not regained*' category was noticeably high in the Neonatal review group (Figure 5) and statistically significantly higher than average after adjusting for review size and model type (Supplementary Table 1), with a statistical significant increase also observed for the Pregnancy and Childbirth group.

The GLS model had unstable parameter estimates in seven meta-analyses, which are therefore excluded from the following results. Of the remaining 1,281 meta-analyses, 717 (56%) possessed a trend, as indicated by a statistically significant estimate of the slope coefficient, of which 414 (32% of the total) were classified as '*Diminishing*', fewer (135 (11%)) were classified as '*Diverging*' and 168 (13%) were classified as '*Crossing*' (see Figure 6 for an illustration of the four qualitative patterns of temporal trend). Hence, 57.7% (414/717, 95% confidence interval: [54.0%, 61.4%]) of those that showed a statistically significant trend had a fitted regression line that was directed towards the null effect and did not cross the null line. '*Crossing*' was more frequent in smaller than in larger meta-analyses, but '*Diminishing*' patterns were more common than '*Diverging*' patterns irrespective of meta-analysis size.

Figure 7 shows the relative frequency of each by review group ($p=0.64$, chi-squared test). In a logistic regression of '*Diminishing*' (outcome variable) against review group, adjusting for meta-analysis size and model type, the Stroke review group was the only statistically significant explanatory variable, indicating significantly lower occurrence of the pattern in that group (Supplementary Table 3).

A comparison between classification methods can be found in Supplementary Table 4. Of the 152 meta-analyses classified as 'Lost, not regained', 86 (57%) have a '*Diminishing*' slope in the GLS model (compared with 32% for all meta-analyses), while only 1 (0.7%) is '*Diverging*' (compared with 11% overall). Similar relationships are observed for the 735 meta-analyses classified as '*More extreme/Same side*' for the comparison between first and final estimates: 365 (50%) are '*Diminishing*', whilst only 21 (3%) are '*Diverging*'.

Estimates of effect size per year, using the estimated GLS slopes, are shown in Table 2, separately for studies that used RR, OR and SMD. Effect sizes for raw mean differences cannot be compared between studies because of differences between units. As the direction of the effect depends on the way the outcome is specified, all effect sizes have been converted to the 'positive' direction (ratios greater than 1 and positive SMDs). Average effect size changes per year are larger for meta-analyses showing a '*Diminishing*' trend than for those showing a '*Diverging*' trend. For '*Diminishing*' meta-analyses, the median change in RR per year was 3.2% and the median change in OR per year was 4.6%. Additionally, the proportion with a change in RR of at least 10% per year was much higher for '*Diminishing*' meta-analyses than for '*Diverging*' meta-analyses (31/213 (15%) versus 3/73 (4%)).

The CUSUM and \bar{X} methods are illustrated for a single example by Supplementary Figure 4, which shows that violations may occur for both methods simultaneously (e.g. an \bar{X} upper violation and a CUSUM upper violation at Study 18) or for one method but not the other (e.g. an \bar{X} violation at Study 9). Of the 705 fixed-effects meta-analyses, 126 (18%) contained at least one CUSUM violation and 165 (23%) contained at least one \bar{X} violation, and 87% (109/126) of meta-analyses that had a CUSUM violation also had an \bar{X} violation (Supplementary Table 5). The proportions of studies with zero, one, or more than one violation, split by review group, are shown in Supplementary Figures 5 and 6. After controlling for meta-analysis size, the Gynaecology and Fertility; Neonatal; and Pain, Palliative and Supportive Care review groups were all associated with a statistically significantly higher occurrence of one or more \bar{X} violations, and the Pain, Palliative and Supportive Care group was also significant for CUSUM violations (Supplementary Tables 6 & 7). Of the smallest meta-analyses (those with only 4 or 5 studies), 7% (15/212) contained at least one \bar{X} violation and only one contained a CUSUM violation.

Discussion

In this paper we have assessed the extent and characteristics of temporal trends in a large sample of meta-analyses published in the Cochrane Library. We have drawn from a very large evidence base, ensured high methodological standards and replicated a realistic cumulative meta-analysis approach by following the analytic decisions chosen by the review authors as closely as possible. Moreover, by including all of the intervention reviews

published in the Cochrane Library from a specific time period, we have used a broad scope of all potential clinical areas and avoided the possibility of selective reporting.

In all of the methods investigated, more trends were observed than would be expected by chance in the absence of any true temporal effects. Specifically, 57% of the first study results were more extreme and in the same direction as the final pooled estimate, and 13% of first studies exhibited results that were statistically significantly different from those of the subsequent studies. Of the meta-analyses with statistically significant GLS trends, 58% had slopes that fitted a diminishing effect, and there was evidence of both outliers (in 18% of meta-analyses) and temporal changes (in 23% of meta-analyses) using quality control charts.

Among meta-analyses that showed a statistically significant trend, the estimates of effect size per year were typically small although, for example, there were several examples of trends in RR corresponding to changes of more than 10% per year. In principle, information of this kind might be used in deciding priorities for systematic review updates, although this is just one consideration in a topic with many facets, as discussed in a recent overview article (Garner et al., 2016). Likewise, following Barrowman et al. (2003), information about the temporal trend might be considered in deciding when to update null meta-analyses, especially if the trend is diverging in nature.

When classifying the patterns of statistical significance within cumulative meta-analysis, 12% displayed significance that was subsequently lost and not regained in multiple hypothesis testing. The proportions of the different patterns of statistical significance are roughly similar to those seen in a previously published subset of mental health trials (Trikalinos et al., 2004), and any differences are consistent with our study accessing later issues of the Cochrane Library. This led, for example, to more interim analyses per meta-analysis (a mean of 11, compared with 7 in the previous study), which may explain the increased proportion in the '*Lost, regained*' at the expense of '*Only at end*' and '*Never reached*'.

Although our analysis showed some associations with specific review groups, these were not consistent between different methods, and it is unclear why the Pain, Palliative and Supportive Care group, for instance, should be associated with a higher proportion of control chart violations than the other review groups. The review group classification in isolation may be too broad a measure to indicate specific reasons for temporal patterns, as studies falling within a review group may differ quite substantially in their interventions and outcome measures. The only factors that were consistently associated with the presence of temporal patterns was the number of studies in the meta-analysis, which is to be expected as larger meta-analyses are likely to have greater power and are more affected by multiple testing; and the use of a random-effects model, which the original study authors may have been more likely to choose if they observed high heterogeneity.

The relationships we discovered from applying different methods to the same meta-analyses were largely as expected – that is, a strong but not perfect association. For example, meta-analyses whose first point estimate is more extreme than the final pooled estimate are more likely to, but do not inevitably, show a diminishing linear trend across the whole temporal range. Practitioners who wish to adopt one or more of these methods should therefore be aware of the type of pattern that they set out to detect, perhaps with the help of the rough categorisation of methods outlined in the Methods section of the current paper.

This study has also revealed some limitations with certain methodological approaches. The GLS method (Bagos and Nikolopoulos, 2009) appears not to adequately make allowance for the precision of pooled estimates, as highlighted by the lowest panel of Figure 6. It is highly likely to give statistically significant results if point estimates follow a near-linear trend even if their precision is low, and this seems more likely to occur by chance for smaller meta-analyses. However, we do not think this is likely to affect our main findings, as ‘*Diminishing*’ trends were more common than ‘*Diverging*’ trends across all meta-analysis sizes. Additionally, a non-linear trend may be more appropriate in certain cases, although this is more a practical disadvantage of our need to apply a single model to a large selection of meta-analyses than a drawback of the method in itself. This consideration also guided our decision to classify trends as ‘*Diminishing*’ or ‘*Diverging*’ relative to the null, or as ‘*More extreme*’ or ‘*Less extreme*’, rather than express trends in terms of the direction of the beneficial effect of the intervention, which may be more desirable for single case studies.

A similar collective decision was necessary in setting target and control levels for the \bar{X} and CUSUM charts. In individual cases, this target level would preferably be set at the ‘true’ treatment effect or at a clinically important value. While the final meta-analysis pooled result would be the most reliable available estimate of this effect size, this would introduce a high level of data-dependence, as the same study results would be used in estimating the effect size and in retroactively assessing whether any of them are outliers. For this reason, we instead chose the pooled effect after four studies, to balance the amount of dependence with the desire to mimic a quality control monitoring process. The dependence is likely to be small for large meta-analyses, and because of low power the \bar{X} and CUSUM methods rarely lead to violations in small meta-analyses for any reasonable choice of the target level.

In this study we chose to analyse the largest meta-analysis within each systematic review. While this was objective and facilitated the application of cumulative analysis methods, it may not always be the most appropriate choice. The chosen outcome might not be primary, or even clinically important, and may have particularly low statistical power.

There are a number of possible causes of temporal trends in meta-analysis, of which publication bias provides just one explanation. There is ample evidence that publications that contain extreme results tend to be published faster than those whose results are more modest (Dickersin et al., 2002; Stern and Simes, 1997; Suñé et al., 2013). Regression to the

mean could also be responsible for initial extreme results being followed by a shift to a smaller pooled effect if researchers are more likely to perform further studies of interventions that demonstrate more eye-catching early estimates. This appears more likely if meta-analyses are updated sequentially, as a living systematic review (Higgins et al., 2011; Spence et al., 2016), and further work in this area in the context of potentially informative sample size would be welcome. This connection with sequential updating guided our decision to use publication order, rather than publication year, as the unit of analysis.

Additionally, trends in the treatment effect could be the result of changes in study populations. For example, an intervention might be studied initially in populations suspected of showing the greatest response, before expanding to more diverse or less responsive populations. Sometimes systematic review authors might treat these different subgroups, but in many meta-analyses there is likely to be some level of clinical heterogeneity, even if a fixed-effects model is chosen (Gagnier et al., 2012).

Regardless of the cause, the presence of temporal trends in a significant proportion of meta-analyses with a substantial bias towards effect sizes that diminish over time could lead to a systematic over-estimation of effect sizes. This finding should persuade researchers to exercise caution when drawing conclusions from meta-analyses with relatively few studies, and encourage review authors to investigate trends using cumulative meta-analysis to identify possible causes.

Acknowledgements

This paper presents independent research funded by a National Institute of Health Research (NIHR) Research Methods Fellowship (GS) and Internship (LS) (NIHR-RMFI-2015-05-015). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. We thank the Cochrane Collaboration for access to the Review Manager files. We also thank two anonymous reviewers and an associate editor for comments that helped to improve the clarity of the paper.

References

- Bagos, P. G. & Nikolopoulos, G. K. (2009). Generalized least squares for assessing trends in cumulative meta-analysis with applications in genetic epidemiology. *Journal of Clinical Epidemiology*, **62**, 1037-1044.
- Baker, R. & Jackson, D. (2010). Inference for meta-analysis with a suspected temporal trend. *Biometrical Journal*, **52**, 538-551.
- Barrowman, N. J., Fang, M., Sampson, M. & Moher, D. (2003). Identifying null meta-analyses that are ripe for updating. *BMC Medical Research Methodology*, **3**, 13.

- Berkey, C. S., Mosteller, F., Lau, J. & Antman, E. M. (1996). Uncertainty of the time of first significance in random effects cumulative meta-analysis. *Controlled Clinical Trials*, **17**, 357-371.
- Chalmers, T. C. & Lau, J. (1993). Meta-analytic stimulus for changes in clinical trials. *Statistical Methods in Medical Research*, **2**, 161-172.
- Dickersin, K., Olson, C. M., Rennie, D. & et al. (2002). Association between time interval to publication and statistical significance. *JAMA*, **287**, 2829-2831.
- Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P. T., Mavergames, C. & Gruen, R. L. (2014). Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. *PLoS Med*, **11**, e1001603.
- Elvik, R. (2011). Publication bias and time-trend bias in meta-analysis of bicycle helmet efficacy: A re-analysis of Attewell, Glase and McFadden, 2001. *Accident Analysis & Prevention*, **43**, 1245-1251.
- French, S. D., McDonald, S., McKenzie, J. E. & Green, S. E. (2005). Investing in updating: how do conclusions change when Cochrane systematic reviews are updated? *BMC Medical Research Methodology*, **5**, 33.
- Gagnier, J. J., Moher, D., Boon, H., Beyene, J. & Bombardier, C. (2012). Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature. *BMC Medical Research Methodology*, **12**, 1-15.
- Garner, P., Hopewell, S., Chandler, J., MacLehose, H., Akl, E. A., Beyene, J., Chang, S., Churchill, R., Dearness, K., Guyatt, G., Lefebvre, C., Liles, B., Marshall, R., Martínez García, L., Mavergames, C., Nasser, M., Qaseem, A., Sampson, M., Soares-Weiser, K., Takwoingi, Y., Thabane, L., Trivella, M., Tugwell, P., Welsh, E., Wilson, E. C. & Schünemann, H. J. (2016). When and how to update systematic reviews: consensus and checklist. *BMJ*, **354**.
- Gehr, B. T., Weiss, C. & Porzsolt, F. (2006). The fading of reported effectiveness. A meta-analysis of randomised controlled trials. *BMC Medical Research Methodology*, **6**, 1-12.
- Herbeck, J. T., Müller, V., Maust, B. S., Ledergerber, B., Torti, C., Di Giambenedetto, S., Gras, L., Günthard, H. F., Jacobson, L. P., Mullins, J. I. & Gottlieb, G. S. (2012). Is the virulence of HIV changing? A meta-analysis of trends in prognostic markers of HIV disease progression and transmission. *AIDS*, **26**, 193-205.
- Higgins, J. P. T., Whitehead, A. & Simmonds, M. (2011). Sequential methods for random-effects meta-analysis. *Statistics in Medicine*, **30**, 903-921.
- Ioannidis, J. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, **294**, 218-228.
- Ioannidis, J. P. A. (2006). Evolution and Translation of Research Findings: From Bench to Where? *PLoS Clinical Trials*, **1**, e36.
- Ioannidis, J. P. A. & Lau, J. (2001). Evolution of treatment effects over time: Empirical insight from recursive cumulative metaanalyses. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 831-836.
- Ioannidis, J. P. A. & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, **58**, 543-549.
- Jennions, M. D. & Møller, A. P. (2002). Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society of London B: Biological Sciences*, **269**, 43-48.
- Kang, C. W. (2011). *Basic statistical tools for improving quality*. Hoboken: Wiley.
- Kulinskaya, E. & Koricheva, J. (2010). Use of quality control charts for detection of outliers and temporal trends in cumulative meta-analysis. *Research Synthesis Methods*, **1**, 297-307.
- Lau, J., Schmid, C. H. & Chalmers, T. C. (1995). Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. *Journal of Clinical Epidemiology*, **48**, 45-57.

- Lucas, J. M. (1985). Cumulative sum (cusum) control schemes. *Communications in Statistics - Theory and Methods*, **14**, 2689-2704.
- Olson, C. M., Rennie, D., Cook, D. & et al. (2002). Publication bias in editorial decision making. *JAMA*, **287**, 2825-2828.
- Pattanittum, P., Laopaiboon, M., Moher, D., Lumbiganon, P. & Ngamjarus, C. (2012). A Comparison of Statistical Methods for Identifying Out-of-Date Systematic Reviews. *PLoS ONE*, **7**, e48894.
- Schwarzer, G. (2015). meta: General Package for Meta-Analysis. R package version 4.3-0. <https://CRAN.R-project.org/package=meta>.
- Spence, G. T., Steinsaltz, D. & Fanshawe, T. R. (2016). A Bayesian approach to sequential meta-analysis. *Statistics in Medicine*, doi: 10.1002/sim.7052.
- Stern, J. M. & Simes, R. J. (1997). Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ*, **315**, 640-645.
- Suñé, P., Suñé, J. M. & Montoro, J. B. (2013). Positive Outcomes Influence the Rate and Time to Publication, but Not the Impact Factor of Publications of Clinical Trial Results. *PLoS ONE*, **8**, e54583.
- Trikalinos, T. A., Churchill, R., Ferri, M., Leucht, S., Tuunainen, A., Wahlbeck, K. & Ioannidis, J. P. A. (2004). Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *Journal of Clinical Epidemiology*, **57**, 1124-1130.
- Tu, Y.-K., Tugnait, A. & Clerehugh, V. (2008). Is there a temporal trend in the reported treatment efficacy of periodontal regeneration? A meta-analysis of randomized-controlled trials. *Journal of Clinical Periodontology*, **35**, 139-146.
- von Dadelszen, P., Ornstein, M. P., Bull, S. B., Logan, A. G., Koren, G. & Magee, L. A. (2000). Fall in mean arterial pressure and fetal growth restriction in pregnancy hypertension: a meta-analysis. *The Lancet*, **355**, 87-92.
- Wickham, H. (2016). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.2. <https://cran.r-project.org/web/packages/rvest/index.html>.

Table 1: Summary characteristics of the full sample (consisting of the largest valid meta-analysis in each systematic review) and the analysed sample (those with at least four studies in the largest valid meta-analysis). Values are presented as number (percentage) or median [inter-quartile range].

	Full sample	Analysed sample
Number of systematic reviews	2008	1288
Number of studies per largest meta-analysis	5 [3 to 9]	7 [5 to 13]
< 4	664 (33.1%)	0 (0.0%)
4 – 5	390 (19.4%)	359 (27.9%)
6 – 10	527 (26.2%)	505 (39.2%)
11 – 20	293 (14.6%)	292 (22.7%)
21+	134 (6.7%)	132 (10.2%)
Number of participants per largest meta-analysis	603 [216 to 1665]	993 [468 to 2642]
Year range	10 [5 to 17]	14 [8 to 20]
Review Group (ordered)		
Pregnancy and Childbirth	185 (9.2%)	117 (9.1%)
Neonatal	99 (4.9%)	61 (4.7%)
Airways	74 (3.7%)	56 (4.3%)
Gynaecology and Fertility	86 (4.3%)	54 (4.2%)
Pain, Palliative and Supportive Care	72 (3.6%)	53 (4.1%)
Stroke	72 (3.6%)	51 (4.0%)
Kidney and Transplant	57 (2.8%)	48 (3.7%)
Colorectal Cancer	54 (2.7%)	45 (3.5%)
Heart	59 (2.9%)	45 (3.5%)
Hepato-Biliary	66 (3.3%)	43 (3.3%)
Acute Respiratory Infections	71 (3.5%)	42 (3.3%)
Gynaecological, Neuro-oncology and Orphan Cancer	59 (2.9%)	39 (3.0%)
Anaesthesia, Critical and Emergency Care	48 (2.4%)	36 (2.8%)
Common Mental Disorders	48 (2.4%)	34 (2.6%)
Summary measure		
Relative Risk	1036 (51.6%)	691 (53.6%)
Odds Ratio	363 (18.1%)	230 (17.9%)
Mean Difference	360 (17.9%)	197 (15.3%)
Standardised Mean Difference	164 (8.2%)	119 (9.2%)
Hazard Ratio	44 (2.2%)	23 (1.8%)
Risk Difference	25 (1.2%)	20 (1.6%)
Rate Ratio	9 (0.4%)	6 (0.5%)
Other	7 (0.3%)	2 (0.2%)
Method of pooling		
Mantel-Haenszel	1234 (61.5%)	810 (62.9%)
Inverse	660 (32.9%)	402 (31.2%)
Peto	114 (5.7%)	76 (5.9%)
Model		
Fixed Effects	1211 (60.3%)	705 (54.7%)
Random Effects	797 (39.7%)	583 (45.3%)

Table 2: Rate of change of effect size estimates per year, expressed as median [inter-quartile range].

	RR	OR	SMD
Diminishing	1.032 [1.015 to 1.069]	1.046 [1.022 to 1.100]	0.037 [0.015 to 0.087]
Diverging	1.019 [1.011 to 1.042]	1.016 [1.011 to 1.036]	0.022 [0.015 to 0.038]

Figure Legends

Figure 1. Systematic review and meta-analysis flowchart.

Figure 2: Distribution of the number of studies in each of the analysed meta-analyses. Note: differing bin widths on the x-axis.

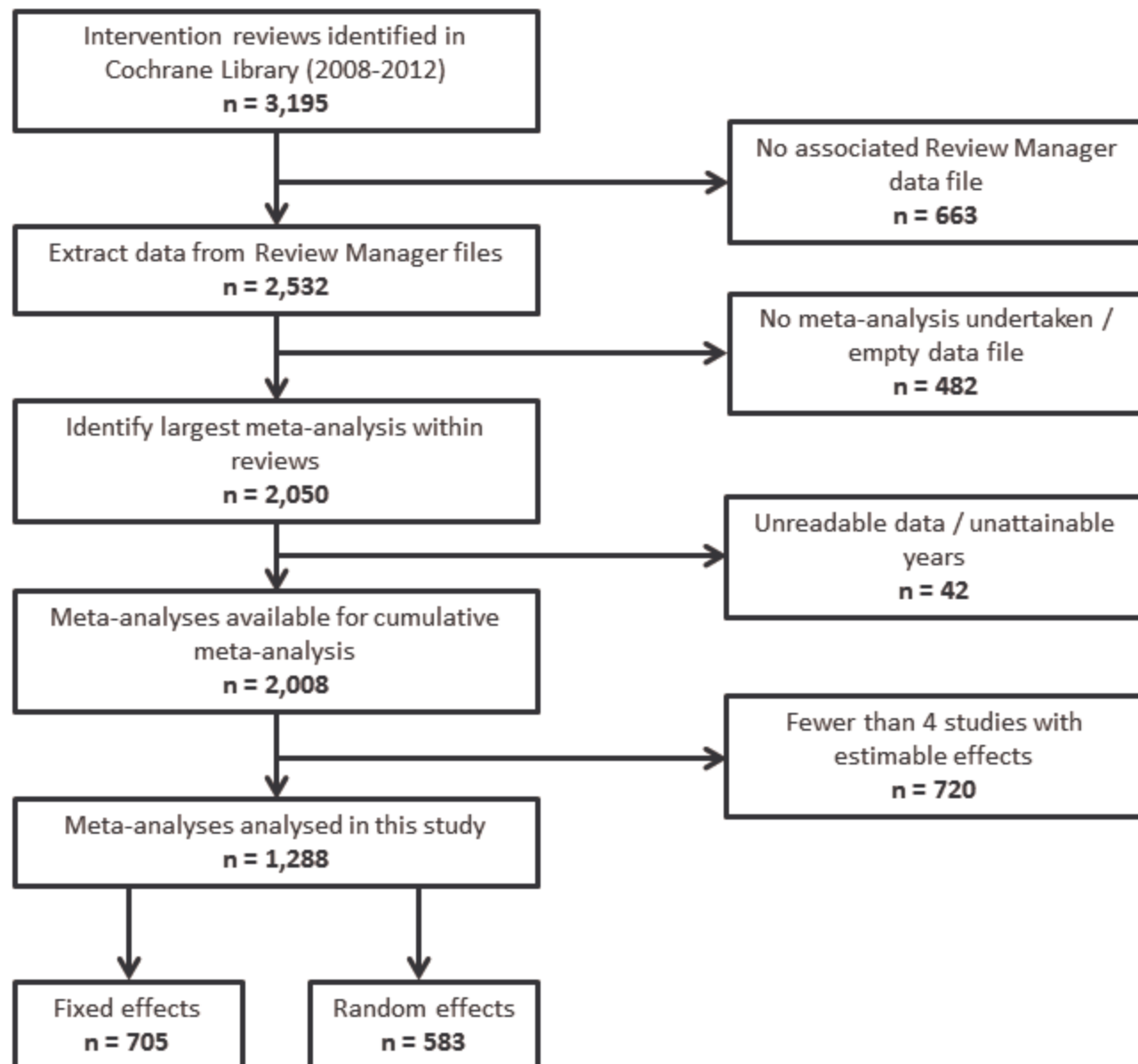
Figure 3: Illustration of the five statistical significance patterns of cumulative meta-analyses. 'Never reached' (CD000051), 'Only at end' (CD000025), 'Retained' (CD000067), 'Lost, not regained' (CD008099), 'Lost, regained' (CD002026).

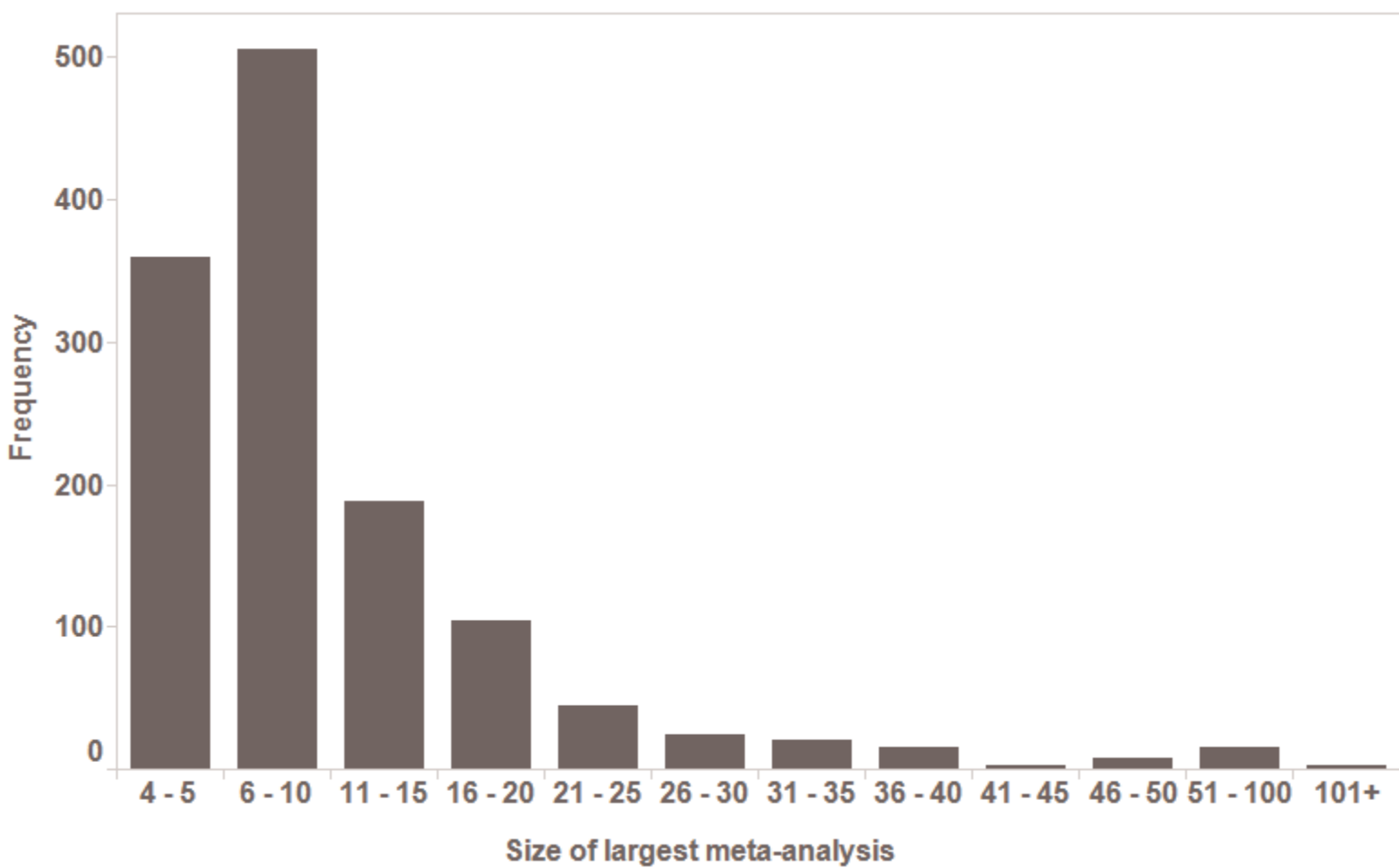
Figure 4: Relationship between total number of meta-analyses and average size of meta-analysis (number of studies) for each of the five categories of statistical significance (see text for definitions).

Figure 5: Relative frequency of five patterns of statistical significance, by review group and size of meta-analysis.

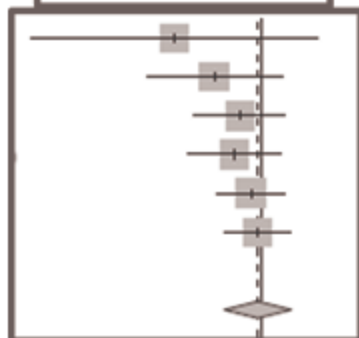
Figure 6: Illustration of the four qualitative patterns of temporal trend that can be derived from the GLS model. A: 'Flat' (CD005370), B: 'Crossing' (CD006574), C: 'Diminishing' (CD001100), D: 'Diverging' (CD000193), E: 'Flat' (CD002067), F: 'Diminishing' (CD004685).

Figure 7: Relative frequency of four qualitative patterns of results from the GLS analysis, by review group and size of meta-analysis.

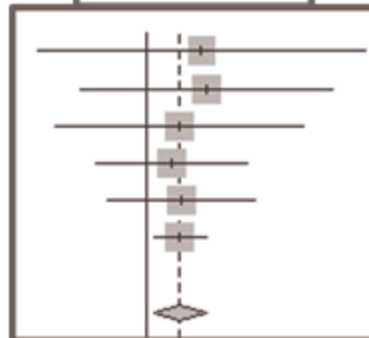




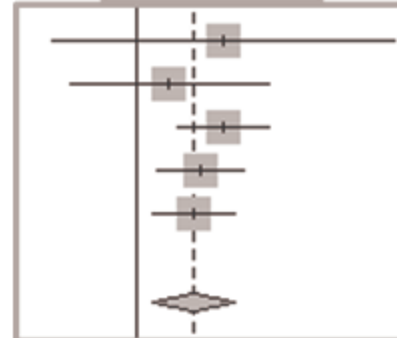
Never reached



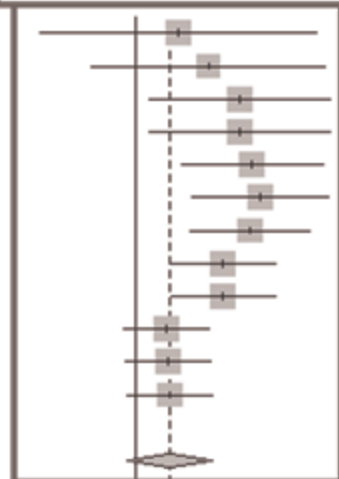
Only at end



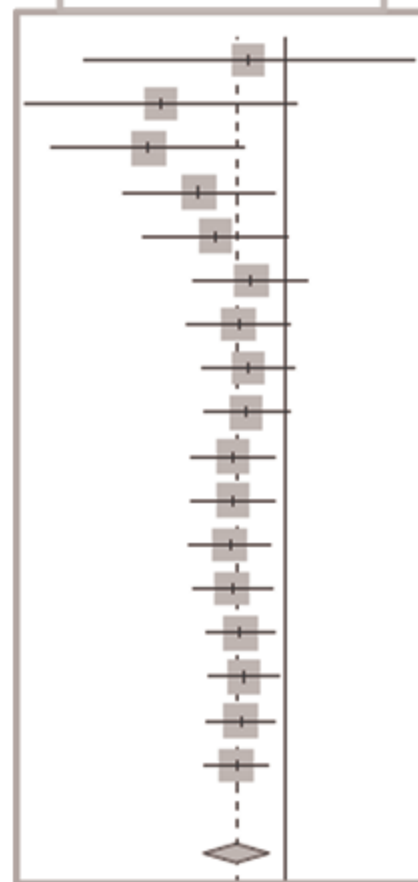
Retained



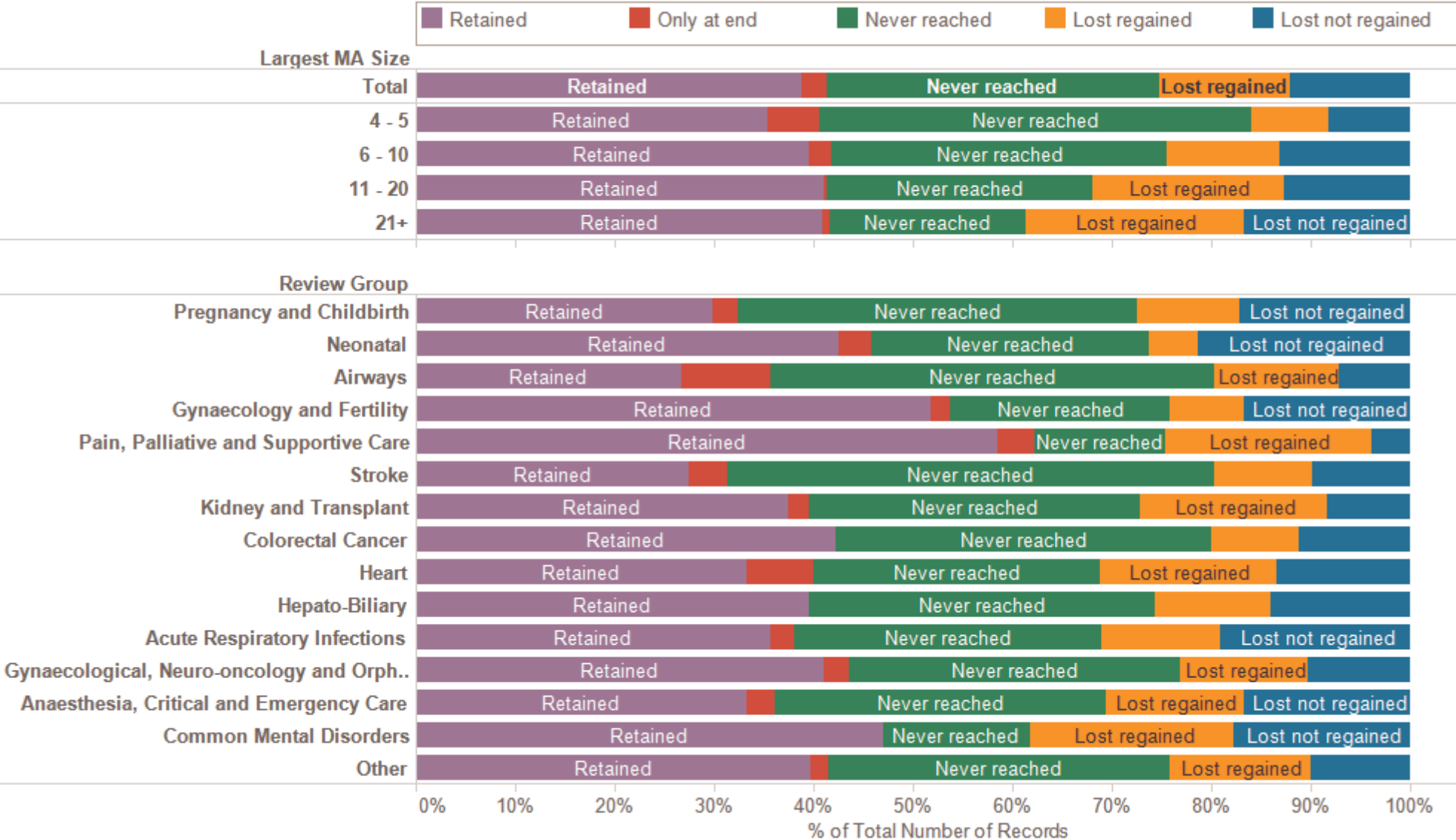
Lost not regained

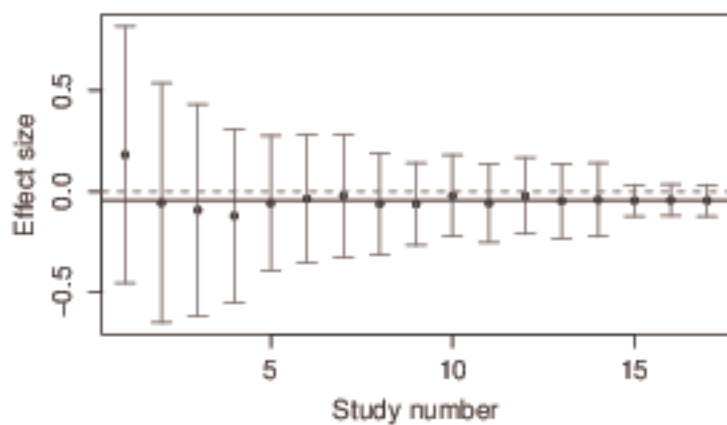
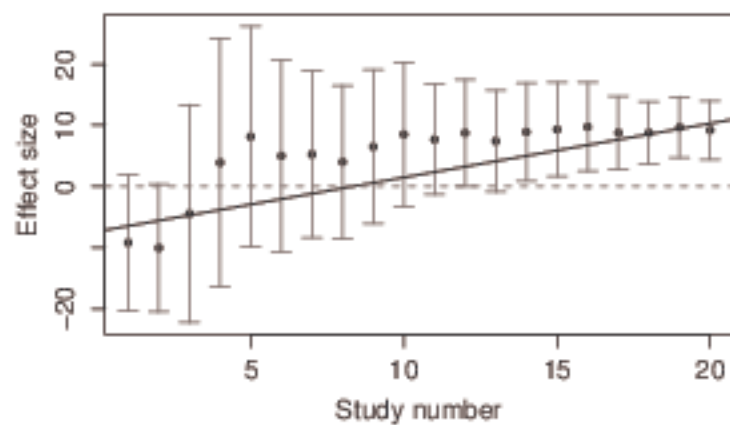
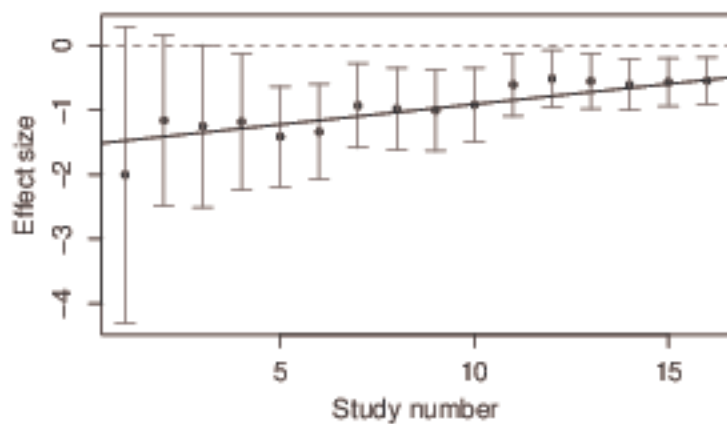
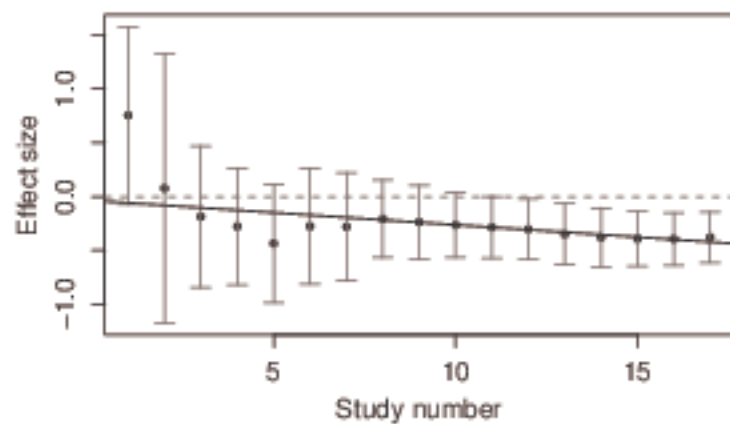
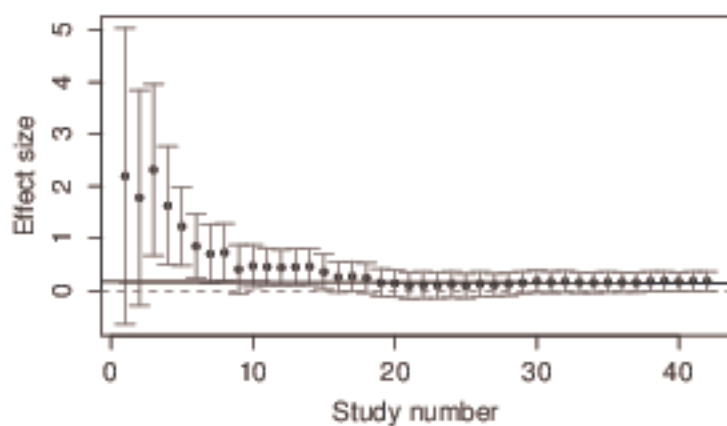
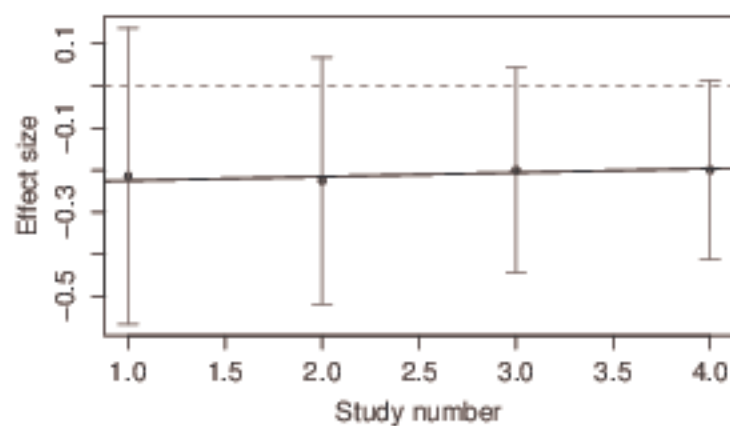


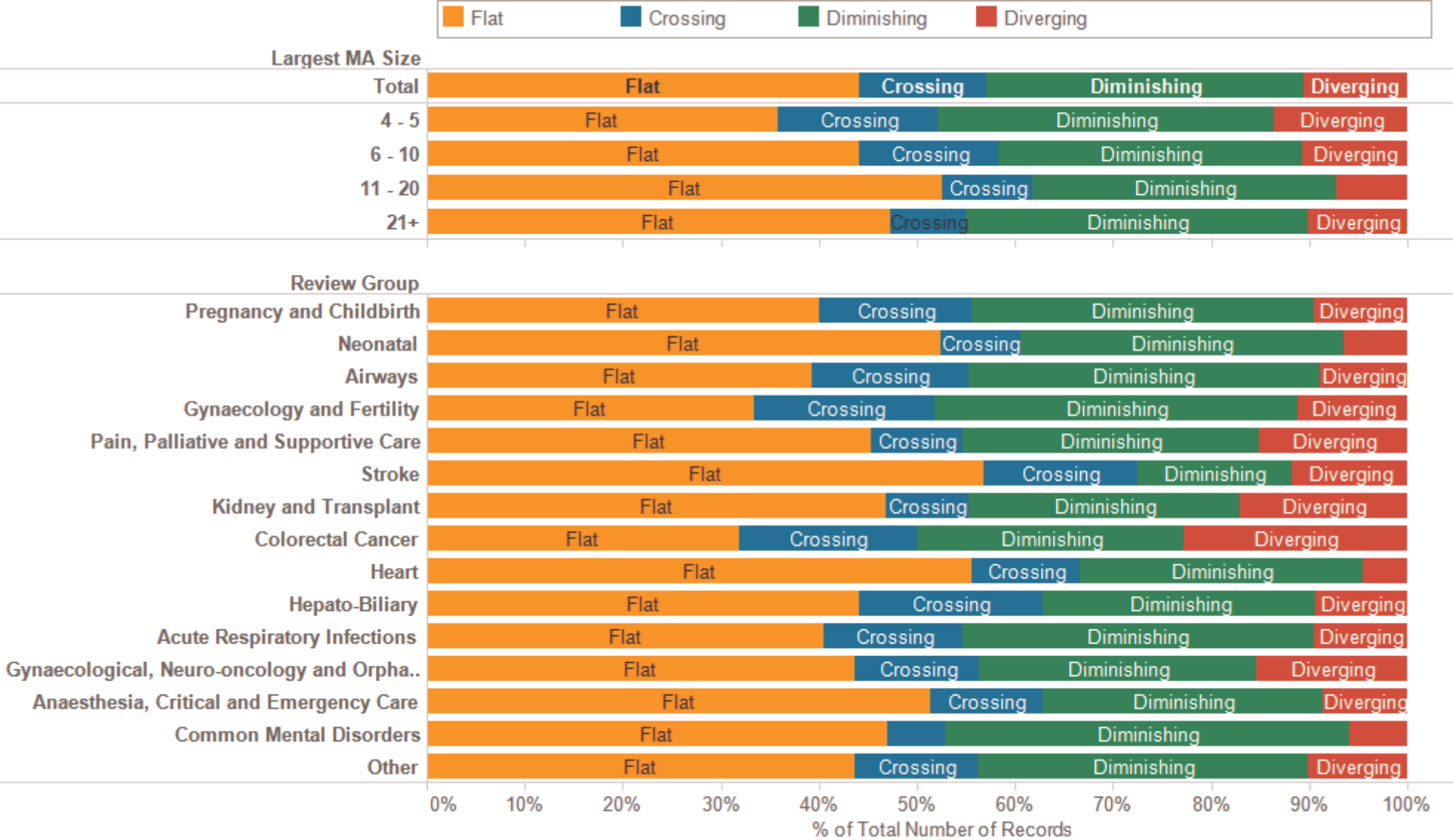
Lost regained







A**B****C****D****E****F**



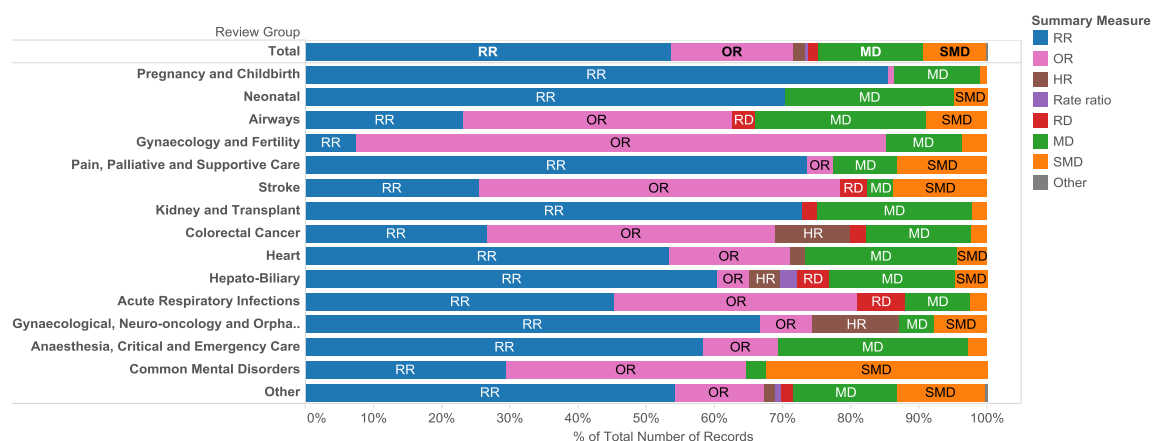
A Large-Scale Assessment of Temporal Trends in Meta-Analyses using Systematic Review Reports from the Cochrane Library – Supplementary Material

Thomas R. Fanshawe

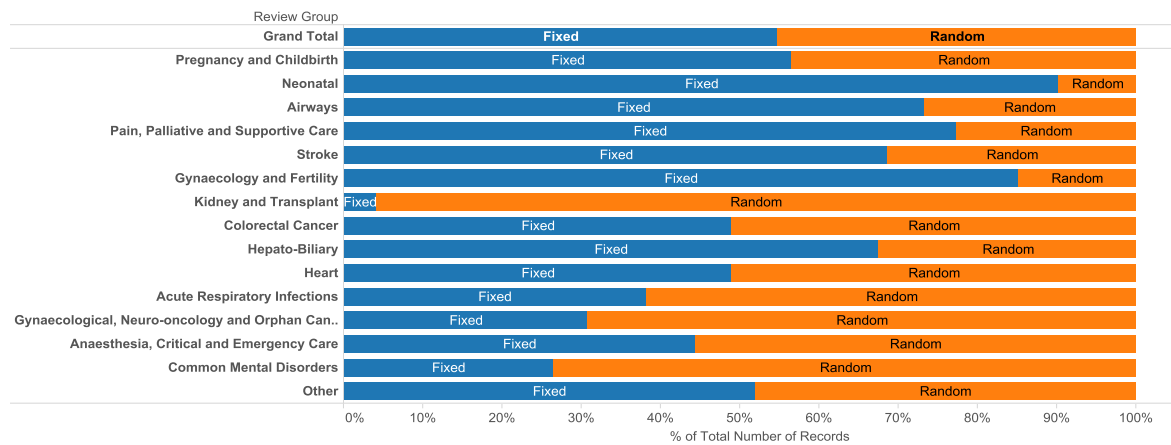
Luke F. Shaw

Graeme T. Spence

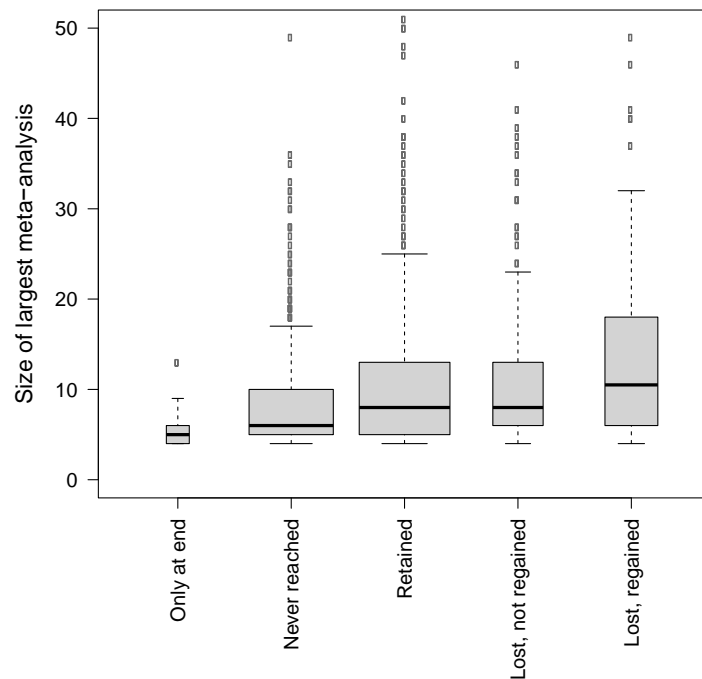
Nuffield Department of Primary Care Health Sciences, University of Oxford



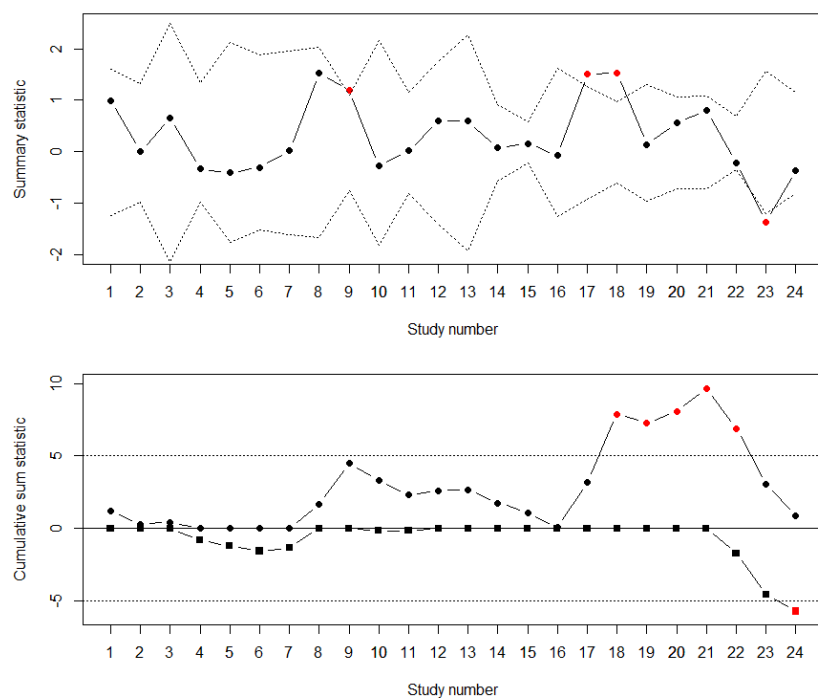
Supplementary Figure 1: Relative frequency of use of each summary measure, by review group.



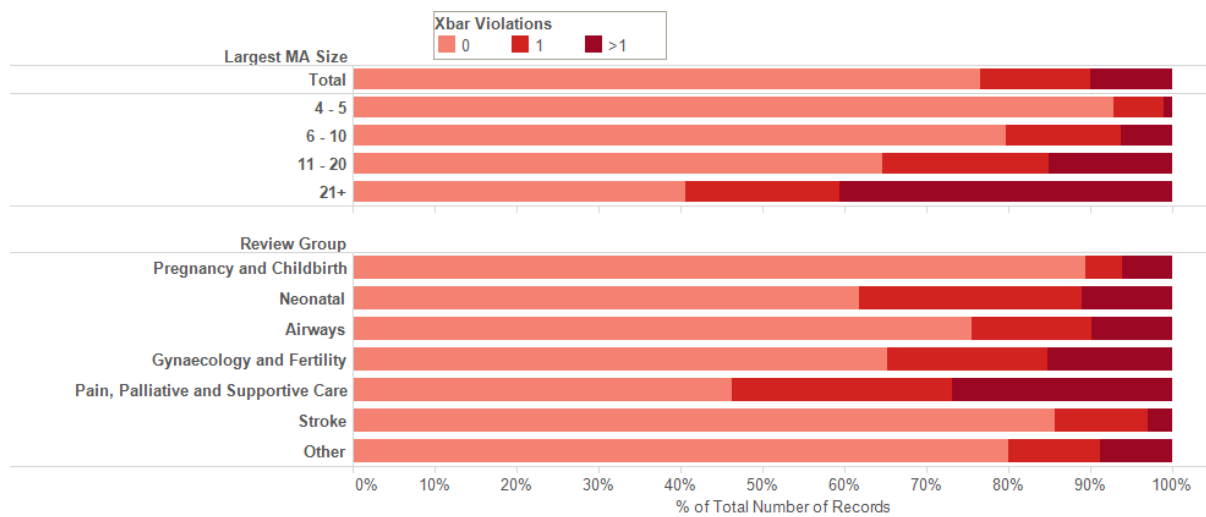
Supplementary Figure 2: Relative frequency of use of fixed and random effects meta-analysis models, by review group.



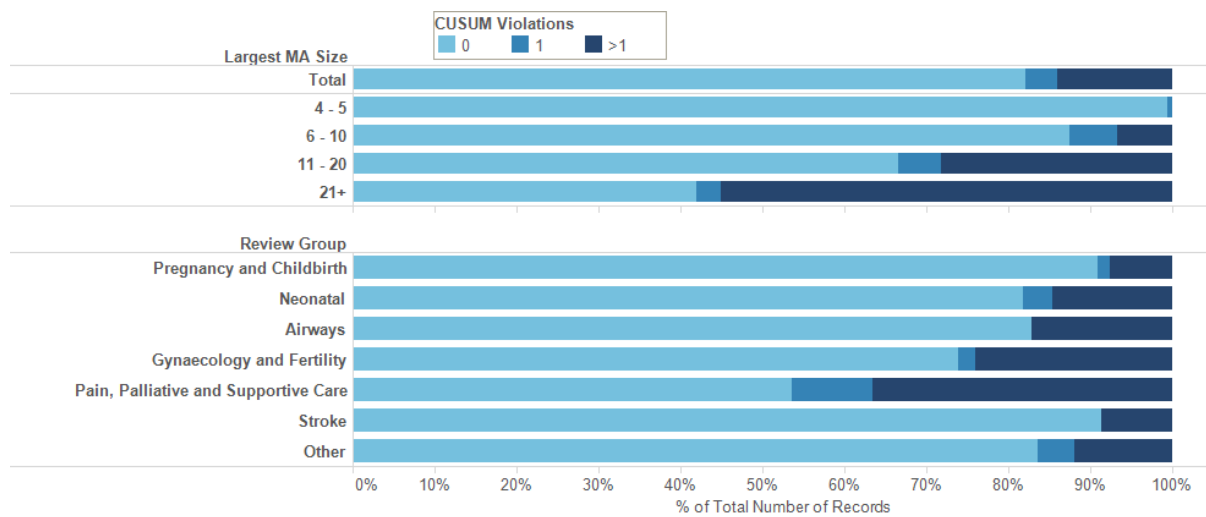
Supplementary Figure 3: Boxplots of the size of meta-analysis for the five categories of statistical significance defined in the main text. The central lines correspond to the sample medians, the boxes represent the interquartile ranges (IQR), and the points outside the whiskers are defined as outliers (greater than 1.5 x IQR away from the box edges). The width of each box is proportional to the square root of the number of meta-analyses in the relevant category.



Supplementary Figure 4: \bar{X} (top) and CUSUM (bottom) quality control charts for CD004872.



Supplementary Figure 5: Relative frequency of CUSUM violations, by review group and size of meta-analysis.



Supplementary Figure 6: Relative frequency of \bar{X} violations, by review group and size of meta-analysis.

Supplementary Table 1: Results of logistic regression for the outcome of 'Lost, not regained' cumulative meta-analysis classification. Asterisks mark coefficient significance levels.

Explanatory variables	OR	95% Confidence Interval	p-value
MA size (coefficient expressed per 5 studies)	1.04	[0.97, 1.10]	0.23
Model: random effects	1.60	[1.11, 2.32]	0.01 *
Acute Respiratory Infections	2.01	[0.83, 4.38]	0.10
Airways	0.77	[0.23, 1.99]	0.63
Anaesthesia, Critical and Emergency Care	1.74	[0.63, 4.12]	0.24
Colorectal Cancer	1.13	[0.38, 2.76]	0.80
Common Mental Disorders	1.71	[0.62, 4.09]	0.26
Gynaecological, Neuro-oncology and Orphan Cancer	0.96	[0.28, 2.52]	0.94
Gynaecology and Fertility	2.16	[0.93, 4.54]	0.05
Heart	1.32	[0.48, 3.07]	0.55
Hepato-Biliary	1.57	[0.57, 3.66]	0.33
Kidney and Transplant	0.68	[0.20, 1.78]	0.48
Neonatal	3.08	[1.49, 6.05]	0.002 **
Pain, Palliative and Supportive Care	0.41	[0.07, 1.36]	0.22
Pregnancy and Childbirth	1.93	[1.08, 3.32]	0.02 *
Stroke	1.07	[0.36, 2.58]	0.90

Supplementary Table 2: Results of logistic regression for the outcome of 'Lost, regained' cumulative meta-analysis classification. Asterisks mark coefficient significance levels.

Explanatory variables	OR	95% Confidence Interval	p-value
MA size (coefficient expressed per 5 studies)	1.08	[1.02, 1.15]	0.005 **
Model: random effects	2.90	[2.02, 4.21]	<0.001 ***
Acute Respiratory Infections	0.71	[0.24, 1.74]	0.50
Airways	1.09	[0.43, 2.40]	0.84
Anaesthesia, Critical and Emergency Care	0.89	[0.30, 2.22]	0.83
Colorectal Cancer	0.58	[0.17, 1.51]	0.32
Common Mental Disorders	1.20	[0.46, 2.76]	0.69
Gynaecological, Neuro-oncology and Orphan Cancer	0.75	[0.25, 1.84]	0.57
Gynaecology and Fertility	0.71	[0.21, 1.87]	0.54
Heart	1.16	[0.48, 2.54]	0.72
Hepato-Biliary	0.90	[0.30, 2.23]	0.84
Kidney and Transplant	0.95	[0.41, 1.98]	0.90
Neonatal	0.53	[0.12, 1.51]	0.29
Pain, Palliative and Supportive Care	2.22	[1.02, 4.49]	0.03 *
Pregnancy and Childbirth	0.72	[0.36, 1.34]	0.33
Stroke	0.78	[0.26, 1.88]	0.61

Supplementary Table 3: Results of logistic regression for the outcome of ‘*Diminishing*’ GLS regression classification. Asterisks mark coefficient significance levels.

Explanatory variables	OR	95% Confidence Interval	p-value
MA size (coefficient expressed per 5 studies)	0.98	[0.92, 1.03]	0.39
Model: random effects	1.40	[1.09, 1.80]	0.009 **
Acute Respiratory Infections	1.05	[0.53, 2.00]	0.89
Airways	1.19	[0.65, 2.09]	0.56
Anaesthesia, Critical and Emergency Care	0.74	[0.33, 1.53]	0.44
Colorectal Cancer	0.79	[0.39, 1.51]	0.48
Common Mental Disorders	1.28	[0.62, 2.59]	0.49
Gynaecological, Neuro-oncology and Orphan Cancer	0.72	[0.33, 1.44]	0.36
Gynaecology and Fertility	1.30	[0.71, 2.33]	0.37
Heart	0.81	[0.40, 1.55]	0.54
Hepato-Biliary	0.81	[0.39, 1.57]	0.55
Kidney and Transplant	0.62	[0.31, 1.19]	0.17
Neonatal	1.08	[0.60, 1.90]	0.79
Pain, Palliative and Supportive Care	0.92	[0.49, 1.69]	0.81
Pregnancy and Childbirth	1.03	[0.68, 1.58]	0.86
Stroke	0.39	[0.17, 0.80]	0.02 *

Supplementary Table 4: Comparison of cumulative meta-analysis patterns and category from the regression line of the fitted GLS model. Percentages are row percentages.

		Category (GLS model)				Total
		Flat	Crossing	Diminishing	Diverging	
Pattern of statistical significance	Retained	232 (47%)	10 (2%)	166 (33%)	90 (18%)	498
	Only at end	18 (56%)	4 (12%)	1 (3%)	9 (28%)	32
	Never reached	188 (43%)	136 (32%)	85 (20%)	20 (5%)	429
	Lost, regained	73 (43%)	6 (4%)	76 (45%)	15 (9%)	170
	Lost, not regained	53 (35%)	12 (8%)	86 (57%)	1 (1%)	152
	Total	564 (44%)	168 (13%)	414 (32%)	135 (11%)	1281
First study result vs. final pooled estimate	More extreme /Same side	311 (42%)	38 (5%)	365 (50%)	21 (3%)	735
	Less extreme /Same side	131 (46%)	27 (9%)	35 (12%)	94 (33%)	287
	Less extreme /Different side	41 (50%)	23 (28%)	1 (1%)	17 (21%)	82
	More extreme /Different side	81 (46%)	80 (45%)	13 (7%)	3 (2%)	177
	Total	564 (44%)	168 (13%)	414 (32%)	135 (11%)	1281

Supplementary Table 5: Comparison of the presence of one or more \bar{X} and CUSUM violations. Percentages are total percentages.

		CUSUM Violations		
		0	1+	Total
\bar{X} Violations	0	523 (74.2%)	17 (2.4%)	540 (76.6%)
	1+	56 (7.9%)	109 (15.5%)	165 (23.4%)
	Total	579 (82.1%)	126 (17.9%)	705 (100%)

Supplementary Table 6: Results of logistic regression, outcome 'At least one \bar{X} control chart violation'. Asterisks mark coefficient significance levels.

Explanatory variables	OR	95% Confidence Interval	<i>p</i> -value
MA size (coefficient expressed per 5 studies)	1.64	[1.46, 1.86]	<0.001***
Airways	1.06	[0.44, 2.37]	0.89
Gynaecology and Fertility	2.14	[1.03, 4.30]	0.04 *
Neonatal	3.64	[1.93, 6.78]	<0.001***
Pain, Palliative and Supportive Care	5.80	[2.90, 11.71]	<0.001***
Pregnancy and Childbirth	0.44	[0.16, 1.01]	0.08
Stroke	0.68	[0.21, 1.75]	0.46

Supplementary Table 7: Results of logistic regression, outcome 'At least one CUSUM control chart violation'. Asterisks mark coefficient significance levels.

Explanatory variables	OR	95% Confidence Interval	<i>p</i> -value
MA size (coefficient expressed per 5 studies)	1.73	[1.53, 1.97]	<0.001***
Airways	0.78	[0.27, 1.95]	0.62
Gynaecology and Fertility	1.76	[0.77, 3.77]	0.16
Neonatal	1.73	[0.77, 3.62]	0.16
Pain, Palliative and Supportive Care	5.87	[2.85, 12.02]	<0.001***
Pregnancy and Childbirth	0.47	[0.15, 1.17]	0.14
Stroke	0.47	[0.10, 1.48]	0.25