Living systematic reviews: 3. Statistical methods for updating meta-analyses

Mark Simmonds[a,*], Georgia Salanti[b], Joanne McKenzie[c], Julian Elliott[c], On behalf of the Living Systematic Review Network

[a]Centre for Reviews and Dissemination, University of York, York YO10 5DD, UK

[b]Institute of Social and Preventive Medicine (ISPM), University of Bern, Niesenweg 6, Bern 3012, Switzerland

[c]Cochrane Australia School of Public Health & Preventive Medicine, Monash University, Level 4, 553 St Kilda Road, Melbourne, Victoria 3004, Australia

*Corresponding author:

Tel.: 01904 321091.

E-mail address: mark.simmonds@york.ac.uk (M. Simmonds).

Funding

All the authors of this paper were funded to produce this research by a grant from the Cochrane Methods Innovation Fund. The Living Systematic Review Network is supported by funding from Cochrane and the Australian National Health and Medical Research Council (Partnership Project grant APP1114605). Georgia Salanti is supported by a Marie Sk1odowska-Curie fellowship (MSCA-IF-703254).

Keywords

Living systematic review; Meta-analysis; Type I error; Type II error; Heterogeneity

Published in final form edited form as: *J Clin Epidemiol.* 2017 Nov;91:38-46. doi: 10.1016/j.jclinepi.2017.08.008

1 Table and 3 Figures

Table 1: Key properties of the updating methods

Figure 1: Type I error rate as the number of studies or updates in a meta analysis increases.

Figure 2: Cumulative meta-analysis of the peptic ulcer data.

Figure 3: Applying the four sequential methods to the peptic ulcer meta-analysis

Published in final form edited form as: *J Clin Epidemiol.* 2017 Nov;91:38-46. doi: 10.1016/j.jclinepi.2017.08.008

ABSTRACT

A living systematic review (LSR) should keep the review current as new research evidence emerges. Any meta-analyses included in the review will also need updating as new material is identified. If the aim of the review is solely to present the best current evidence standard meta-analysis may be sufficient, provided reviewers are aware that results may change at later updates. If the review is used in a decision-making context, more caution may be needed. When using standard meta-analysis methods, the chance of incorrectly concluding that any updated meta-analysis is statistically significant when there is no effect (the type I error) increases rapidly as more updates are performed. Inaccurate estimation of any heterogeneity across studies may also lead to inappropriate conclusions. This paper considers four methods to avoid some of these statistical problems when updating meta-analyses: two methods, that is, law of the iterated logarithm and the Shuster method control primarily for inflation of type I error and two other methods, that is, trial sequential analysis and sequential meta-analysis control for type I and II errors (failing to detect a genuine effect) and take account of heterogeneity. This paper compares the methods and considers how they could be applied to LSRs.

Box "What is new?"

- Living systematic reviews will require updating of any included meta-analyses at each review update.
- If a living systematic review is used as part of a decision-making process, the frequent updating of the meta-analysis could lead to inappropriate conclusions being drawn, due to an inflated risk of falsely concluding statistical significance (type I error).
- Four statistical methods exist to avoid type I error inflation, and other statistical problems, that arise in repeated meta-analyses.
- This paper gives an overview of these methods and how meta-analyses should be performed in a living systematic review.

Box 1 Living systematic reviews

- A systematic review which is continually updated, incorporating relevant new evidence as it becomes available
- An approach to review updating not a formal review methodology
- Can be applied to any type of review
- Uses standard systematic review methods
- Explicit and a priori commitment to a predetermined frequency of search and review updating

Box 2 An example meta-analysis of peptic ulcer trials

As an example of how the methods might be applied, we apply these methods to a meta-analysis of 23 trials comparing endoscopic hemostasis to a control treatment for treatment of bleeding peptic ulcers[23]. This was originally used as an example to illustrate sequential meta-analysis[19] but is applied to all methods here.

A random-effects cumulative meta-analysis is shown in Fig. 2. This shows the results of the meta-analysis if it were updated once for every new trial, from the first-published trial at the top, to the last,

at the bottom. Each row of the forest plot representing the meta-analysis of all trials up to that point. It can be seen that a conventionally statistically significant result is achieved once only four trials have been included. We compare this to applying the four methods considered, assuming we wish to control for the standard type I error rate of 5%. For trial sequential analysis and sequential meta-analysis, we also assume we wish to have 90% power to detect a relative risk of 0.5 (which is that found from a meta-analysis of all the trials). In this example, we do not use the ''approximate Bayes'' heterogeneity estimation for sequential meta-analysis.

Fig. 3 shows the results for the four methods, respectively, (A) trial sequential analysis, (B) sequential meta-analysis, (C) Shuster, and (D) law of the iterated logarithm. In each case, the red dots and line show the progress of the updated meta-analyses after adding each trial, starting at the third trial, since a random-effects meta-analysis of two trials cannot reliably estimate heterogeneity. The black lines show the stopping boundaries for each method. Trial sequential analysis and sequential meta-analysis cross both the boundary for demonstrating treatment benefit and the maximum required sample size or information boundary after 10 trials for trial sequential analysis and 11 for sequential meta-analysis, although trial sequential analysis just touches the boundary after 6 and 9 trials. This shows that the required information or sample size has been reached after 10 or 11 trials, so had this analysis been run as a living systematic review, updating could reasonably have been stopped or slowed at that point. The law of the iterated logarithm and the Shuster methods take longer to find in favor of the treatment, requiring 16 or 17 trials to cross a boundary.

These analyses have been shown as if there were an update to the LSR after every new trial. If updates are less frequent, so multiple trials are added at each update, the analyses and their results are the same. It is currently conventional to display the results of trial sequential analysis and sequential meta-analysis methods as if an update had been performed for every trial, but this is not required. All analyses were performed in R, and the code is available from the authors on request. Code for trial sequential analysis is also available from the project website[24].

Published in final form edited form as: *J Clin Epidemiol.* 2017 Nov;91:38-46. doi: 10.1016/j.jclinepi.2017.08.008

1 - BACKGROUND

The key intention of a living systematic review (LSR, see Box 1), which differentiates it from a standard systematic review, is that it will be updated frequently, ideally as soon as any new relevant study is published or identified[1-3]. Over time the information available to be included may increase, requiring the review to be-updated to ensure it is presenting the best available evidence. In many updates, this will require updating one or more of the meta-analyses included in the review.

There are two purposes for undertaking an LSR, which while subtly different have implications for the methods used to update meta-analyses. The first purpose is to present a summary of the evidence at the time of the most recent update. For this purpose, simply repeating each meta-analysis (whether fixed or random effects), adding the newly identified studies and presenting new forest plots and summary estimates, may be the most appropriate approach. All other components of the meta-analyses such as assessment of heterogeneity, subgroup analysis, and investigations of reporting bias will also have to be updated and repeated. Provided the meta-analysis methods used are appropriate, this approach will give the best estimate of the effect of interest at that point in time[4]. However, both the reviewers and readers should be aware that the results may change at later updates, and findings may be highly uncertain if there are few studies or participants included in the analysis.

Systematic reviews and meta-analyses are also used for clinical decision-making, guideline development, and reimbursement decisions. Typically, the level of credibility for the meta-analyses of many beneficial and harmful outcomes is considered before making recommendations for practice. An LSR in particular might be used to support the creation of ‘‘living guidelines’’[5], in which the best available evidence about the benefits and harms of an intervention is used to inform frequently updated recommendations about the use of the intervention. The effect estimate from the meta-analysis and its precision (or confidence interval) is one of the deciding factors in grading the existing evidence, and in this paper, we discuss the implications of continually or frequently updating meta-analyses for the statistical precision of the summary effects.

In a meta-analysis of clinical trials, we may wish to determine if an experimental treatment is superior, inferior, or equivalent to a control treatment. If the review presents assessments of statistical significance with a conventional 95% confidence interval or a P-value of 0.05, then updating of the meta-analyses may overestimate the number of meta-analyses considered statistically significant. While each individual analysis has only a 5% chance of finding a statistically significant result when, in fact, there is none (type I error), the chance of finding a false statistical significant result in any one meta-analysis increases as we repeat these analyses with each review update[6].

As an example, consider a sequence of clinical trials of a new intervention compared to a control, with an updated meta-analysis conducted as soon as each new trial is published. Suppose that there is no true difference in effect between intervention groups on a particular outcome. In this circumstance, the type I error rate, of incorrectly getting a statistically significant result, rises rapidly with each new analysis, as shown in Fig. 1. Similarly, the confidence intervals that often accompany the summary effect will be too narrow if calculated using a conventional meta-analysis. Therefore, using assessments of statistical significance at any individual update of a meta-analysis carries a substantial risk of erroneously concluding that the new intervention is beneficial (or harmful). More formally, repeating a meta-analysis inflates the type I error.

In an LSR, we may also wish to determine when there is sufficient evidence such that we can be confident there is no meaningful effect to detect (such as no important difference in effect between new intervention and the control). This should be achieved so that a type II error is avoided, that is, the error of failing to detect a genuine effect and so that no future update will detect any evidence of a clinically meaningful effect. In a clinical trial, we might select an effect size to identify, such as a minimal clinically meaningful effect, a statistical power to detect that effect (e.g., 80% or 90%) and calculate the required sample size for the trial. We might conclude that the true effect size is less than the clinically meaningful effect if no statistically significant result is found once the specified sample size has been reached[7]. A similar approach can be taken with meta-analyses, including those in an LSR. However, previous analyses have found that few meta-analyses ever reach a sufficient sample size[8].

When an LSR is used only to summarize the best evidence on a topic over time, using standard meta-analysis methods should be sufficient as the review is updated. However, if the LSR is being used to make decisions or readers will use it to do so, then we may wish to consider approaches to avoid inadvertent type I and II errors. This paper considers four methods that have been proposed to correct for these potential errors when updating a meta-analysis. While this paper focuses on LSRs, the same issues apply to all systematic reviews which may be updated. For example, Cochrane recommends that all Cochrane reviews be kept up to date, with revisions at least every 2 years if new trials have been published.

2 - ANALYSIS METHODS FOR REPEATED META-ANALYSES

Updating a meta-analysis has some similarities with interim analyses of clinical trials[9-11]. Interim analyses are often performed in trials so the trial can be stopped early if there is convincing evidence that the intervention is beneficial or harmful. Methods have been developed to avoid type I and II errors and produce robust conclusions for these trial sequential analyses. These methods have been adapted for the analysis of repeated meta-analyses and more recently for the updating of network meta-analysis.

Heterogeneity is also of particular concern in repeated meta-analyses. Heterogeneity should be considered in any meta-analysis, but it cannot be estimated accurately with few studies, and its estimation may vary substantially as a meta-analysis is updated. Incorrect estimation of heterogeneity may affect the conclusions drawn if the level of variability across studies is overestimated or underestimated. Heterogeneity also affects the required sample size, as greater heterogeneity reduces statistical certainty in the evidence and so increases the sample size required to detect a specified effect size.

2.1 -Trial sequential analysis

Trial sequential analysis seeks to control the type I error by ensuring that the cumulative type I error rate across all updates remains at the desired level (usually 5%). To do this, the method uses the principle of alpha spending, that is, penalizing the type I error rate (alpha) at each analysis[12-14]. To avoid type II error, a maximum required sample size to detect some assumed effect size is also specified. This sample size is calculated in the same way as if the meta-analysis was a single clinical trial, by setting a desired type I error, an assumed effect size, and the desired statistical power to detect that effect.

In order to avoid inflated type I error prior to achieving the maximum sample size, alpha-spending boundaries are applied to the meta-analysis. In trial sequential analysis, the O'Brien-Fleming boundaries are applied to the sample size[15]. At each update of the meta-analysis, the Z score (estimated treatment effect divided by its standard error) is calculated. If this exceeds the upper alpha-spending boundary, then the result can be considered conclusive. For example, in a clinical trial, this would lead to a conclusion that the experimental intervention was superior to the control. Correspondingly, if the Z score were less than the lower alpha-spending boundary, the experimental intervention is worse than the control. If the maximum sample size is exceeded without crossing an alpha-spending boundary, we would conclude that any effect of the intervention is less than the specified effect. Additional stopping boundaries can be added to test for futility, so the updating process can be stopped if it is unlikely that a meaningful effect will be found.

Ideally, the assumed effect size would be the minimal clinically important effect size, as recommended by experts in the relevant field [16]. Alternatively, the effect size may be based on the trials currently in the meta-analysis. If this approach is used, it is recommended that only trials judged to be at low risk of bias be used to estimate the desired effect[14]. Heterogeneity across studies increases the sample size because it increases uncertainty in the effect estimates. It is therefore recommended that the sample size be adjusted for heterogeneity, using either some prespecified estimate of heterogeneity or the best current estimate of heterogeneity in the meta-analysis. In trial sequential analysis, the

heterogeneity adjustment is generally made using the D^2 statistic, which is mathematically correct and produces a larger required sample size, although the more widely used I^2 statistic may be used instead[17].

2.2 - Sequential meta-analysis

Sequential meta-analysis, in a similar way to trial sequential analysis, uses methods adapted from sequential trial monitoring and applies them to a meta-analysis[10]. Sequential meta-analysis uses Whitehead's sequential trial boundaries approach to control type I error inflation and also type II error (failing to detect a genuine effect)[18,19].

Sequential meta-analysis is based around calculating the cumulative Z score (the sum of the study effect estimates times their meta-analytic weights) and the cumulative statistical information V (the sum of the inverse of the study weights) at each update. A conclusive result is deemed to be achieved if the Z/V pair lies outside some prespecified boundary. For meta-analysis, a rectangular boundary is recommended, as this reduces the chance of crossing a boundary very early. Hence, if Z exceeds some boundary value Z_{MAX}, then there is evidence of a beneficial effect (as when crossing an alpha-spending boundary in trial sequential analysis). If V exceeds a boundary V_{MAX}, then the updating can be stopped as no conclusive result is ever likely to be found, as the maximum required statistical information or sample size has been reached. The Z_{MAX} and V_{MAX} values are calculated based on setting a desired type I error, an assumed effect size, and the desired statistical power to detect that effect.

Sequential meta-analysis implicitly adjusts for heterogeneity because as heterogeneity increases, the information contained in the meta-analysis decreases. This means the cumulative information V can decrease between updates as well as increase. Sequential meta-analysis can also control for misestimation of heterogeneity using an ''approximate Bayesian'' approach[19]. The DerSimoniane – Laird estimate of heterogeneity used at each update of the random-effects meta-analysis is replaced by a weighted average of the DerSimonianeLaird estimate and a prior estimate of heterogeneity. If

this prior estimate is suitably large, the method can control for underestimation of heterogeneity (and consequent overestimation of statistical information) early in the updating process.

2.3 - The Shuster method

The Shuster method is a newer alternative to the above two methods, designed by Shuster and Neu[20]. This method also uses alpha-spending boundaries but with the more conservative Pocock boundaries used in place of the O'Brien-Fleming boundaries used in trial sequential analysis[20]. The Pocock boundaries were chosen as they are considered more robust to possible changes over time in the effect size and to the fact that the required sample size is estimated rather than known.

Rather than a Z score, a modified t statistic is used. The result is only considered conclusive if the t statistic crosses the Pocock alpha-spending boundary. The method controls only for type I error inflation, so an assumed treatment effect and power are not required, and no sample size or statistical information estimate is needed. This method requires prespecifying the number of meta-analysis updates that will be performed. As this may not be known for an LSR, a reasonable guess will have to be made.

The Shuster method makes no explicit adjustment for heterogeneity, but in a random-effects analysis, the t statistic is a function of heterogeneity, decreasing as heterogeneity increases.

2.4 - Law of the iterated logarithm

Unlike the preceding methods, the law of the iterated logarithm approach is not based on sequential trial analysis[21,22]. Instead, it seeks to adjust the usual Z statistic so that the desired type I error (e.g., 5%) is maintained across all updates. To do this, the method utilizes the fact that a modified form of the conventional Z statistic can be constructed to be bounded as the sample size N tends to infinity: The law of the iterated logarithm approach therefore recommends replacing the standard Z statistic at update k with a similar penalized statistic which is bounded as the statistical information (inverse of the sum of the meta-analytic weights) increases:

The formulae also require a further penalty term λ in the denominator. For an appropriate choice of λ, we can ensure that this penalized statistic is bounded by some suitable value, such as 1.96 for a conventional 95% confidence interval. Comparing this penalized statistic to 1.96 ensures that the standard 5% type I error is maintained across updates. The suggested values of λ are 2 for analyses of odds ratios, risk ratios, and mean differences and 1.5 for risk differences[21]. As with the Shuster method, the law of the iterated logarithm method only controls for type I error inflation, so does not require specification of sample size, an assumed effect estimate, or power. As with the Shuster method, no explicit adjustment for heterogeneity is made, other than the impact on the adjusted Z statistic from heterogeneity when using a random-effects analysis.

An application of the four methods to a meta-analysis of peptic ulcer trials is presented in Box 2.

3 - METHODS FOR NETWORK META-ANALYSIS

A multivariate extension of the alpha-spending boundaries method has been proposed for updating network meta-analysis under the assumption of consistency[25]. Despite the computational complexity in the presence of multiple interventions, the approach is essentially the same as in pairwise meta-analysis. Relative treatment effects between the compared treatments need to be set so as to satisfy the consistency assumptions. Then successively, monitoring boundaries for a predefined level of power are calculated so that overall, the type I error is at the nominal level. Comparison-specific treatment effects are updated after a study is added to the network as it contributes indirect evidence. In the method presented by Nikolakopoulou et al., informative priors are used for heterogeneity throughout.

Updating a network meta-analysis requires additional considerations. The addition of a trial examining a given comparison updates the treatment effects for all other treatment comparisons examined in the network. The assumption of consistency underlying this method needs to be reassessed after each update and the inflation of type I error needs to be controlled for in the inferences. In the early phases

of the network where few studies are included, estimation of inconsistency and heterogeneity will be problematic[26].

4 - COMMENTARY ON THE METHODS

The key properties of each method are outlined in Table 1. Most of the methods for handling repeated meta-analysis are based on an analogy between repeating meta-analysis and sequential analysis of a single clinical trial. While this analogy is generally reasonable, it has some limitations because meta-analyses are based on multiple studies and are not a single controlled trial. Heterogeneity between studies is an obvious key difference. In all methods, if a random-effects meta-analysis is used, the test score incorporates the extra uncertainty and decreases as heterogeneity increases. In sequential meta-analysis, the observed information decreases if the observed heterogeneity increases, and in trial sequential analysis, the required sample size is adjusted for heterogeneity, so will increase if heterogeneity increases. Neither law of the iterated logarithm nor the Shuster method makes any explicit adjustment for heterogeneity, other than its effect on the t statistic or adjusted Z statistic. Currently, only sequential meta-analysis accounts for poor estimation of heterogeneity, particularly when there are few studies, by using the approximate Bayesian adjustment. However, as this adjustment is essentially an alternative estimator for heterogeneity, it could, in principle, be used in any of the methods.

The methods have been described here as reaching a conclusion when some specified boundary is crossed (as seen in Fig. 3). It is also possible to represent the methods in a conventional forest plot, as with the cumulative plot in Fig. 2. This is achieved by adjusting the conventional 95% confidence intervals using the stopping boundaries so that the adjusted confidence interval excludes the null value only if a stopping boundary is crossed. Trial sequential analysis-adjusted confidence intervals can be generated, and the principle has been illustrated elsewhere for the sequential meta-analysis method[19] but can be similarly used for all four methods discussed here.

Although sequential meta-analysis and trial sequential analysis appear different on the surface, they are, in fact, based on the same underlying statistical theory of using O'Brien-Fleming alpha-spending boundaries to adjust the significance level required to judge that an effect is statistically significant. As such, the methods should, in principle, have similar properties, although results may differ in any particular meta-analysis[27].

The primary difference between the methods is that sequential meta-analysis is based on the required statistical information to detect a desired effect, whereas trial sequential analysis generally uses the required sample size. Sample size depends on properties of the studies, such as the risk of an event in the control group. This may vary across studies and its estimate may change as the meta-analysis is updated, and so, the required sample size may not be constant across updates. Sample size should also be adjusted for heterogeneity. This could be done using the estimated heterogeneity at the current update, in which case sample size may vary substantially between updates. Alternatively, some prior estimate of expected heterogeneity could be used, but the sample size may be inappropriate if this estimate does not reflect the observed heterogeneity. Using required statistical information instead (as in sequential meta-analysis) has the advantage that it is independent of the properties of the trials, and of the heterogeneity, so, it does not vary across updates and can be calculated before trials are identified (e.g., in the protocol). Statistical information is, however, more difficult to interpret than sample size, and the total information may decrease between updates if the heterogeneity increases substantially. Although trial sequential analysis generally uses the sample size in its calculations, it is possible to use statistical information instead without any change to the underlying method.

As law of the iterated logarithm and the Shuster method control only for type I error inflation, they do not specify a required sample size or statistical information, nor a desired effect size or statistical power to detect it. This may make them simpler to implement as the stopping boundaries are not dependent on the properties of the studies included in the analysis or of external factors such as a clinically meaningful effect size. However, it does mean that these two methods have no stopping

conditions if there is no observable effect, so the methods cannot easily recommend that the updating of an LSR shall be stopped for futility. While trial sequential analysis and sequential meta-analysis do allow for stopping for futility, they require specification of a desired effect size, which may require specialist knowledge to determine and may be arbitrary or overestimate the true effect.

The methods could also be used to make judgments about when to update the LSR and its meta-analysis. Informally, if the current results are close to a stopping boundary, then an update might be needed soon, but if the results are a long way from a boundary, then it may be appropriate to wait longer. In the sequential meta-analysis and trial sequential analysis methods, it is possible to estimate how much statistical information or additional sample size might be needed before a boundary is crossed, and so, time future updates for when that level of information might become available from new trials. To our knowledge, these methods have not yet been used in this way so any use of these methods to plan future update should be cautious. Other methods for determining when and if a meta-analysis should be updated have been developed and could be used alongside the sequential methods considered here[7,28,29].

5 – CONCLUSIONS AND RECOMMENDATIONS

The aim of an LSR is to provide the best available evidence to support decision-making by updating frequently, potentially as soon as a single relevant new study is identified. As with conventional approaches to updating, it is to be expected that the findings of the meta-analyses may change between updates and so reviewers should be suitably cautious when drawing conclusions from a meta-analysis in an LSR, particularly why n considering if a result is statistically significant.

The methods discussed in this paper should, in principle, increase the chance that conclusions drawn from a repeated meta-analysis are robust. The use of these methods in LSRs could therefore help prevent reviewers and readers from drawing inappropriate conclusions about the effectiveness of interventions. If these methods are used in an LSR, they should be clearly set out in the review protocol, including specification of desired type I error, assumed effect size, and the desired statistical

power. All the methods considered have been shown to avoid type I error inflation, as demonstrated in simulation studies for each method, and, to a somewhat lesser extent, in practical application in real meta-analyses. While this paper has focused on LSRs, the need to avoid errors of interpretation applies to all meta-analyses that are updated, even if less frequently than in an LSR. If a meta-analysis receives only one or two updates, however, the type I error inflation is modest, and there may be less need for these methods.

The frequent updating in LSRs may make them more resource intensive, expensive, and time-consuming to perform than a conventional review which might be updated infrequently or never. Given this, it is likely that in any LSR, decisions will have to be made about when to perform updates and if regular updating could be made less frequent or stopped. A possible benefit of the methods is that they could provide guidance as to when ceasing to update an LSR, or reducing update frequency, is statistically justifiable. The high risk of type I error means that conventional statistical significance is unsuitable for this[30]. When a stopping boundary is crossed in the methods considered here, however, the conclusions of the analysis are unlikely (up to the specified type I error) to change at future updates.

In an LSR, it would also be useful to know that updating could be stopped because no meaningful effect will ever be found. Reaching the maximum sample size or statistical information (without crossing any other boundary) in trial sequential analysis and sequential meta-analysis provides a possible means for making such a decision. It should be noted, however, that the properties of using these methods to decide on when and how to update an LSR has not yet been formally investigated.

Heterogeneity across studies in a meta-analysis will always be of concern, particularly when there are few studies so any estimation of heterogeneity is uncertain. This is a particular issue in LSRs as misestimation of heterogeneity will lead to incorrect confidence intervals and wrong judgments about the required sample size or amount of statistical information contained in the analysis. The approximate Bayes estimation of heterogeneity used in sequential metaanalysis may help to prevent such misestimation when there are few studies. However, any meta-analysis in an LSR which shows a

statistically significant result based on few studies, little information, or where there is evidence of substantial heterogeneity should be treated with caution, and further updates considered.

The methods described can correct for the statistical errors of type I and II errors, but they do not prevent other nonstatistical errors of analysis or interpretation. In particular, they do not correct for bias, and analysts should still consider the possibility of publication and selective reporting biases, as well as potential for bias due to including poor-quality studies.

This paper has only considered applying the methods to a single outcome, but most LSRs will meta-analyze multiple outcomes. Conclusions drawn from the LSR and decisions regarding stopping updating will, naturally, have to consider the findings across all outcomes and potentially on any subgroup analyses. The methods discussed here could potentially be used simultaneously on multiple outcomes, but the value of doing this is currently unclear. Similarly, all the methods are designed for the analyses of trials comparing interventions. How to avoid statistical errors when updating other types of review, such as in diagnostic test accuracy or prognostic testing, remains uncertain.

Some issues relating to the use of these methods remain uncertain and require further research. These include how the methods behave for different effect metrics (mean differences, relative risk, risk difference), their properties when data are sparse or highly heterogeneous, and how robust methods are when a boundary is crossed.

All the methods considered here are designed to achieve correct type I errors or P-values across repeated meta-analyses. Of course, making judgments about the value of an intervention based on the P-value alone is, rightly, widely criticized[31]. In any statistical analysis, it would be wrong to assume that an intervention is beneficial simply because a P-value of below 0.05 has been found. The same applies to sequential methods; when a boundary is crossed, the full evidence should be considered, including effect size, confidence intervals and heterogeneity, and the evidence from other outcomes or subgroups. The main purpose of these methods is, perhaps, not so much to demonstrate a beneficial effect as to avoid misinterpretation of conventional meta-analyses and confidence intervals

in LSRs where frequent updating means the risk of type I error is high and to guide the need for updating.

ACKNOWLEDGMENTS

The authors would like to thank the members of the Living Systematic Review Network for their comments on drafts of this paper, particularly Philippe Ravaud, Andrew Maas, Kurinichi Gurusamy, Laura Martinez, Joerg Meerpohl and Stefania Mondello.

REFERENCES

1. Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JP, Mavergames C, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med 2017;11(2): e1001603.
2. Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, et al. Living systematic reviews: 1. Introductiondthe why, what, when and how. J Clin Epidemiol 2017;91:23e30.
3. Thomas J, Noel-Storr A, Marshall I, Wallace B, McDonald S, Mavergames C, et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol 2017;91:31e7.
4. Ioannidis JP, Contopoulos-Ioannidis DG, Lau J. Recursive cumulative meta-analysis: a diagnostic for the evolution of total randomized evidence from group and individual patient data. J Clin Epidemiol 1999;52:281e91.
5. Akl EA, Meerpohl JJ, Elliott J, Kahale LA, Schunemann HJ. Living systematic reviews: 4. Living guideline recommendations. J Clin Epidemiol 2017;91:47e53.
6. Borm GF, Donders AR. Updating meta-analyses leads to larger type I errors than publication bias. J Clin Epidemiol 2009;62:825e830.e10.
7. Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR, Abrams KR. Evidence-based sample size calculations based upon updated meta-analysis. Stat Med 2007;26:2479e500.
8. Turner RM, Bird SM, Higgins JP. The impact of study size on metaanalyses: examination of underpowered studies in Cochrane reviews. PLoS One 2013;8:e59202.
9. Lan KKG, Demets DL. Discrete sequential boundaries for clinicaltrials. Biometrika 1983;70(3):659e63.
10. Whitehead J. A unified theory for sequential clinical trials. Stat Med 1999;18:2271e86.
11. Pogue JM, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. Control Clin Trials 1997;18:580e93.
12. Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be inconclusiveetrial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. Int J Epidemiol 2009;38:287e98.
13. Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JP, Thabane L, et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? Int J Epidemiol 2009;38:276e86.

14. Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative metaanalysis. J Clin Epidemiol 2008;61:64e75.
15. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549e56.
16. Cook JA, Hislop J, Altman DG, Fayers P, Briggs AH, Ramsay CR, et al. Specifying the target difference in the primary outcome for a randomised controlled trial: guidance for researchers. Trials 2015; 16:12.
17. Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in random-effects model meta-analyses. BMC Med Res Methodol 2009;9:86.
18. Whitehead A. A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. Stat Med 1997;16: 2901e13.
19. Higgins JP, Whitehead A, Simmonds M. Sequential methods for random-effects meta-analysis. Stat Med 2011;30:903e21.
20. Shuster JJ, Neu J. A Pocock approach to sequential meta-analysis of clinical trials. Res Synth Methods 2013;4(3):269e79.
21. Hu M, Cappelleri JC, Lan KK. Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. Clin Trials 2007;4:329e40.
22. Lan KKG, Hu M, Cappelleri JC. Applying the law of iterated logarithm to cumulative meta-analyses of a continuous endpoint. Stat Sin 2003;13(4):1135e45.
23. Sacks HS, Chalmers TC, Blum AL, Berrier J, Pagano D. Endoscopic hemostasis. An effective therapy for bleeding peptic ulcers. JAMA 1990;264:494e9.
24. Trial Sequential Analysis 2017. Available at http://www.ctu.dk/toolsand-links/trial-sequential-analysis.aspx. Accessed September 9, 2017.
25. Nikolakopoulou A, Mavridis D, Egger M, Salanti G. Continuously updated network meta-analysis and statistical monitoring for timely decision-making. Stat Methods Med Res 2016. Available at http://journals.sagepub.com/doi/pdf/10.1177/0962280216659896. Accessed September 9, 2017.
26. Veroniki AA, Straus SE, Soobiah C, Elliott MJ, Tricco AC. A scoping review of indirect comparison methods and applications using individual patient data. BMC Med Res Methodol 2016;16:47.
27. Imberger G, Gluud C,Wetterslev J. Comments on 'Sequential methods for random-effects meta-analysis'. Stat Med 2011;30:2965e6.
28. Roloff V, Higgins JP, Sutton AJ. Planning future studies based on the conditional power of a meta-analysis. Stat Med 2013;32:11e24.
29. Langan D, Higgins JP, Gregory W, Sutton AJ. Graphical augmentations to the funnel plot assess the impact of additional evidence on a meta-analysis. J Clin Epidemiol 2012;65:511e9.
30. Berkey CS, Mosteller F, Lau J, Antman EM. Uncertainty of the time of first significance in random effects cumulative meta-analysis. Control Clin Trials 1996;17:357e71.
31. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31(4):337e50.

Published in final form edited form as: *J Clin Epidemiol.* 2017 Nov;91:38-46. doi: 10.1016/j.jclinepi.2017.08.008

478 **TABLE**

479 **Table 1**

480 Key properties of the updating methods

	Trial sequential analysis	Sequential meta-analysis	Shuster	Law of the iterated logarithm
Corrects for type I error	Yes	Yes	Yes	Yes
Corrects for type II error	Yes	Yes	No	No
Assumed effect size and statistical power required	Yes	Yes	No	No
Need to specify number of updates	No	No	Yes	No
Adjusts information/sample size for heterogeneity	Yes	Yes	No	No
Adjusts for misestimation of heterogeneity	No	Optional	No	No

481

482

Published in final form edited form as: *J Clin Epidemiol.* 2017 Nov;91:38-46. doi: 10.1016/j.jclinepi.2017.08.008

FIGURES

Figure 1

Type I error rate as the number of studies or updates in a meta-analysis increases.

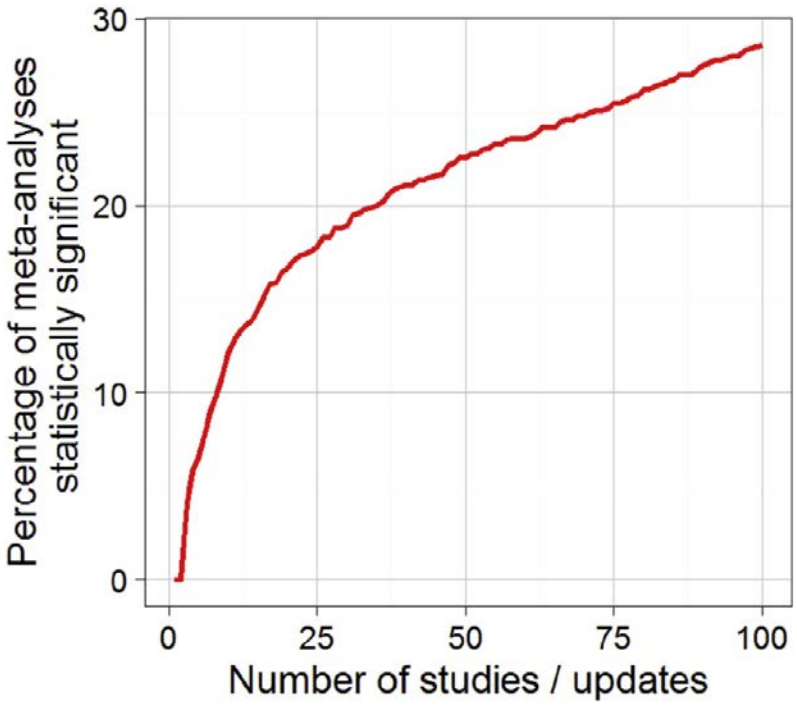
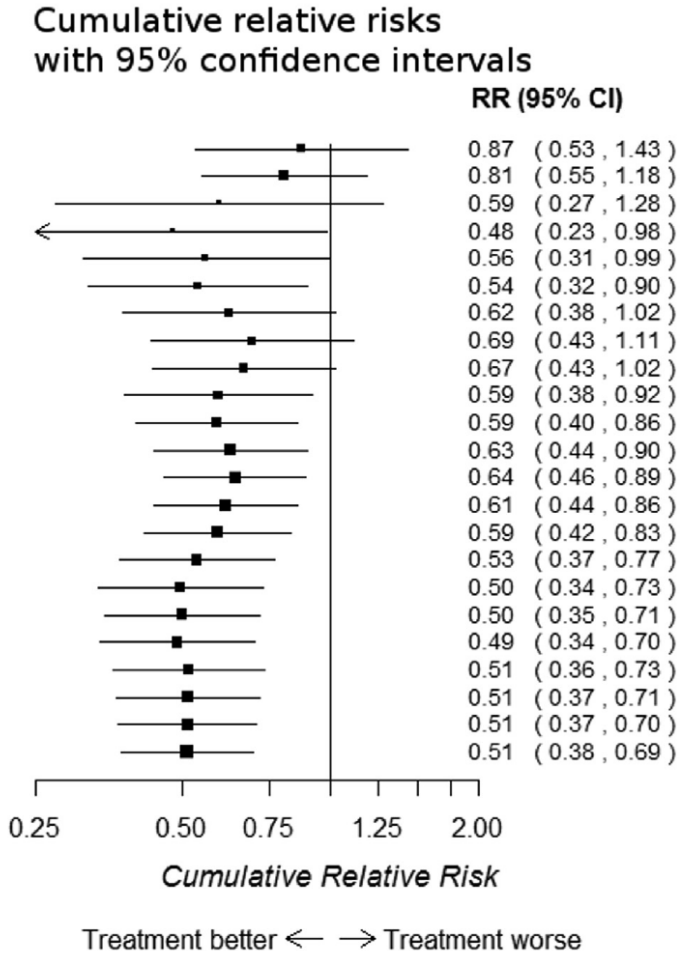


Figure 2

Cumulative meta-analysis of the peptic ulcer data. Each row of the forest plot representing the meta-analysis of all trials up to that point, as if it were updated once for every new trial, from the first-published trial at the top, to the last, at the bottom.



Published in final form edited form as: *J Clin Epidemiol.* 2017 Nov;91:38-46. doi: 10.1016/j.jclinepi.2017.08.008

Figure 3

Applying the four sequential methods to the peptic ulcer meta-analysis. Results of updated meta-analyses are shown for (A) trial sequential analysis, (B) sequential meta-analysis, (C) Shuster, and (D) law of the iterated logarithm. The red dots and line show the progress of the updated meta-analyses after adding each trial, starting at the third trial, since a random-effects meta-analysis of two trials cannot reliably estimate heterogeneity. The black lines show the stopping boundaries for each method. Trial sequential analysis plots the standard Z score against cumulative sample size. Sequential meta-analysis plots the cumulative Z score (the sum of the study effect estimates times their meta-analytic weights) against the cumulative statistical information (the sum of the inverse of the study weights). Law of iterated logarithm plots the penalized Z score at each update or trial and the Shuster method, the adjusted t statistic at each update or trial. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)