

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

2 Mark Simmonds^{a,*}, Georgia Salanti^b, Joanne McKenzie^c, Julian Elliott^c, On behalf of the Living
3 Systematic Review Network

4
5 ^aCentre for Reviews and Dissemination, University of York, York YO10 5DD, UK

6 ^bInstitute of Social and Preventive Medicine (ISPM), University of Bern, Niesenweg 6, Bern 3012,
7 Switzerland

8 ^cCochrane Australia School of Public Health & Preventive Medicine, Monash University, Level 4, 553
9 St Kilda Road, Melbourne, Victoria 3004, Australia

10

11 *Corresponding author:

12 Tel.: 01904 321091.

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

14

15 **Funding**

16 All the authors of this paper were funded to produce this research by a grant from the Cochrane
17 Methods Innovation Fund. The Living Systematic Review Network is supported by funding from
18 Cochrane and the Australian National Health and Medical Research Council (Partnership Project grant
19 APP1114605). Georgia Salanti is supported by a Marie Skłodowska-Curie fellowship (MSCA-IF-
20 703254).

21

22 **Keywords**

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

24

25 **1 Table and 3 Figures**

26 **Table 1:** Key properties of the updating methods

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

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

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

30 **ABSTRACT**

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

44 **Box “What is new?”**

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

54

55 **Box 1 Living systematic reviews**

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

62

63 **Box 2 An example meta-analysis of peptic ulcer trials**

64 As an example of how the methods might be applied, we apply these methods to a meta-analysis of
65 23 trials comparing endoscopic hemostasis to a control treatment for treatment of bleeding peptic
66 ulcers²³. This was originally used as an example to illustrate sequential meta-analysis¹⁹ but is applied
67 to all methods here.

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

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

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

89 These analyses have been shown as if there were an update to the LSR after every new trial. If updates
90 are less frequent, so multiple trials are added at each update, the analyses and their results are the
91 same. It is currently conventional to display the results of trial sequential analysis and sequential meta-
92 analysis methods as if an update had been performed for every trial, but this is not required. All
93 analyses were performed in R, and the code is available from the authors on request. Code for trial
94 sequential analysis is also available from the project website²⁴.

95

96 **1 - BACKGROUND**

97 The key intention of a living systematic review (LSR, see Box 1), which differentiates it from a standard
98 systematic review, is that it will be updated frequently, ideally as soon as any new relevant study is
99 published or identified¹⁻³. Over time the information available to be included may increase, requiring
100 the review to be updated to ensure it is presenting the best available evidence. In many updates, this
101 will require updating one or more of the meta-analyses included in the review.

102 There are two purposes for undertaking an LSR, which while subtly different have implications for the
103 methods used to update meta-analyses. The first purpose is to present a summary of the evidence at
104 the time of the most recent update. For this purpose, simply repeating each meta-analysis (whether
105 fixed or random effects), adding the newly identified studies and presenting new forest plots and
106 summary estimates, may be the most appropriate approach. All other components of the meta-
107 analyses such as assessment of heterogeneity, subgroup analysis, and investigations of reporting bias
108 will also have to be updated and repeated. Provided the meta-analysis methods used are appropriate,
109 this approach will give the best estimate of the effect of interest at that point in time⁴. However, both
110 the reviewers and readers should be aware that the results may change at later updates, and findings
111 may be highly uncertain if there are few studies or participants included in the analysis.

112 Systematic reviews and meta-analyses are also used for clinical decision-making, guideline
113 development, and reimbursement decisions. Typically, the level of credibility for the meta-analyses of
114 many beneficial and harmful outcomes is considered before making recommendations for practice.

115 An LSR in particular might be used to support the creation of “living guidelines”⁵, in which the best
116 available evidence about the benefits and harms of an intervention is used to inform frequently
117 updated recommendations about the use of the intervention. The effect estimate from the meta-
118 analysis and its precision (or confidence interval) is one of the deciding factors in grading the existing
119 evidence, and in this paper, we discuss the implications of continually or frequently updating meta-
120 analyses for the statistical precision of the summary effects.

121 In a meta-analysis of clinical trials, we may wish to determine if an experimental treatment is superior,
122 inferior, or equivalent to a control treatment. If the review presents assessments of statistical
123 significance with a conventional 95% confidence interval or a P-value of 0.05, then updating of the
124 meta-analyses may overestimate the number of meta-analyses considered statistically significant.
125 While each individual analysis has only a 5% chance of finding a statistically significant result when, in
126 fact, there is none (type I error), the chance of finding a false statistical significant result in any one
127 meta-analysis increases as we repeat these analyses with each review update⁶.

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

137 In an LSR, we may also wish to determine when there is sufficient evidence such that we can be
138 confident there is no meaningful effect to detect (such as no important difference in effect between
139 new intervention and the control). This should be achieved so that a type II error is avoided, that is,
140 the error of failing to detect a genuine effect and so that no future update will detect any evidence of
141 a clinically meaningful effect. In a clinical trial, we might select an effect size to identify, such as a
142 minimal clinically meaningful effect, a statistical power to detect that effect (e.g., 80% or 90%) and
143 calculate the required sample size for the trial. We might conclude that the true effect size is less than
144 the clinically meaningful effect if no statistically significant result is found once the specified sample
145 size has been reached⁷. A similar approach can be taken with meta-analyses, including those in an LSR.
146 However, previous analyses have found that few meta-analyses ever reach a sufficient sample size⁸.

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

155

156 **2 - ANALYSIS METHODS FOR REPEATED META-ANALYSES**

157 Updating a meta-analysis has some similarities with interim analyses of clinical trials⁹⁻¹¹. Interim
158 analyses are often performed in trials so the trial can be stopped early if there is convincing evidence
159 that the intervention is beneficial or harmful. Methods have been developed to avoid type I and II
160 errors and produce robust conclusions for these trial sequential analyses. These methods have been
161 adapted for the analysis of repeated meta-analyses and more recently for the updating of network
162 meta-analysis.

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

170

171

172

173 **2.1 -Trial sequential analysis**

174 Trial sequential analysis seeks to control the type I error by ensuring that the cumulative type I error
175 rate across all updates remains at the desired level (usually 5%). To do this, the method uses the
176 principle of alpha spending, that is, penalizing the type I error rate (alpha) at each analysis¹²⁻¹⁴. To
177 avoid type II error, a maximum required sample size to detect some assumed effect size is also
178 specified. This sample size is calculated in the same way as if the meta-analysis was a single clinical
179 trial, by setting a desired type I error, an assumed effect size, and the desired statistical power to
180 detect that effect.

181 In order to avoid inflated type I error prior to achieving the maximum sample size, alpha-spending
182 boundaries are applied to the meta-analysis. In trial sequential analysis, the O'Brien-Fleming
183 boundaries are applied to the sample size¹⁵. At each update of the meta-analysis, the Z score
184 (estimated treatment effect divided by its standard error) is calculated. If this exceeds the upper alpha-
185 spending boundary, then the result can be considered conclusive. For example, in a clinical trial, this
186 would lead to a conclusion that the experimental intervention was superior to the control.
187 Correspondingly, if the Z score were less than the lower alpha-spending boundary, the experimental
188 intervention is worse than the control. If the maximum sample size is exceeded without crossing an
189 alpha-spending boundary, we would conclude that any effect of the intervention is less than the
190 specified effect. Additional stopping boundaries can be added to test for futility, so the updating
191 process can be stopped if it is unlikely that a meaningful effect will be found.

192 Ideally, the assumed effect size would be the minimal clinically important effect size, as recommended
193 by experts in the relevant field [16]. Alternatively, the effect size may be based on the trials currently
194 in the meta-analysis. If this approach is used, it is recommended that only trials judged to be at low
195 risk of bias be used to estimate the desired effect¹⁴. Heterogeneity across studies increases the sample
196 size because it increases uncertainty in the effect estimates. It is therefore recommended that the
197 sample size be adjusted for heterogeneity, using either some prespecified estimate of heterogeneity
198 or the best current estimate of heterogeneity in the meta-analysis. In trial sequential analysis, the

199 heterogeneity adjustment is generally made using the D^2 statistic, which is mathematically correct and
200 produces a larger required sample size, although the more widely used I^2 statistic may be used
201 instead¹⁷.

202

203 **2.2 - Sequential meta-analysis**

204 Sequential meta-analysis, in a similar way to trial sequential analysis, uses methods adapted from
205 sequential trial monitoring and applies them to a meta-analysis¹⁰. Sequential meta-analysis uses
206 Whitehead's sequential trial boundaries approach to control type I error inflation and also type II error
207 (failing to detect a genuine effect)^{18,19}.

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

218 Sequential meta-analysis implicitly adjusts for heterogeneity because as heterogeneity increases, the
219 information contained in the meta-analysis decreases. This means the cumulative information V can
220 decrease between updates as well as increase. Sequential meta-analysis can also control for
221 misestimation of heterogeneity using an "approximate Bayesian" approach¹⁹. The DerSimoniane –
222 Laird estimate of heterogeneity used at each update of the random-effects meta-analysis is replaced
223 by a weighted average of the DerSimonianeLaird estimate and a prior estimate of heterogeneity. If

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

226

227 **2.3 - The Shuster method**

228 The Shuster method is a newer alternative to the above two methods, designed by Shuster and Neu²⁰.

229 This method also uses alpha-spending boundaries but with the more conservative Pocock boundaries

230 used in place of the O'Brien-Fleming boundaries used in trial sequential analysis²⁰. The Pocock

231 boundaries were chosen as they are considered more robust to possible changes over time in the

232 effect size and to the fact that the required sample size is estimated rather than known.

233 Rather than a Z score, a modified t statistic is used. The result is only considered conclusive if the t

234 statistic crosses the Pocock alpha-spending boundary. The method controls only for type I error

235 inflation, so an assumed treatment effect and power are not required, and no sample size or statistical

236 information estimate is needed. This method requires prespecifying the number of meta-analysis

237 updates that will be performed. As this may not be known for an LSR, a reasonable guess will have to

238 be made.

239 The Shuster method makes no explicit adjustment for heterogeneity, but in a random-effects analysis,

240 the t statistic is a function of heterogeneity, decreasing as heterogeneity increases.

241

242 **2.4 - Law of the iterated logarithm**

243 Unlike the preceding methods, the law of the iterated logarithm approach is not based on sequential

244 trial analysis^{21,22}. Instead, it seeks to adjust the usual Z statistic so that the desired type I error (e.g.,

245 5%) is maintained across all updates. To do this, the method utilizes the fact that a modified form of

246 the conventional Z statistic can be constructed to be bounded as the sample size N tends to infinity:

247 The law of the iterated logarithm approach therefore recommends replacing the standard Z statistic

248 at update k with a similar penalized statistic which is bounded as the statistical information (inverse

249 of the sum of the meta-analytic weights) increases:

250 The formulae also require a further penalty term λ in the denominator. For an appropriate choice of
251 λ , we can ensure that this penalized statistic is bounded by some suitable value, such as 1.96 for a
252 conventional 95% confidence interval. Comparing this penalized statistic to 1.96 ensures that the
253 standard 5% type I error is maintained across updates. The suggested values of λ are 2 for analyses of
254 odds ratios, risk ratios, and mean differences and 1.5 for risk differences²¹. As with the Shuster
255 method, the law of the iterated logarithm method only controls for type I error inflation, so does not
256 require specification of sample size, an assumed effect estimate, or power. As with the Shuster
257 method, no explicit adjustment for heterogeneity is made, other than the impact on the adjusted Z
258 statistic from heterogeneity when using a random-effects analysis.

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

260

261 **3 - METHODS FOR NETWORK META-ANALYSIS**

262 A multivariate extension of the alpha-spending boundaries method has been proposed for updating
263 network meta-analysis under the assumption of consistency²⁵. Despite the computational complexity
264 in the presence of multiple interventions, the approach is essentially the same as in pairwise meta-
265 analysis. Relative treatment effects between the compared treatments need to be set so as to satisfy
266 the consistency assumptions. Then successively, monitoring boundaries for a predefined level of
267 power are calculated so that overall, the type I error is at the nominal level. Comparison-specific
268 treatment effects are updated after a study is added to the network as it contributes indirect evidence.
269 In the method presented by Nikolakopoulou et al., informative priors are used for heterogeneity
270 throughout.

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

275 of the network where few studies are included, estimation of inconsistency and heterogeneity will be
276 problematic²⁶.

277

278 **4 - COMMENTARY ON THE METHODS**

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

293 The methods have been described here as reaching a conclusion when some specified boundary is
294 crossed (as seen in Fig. 3). It is also possible to represent the methods in a conventional forest plot, as
295 with the cumulative plot in Fig. 2. This is achieved by adjusting the conventional 95% confidence
296 intervals using the stopping boundaries so that the adjusted confidence interval excludes the null
297 value only if a stopping boundary is crossed. Trial sequential analysis-adjusted confidence intervals
298 can be generated, and the principle has been illustrated elsewhere for the sequential meta-analysis
299 method¹⁹ but can be similarly used for all four methods discussed here.

300 Although sequential meta-analysis and trial sequential analysis appear different on the surface, they
301 are, in fact, based on the same underlying statistical theory of using O'Brien-Fleming alpha-spending
302 boundaries to adjust the significance level required to judge that an effect is statistically significant.
303 As such, the methods should, in principle, have similar properties, although results may differ in any
304 particular meta-analysis²⁷.

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

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

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

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

340

341 **5 – CONCLUSIONS AND RECOMMENDATIONS**

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

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

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

358 The frequent updating in LSRs may make them more resource intensive, expensive, and time-
359 consuming to perform than a conventional review which might be updated infrequently or never.
360 Given this, it is likely that in any LSR, decisions will have to be made about when to perform updates
361 and if regular updating could be made less frequent or stopped. A possible benefit of the methods is
362 that they could provide guidance as to when ceasing to update an LSR, or reducing update frequency,
363 is statistically justifiable. The high risk of type I error means that conventional statistical significance is
364 unsuitable for this³⁰. When a stopping boundary is crossed in the methods considered here, however,
365 the conclusions of the analysis are unlikely (up to the specified type I error) to change at future
366 updates.

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

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

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

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

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

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

395 All the methods considered here are designed to achieve correct type I errors or P-values across
396 repeated meta-analyses. Of course, making judgments about the value of an intervention based on
397 the P-value alone is, rightly, widely criticized³¹. In any statistical analysis, it would be wrong to assume
398 that an intervention is beneficial simply because a P-value of below 0.05 has been found. The same
399 applies to sequential methods; when a boundary is crossed, the full evidence should be considered,
400 including effect size, confidence intervals and heterogeneity, and the evidence from other outcomes
401 or subgroups. The main purpose of these methods is, perhaps, not so much to demonstrate a
402 beneficial effect as to avoid misinterpretation of conventional meta-analyses and confidence intervals

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

405

406 **ACKNOWLEDGMENTS**

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

410

411 **REFERENCES**

- 412 1. Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JP, Mavergames C, et al. Living systematic reviews:
413 an emerging opportunity to narrow the evidence-practice gap. *PLoS Med* 2017;11(2): e1001603.
- 414 2. Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, et al. Living systematic reviews: 1.
415 Introductionthe why, what, when and how. *J Clin Epidemiol* 2017;91:23e30.
- 416 3. Thomas J, Noel-Storr A, Marshall I, Wallace B, McDonald S, Mavergames C, et al. Living systematic
417 reviews: 2. Combining human and machine effort. *J Clin Epidemiol* 2017;91:31e7.
- 418 4. Ioannidis JP, Contopoulos-Ioannidis DG, Lau J. Recursive cumulative meta-analysis: a diagnostic for
419 the evolution of total randomized evidence from group and individual patient data. *J Clin Epidemiol*
420 1999;52:281e91.
- 421 5. Akl EA, Meerpohl JJ, Elliott J, Kahale LA, Schunemann HJ. Living systematic reviews: 4. Living
422 guideline recommendations. *J Clin Epidemiol* 2017;91:47e53.
- 423 6. Borm GF, Donders AR. Updating meta-analyses leads to larger type I errors than publication bias. *J*
424 *Clin Epidemiol* 2009;62:825e830.e10.
- 425 7. Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR, Abrams KR. Evidence-based sample size
426 calculations based upon updated meta-analysis. *Stat Med* 2007;26:2479e500.
- 427 8. Turner RM, Bird SM, Higgins JP. The impact of study size on metaanalyses: examination of
428 underpowered studies in Cochrane reviews. *PLoS One* 2013;8:e59202.
- 429 9. Lan KKG, Demets DL. Discrete sequential boundaries for clinicaltrials. *Biometrika*
430 1983;70(3):659e63.
- 431 10. Whitehead J. A unified theory for sequential clinical trials. *Stat Med* 1999;18:2271e86.
- 432 11. Pogue JM, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring
433 boundaries for cumulative meta-analysis. *Control Clin Trials* 1997;18:580e93.
- 434 12. Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be
435 inconclusiveetrial sequential analysis adjustment of random error risk due to repetitive testing of
436 accumulating data in apparently conclusive neonatal meta-analyses. *Int J Epidemiol*
437 2009;38:287e98.
- 438 13. Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JP, Thabane L, et al. Can trial sequential
439 monitoring boundaries reduce spurious inferences from meta-analyses? *Int J Epidemiol*
440 2009;38:276e86.

- 441 14. Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm
442 evidence is reached in cumulative metaanalysis. *J Clin Epidemiol* 2008;61:64e75.
- 443 15. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979;35:549e56.
- 444 16. Cook JA, Hislop J, Altman DG, Fayers P, Briggs AH, Ramsay CR, et al. Specifying the target difference
445 in the primary outcome for a randomised controlled trial: guidance for researchers. *Trials* 2015;
446 16:12.
- 447 17. Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying
448 diversity in random-effects model meta-analyses. *BMC Med Res Methodol* 2009;9:86.
- 449 18. Whitehead A. A prospectively planned cumulative meta-analysis applied to a series of concurrent
450 clinical trials. *Stat Med* 1997;16: 2901e13.
- 451 19. Higgins JP, Whitehead A, Simmonds M. Sequential methods for random-effects meta-analysis. *Stat*
452 *Med* 2011;30:903e21.
- 453 20. Shuster JJ, Neu J. A Pocock approach to sequential meta-analysis of clinical trials. *Res Synth*
454 *Methods* 2013;4(3):269e79.
- 455 21. Hu M, Cappelleri JC, Lan KK. Applying the law of iterated logarithm to control type I error in
456 cumulative meta-analysis of binary outcomes. *Clin Trials* 2007;4:329e40.
- 457 22. Lan KKG, Hu M, Cappelleri JC. Applying the law of iterated logarithm to cumulative meta-analyses
458 of a continuous endpoint. *Stat Sin* 2003;13(4):1135e45.
- 459 23. Sacks HS, Chalmers TC, Blum AL, Berrier J, Pagano D. Endoscopic hemostasis. An effective therapy
460 for bleeding peptic ulcers. *JAMA* 1990;264:494e9.
- 461 24. Trial Sequential Analysis 2017. Available at [http://www.ctu.dk/toolsand-links/trial-sequential-](http://www.ctu.dk/toolsand-links/trial-sequential-analysis.aspx)
462 [analysis.aspx](http://www.ctu.dk/toolsand-links/trial-sequential-analysis.aspx). Accessed September 9, 2017.
- 463 25. Nikolakopoulou A, Mavridis D, Egger M, Salanti G. Continuously updated network meta-analysis
464 and statistical monitoring for timely decision-making. *Stat Methods Med Res* 2016. Available at
465 <http://journals.sagepub.com/doi/pdf/10.1177/0962280216659896>. Accessed September 9, 2017.
- 466 26. Veroniki AA, Straus SE, Soobiah C, Elliott MJ, Tricco AC. A scoping review of indirect comparison
467 methods and applications using individual patient data. *BMC Med Res Methodol* 2016;16:47.
- 468 27. Imberger G, Gluud C, Wetterslev J. Comments on 'Sequential methods for random-effects meta-
469 analysis'. *Stat Med* 2011;30:2965e6.
- 470 28. Roloff V, Higgins JP, Sutton AJ. Planning future studies based on the conditional power of a meta-
471 analysis. *Stat Med* 2013;32:11e24.
- 472 29. Langan D, Higgins JP, Gregory W, Sutton AJ. Graphical augmentations to the funnel plot assess the
473 impact of additional evidence on a meta-analysis. *J Clin Epidemiol* 2012;65:511e9.
- 474 30. Berkey CS, Mosteller F, Lau J, Antman EM. Uncertainty of the time of first significance in random
475 effects cumulative meta-analysis. *Control Clin Trials* 1996;17:357e71.
- 476 31. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values,
477 confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016;31(4):337e50.

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

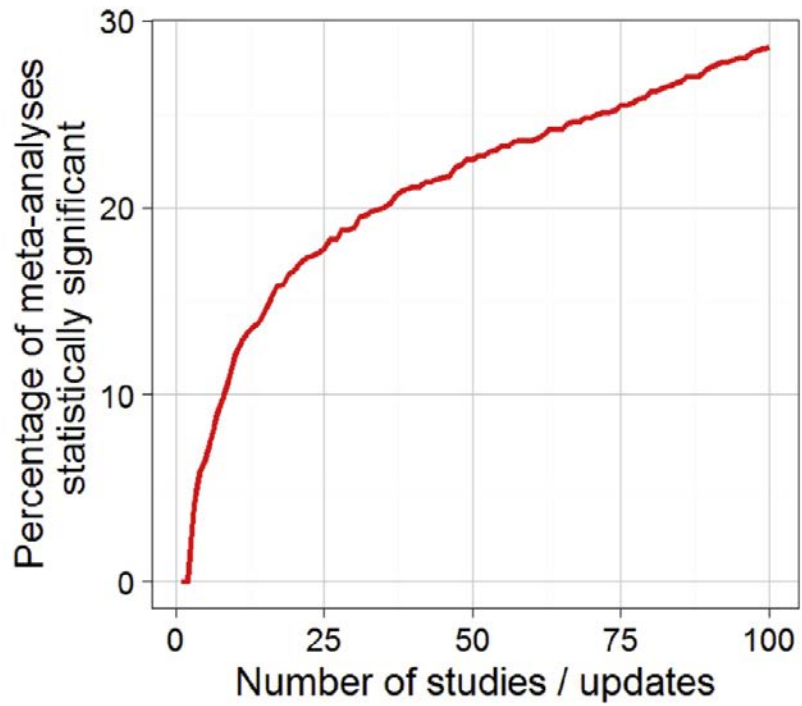
481

482

483 **FIGURES**

484 **Figure 1**

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

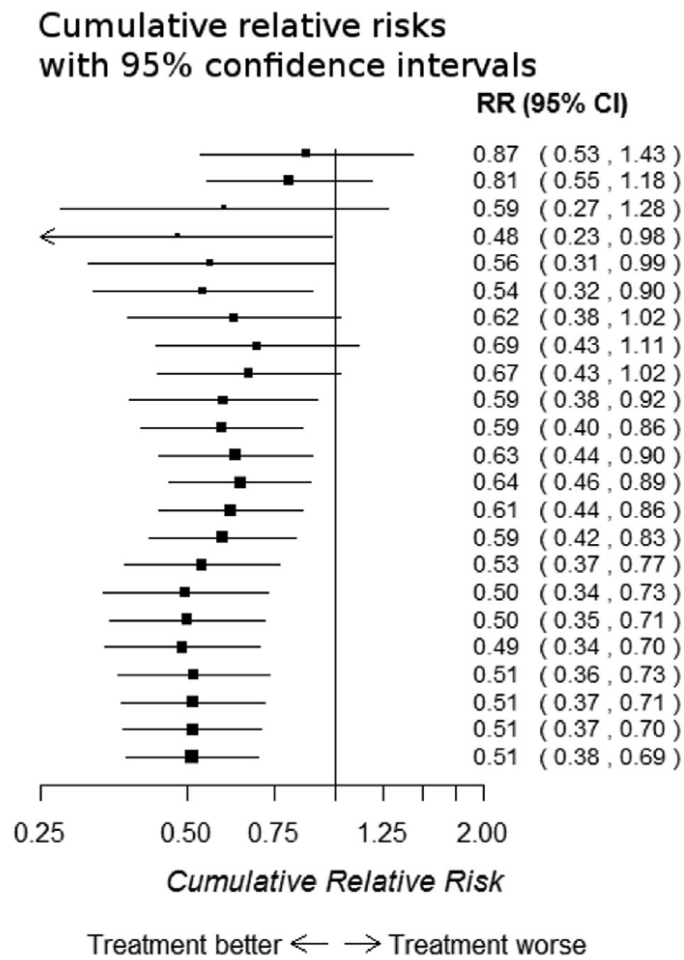


486

487

488 **Figure 2**

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



492

493

494 **Figure 3**

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

