A methodological review of recent meta-analyses has found significant heterogeneity in age between randomized groups

Laura Clarka,*, Caroline Fairhursta, Catherine E. Hewitta, Yvonne Birksb, Sally Brabync, Sarah Cockayneb, Sara Rodgersa, Katherine Hicksa, Robert Hodgsona, Elizabeth Littlewoodc, David J. Torgersona

aDepartment of Health Sciences, York Trials Unit, University of York, York YO10 5DD, United Kingdom
bDepartment of Social Policy, University of York, York YO10 5DD, United Kingdom
cDepartment of Health Sciences, Mental Health and Addictions Research Group, University of York, York YO10 5DD, United Kingdom

Accepted 23 April 2014; Published online xxxx

Abstract

Background: There is evidence to suggest that component randomized controlled trials (RCTs) within systematic reviews may be biased. It is important that these reviews are identified to prevent erroneous conclusions influencing health care policies and decisions.

Purpose: To assess the likelihood of bias in trials in 12 meta-analyses.

Design: A review of 12 systematic reviews.


Study Selection and Data Extraction: Systematic reviews were eligible for inclusion if they included only RCTs. We obtained the full text for the component RCTs of the 12 systematic reviews (in English only). We extracted summary data on age, number of participants in each treatment group, and the method of allocation concealment for each RCT.

Data Synthesis: Five of the 12 meta-analyses exhibited heterogeneity in age differences ($I^2 > 0.30$), when there should have been none. In two meta-analyses, the age of the intervention group was significantly greater than that of the control group. Inadequate allocation concealment was a statistically significant predictor of heterogeneity in one trial as observed by a metaregression.

Conclusions: Most of the sample of recent meta-analyses showed that there were signs of imbalance and/or heterogeneity in ages between treatment groups, when there should have been none. Systematic reviewers might consider using the techniques described here to assess the validity of their findings.

Keywords: Systematic review; Selection bias; Randomized controlled trials; Methods; Meta-analysis; Heterogeneity

1. Introduction

Ideally, a systematic review and meta-analysis of randomized controlled trials (RCTs) should be used to inform clinical practice [1]. However, there is good evidence to suggest that some RCTs in systematic reviews may be biased because of insecure allocation concealment [2–8]. The assessment of whether a meta-analysis contains such biased trials is important in relation to the validity of the findings. One approach outlined by Trowman et al. is to undertake a meta-analysis of an important baseline variable, in their case body weight [9], incorporating all the trials in the review. They found that there was significant imbalance, which explained virtually all the difference between groups at follow-up. This approach of using a baseline variable, which randomization should ensure would differ only by chance, allows us to assess whether a meta-analysis of RCTs is reliable. To date, techniques using baseline information are not routinely used to assess the quality of meta-analyses. Trowman et al. used baseline body weight because this was the most powerful predictor of the outcome (ie, final body weight). However, in many meta-analyses, the outcome variable either is not reported at randomization or is not possible to collect. Another potentially powerful predictor of outcome is age. Age is usually a good predictor of outcome; older people tend to have worse outcomes than younger people. Furthermore, if one wishes to subvert the allocation of a trial then consciously or unconsciously misallocating according to a person’s age is
2. Methods

In editions dating back from May 2012, we identified the first three systematic reviews containing only RCTs published in each of The Annals of Internal Medicine, The British Medical Journal, The Journal of the American Medical Association, and The Lancet. These journals were selected, because they are the highest impact medical journals that frequently publish meta-analyses of RCTs. Other high-impact journals, such as the New England Journal of Medicine do not publish high numbers of meta-analyses.

2.1. Data extraction

Where available, the full-text articles of the component RCTs from the 12 reviews available in English were retrieved. In addition to the method of allocation concealment, the following information was extracted for each trial arm where possible: summary of age (mean or median), its measure of dispersion (SD, standard error [SE], range, and interquartile range), and number of participants. Double, independent data extraction was performed by two researchers (ie, authors of this article); any disagreements were resolved by discussion. No other aspect of trial design was extracted from the articles, and we did not contact the trial authors of randomized trials where data were not available in the published articles.

2.1.1. Allocation concealment

Adequacy of allocation concealment was judged using the Cochrane handbook criteria [10]. Trials were classified as adequate, inadequate, or unclear. The reviewers made a judgment as to the quality of allocation concealment without knowing whether there was an age imbalance for that appraised trial.

2.1.2. Age

If age was not summarized using the mean and SD, for example, if median and range were presented, measures were converted using standard approximation formulas [11]. A fixed effect meta-analysis of age was performed for each review on the assumption that there was a common treatment estimate (ie, zero) across the randomized trials. The P-value for the difference in age between the control and intervention groups for each systematic review was calculated. The $I^2$ value of heterogeneity from the meta-analysis was interpreted in line with the Cochrane handbook guidelines: 0%-40% might not be important; 30%-60% may represent moderate heterogeneity; 50%-90% may represent substantial heterogeneity; and 75%-100% considerable heterogeneity [12]. A metaregression was performed for each review to assess whether allocation concealment adequacy was a predictor of heterogeneity. Meta-analyses and regression were performed in Comprehensive Meta-Analysis 2 (Biostat, Englewood, NJ, USA).

If the null hypothesis is true, the P-values from independent two-sample t-tests should follow a uniform distribution reasonably well provided the data on which the t-tests are carried out are normally distributed and the tests are independent [13]. We applied this theory here. A systematic reviewer can assess for differences in a continuous baseline covariate between the control and intervention groups using a t-test for each RCT in their review. For each test, the null hypothesis should be true; therefore, statistically testing differences in the control and intervention groups at baseline across trials included in a systematic review should yield P-values with a uniform distribution if the randomizations were all faithful. That is, ~10% of the...
P-values should lie within each interval of size 0.1. For each systematic review, we compared the distribution of P-values obtained from a t-test for the age of the control and intervention groups from each component trial to that expected using the chi-square goodness-of-fit test. A further analysis for continuous variables is based on the principle of the central limit theorem (CLT), and its use has previously been described in a similar context [14]. The CLT states that the mean value of samples taken repeatedly from the same population will be approximately normally distributed around the mean of the population. The SD of this curve can be estimated by the standard error of the mean (SEM) of any of the samples used to construct the curve, in the relation SEM = SD/√n_s, where SD_s is the standard deviation of the sample s of size n_s. We calculated the trial population mean age and pooled SD from the summary data for each trial and standardized the mean of the control group X (X - μ)/SEM]. The standardized means should follow a curve with mean 0 and SD 1. We plotted histograms of these standardized sample means for each review overlaid by a standard normal curve and tested whether their distribution differed from the expected distribution using a test for the equality of variance (sdtest, Stata v12; StataCorp LP, College Station, Texas, USA). Deviations from this expected distribution could suggest that participants randomized to the control groups in the component RCTs of a particular review were statistically significantly older or younger than the population mean. Histograms were created in the statistical package R.

Sensitivity analyses were performed as described previously but omitting trials where the mean and/or SD of age had been derived from approximation formulas.

3. Results

Fig. 1 summarizes the flow of systematic reviews and component randomized trials in this review. The 12 systematic reviews included a total of 503 trials [15–26]; 184 trials were not available, so 319 have been included in this analysis. Reasonable attempts were made to retrieve all texts through the University of York library service, but in certain circumstances, this was not possible and it was not deemed necessary to expend excessive monetary or time resources on sourcing texts, because this review is just an exemplar of the techniques described. Of the trials included, 20% (n = 64) had adequate allocation concealment, 4% (n = 14) inadequate, and 76% (n = 243) failed to provide sufficient information for a judgment to be made (see Appendix A at www.jclinepi.com for list of included trials). A further 110 trials were excluded from the meta-analyses because of baseline age data being unavailable for extraction.

![Flow diagram of included systematic reviews](image-url)
3.1. Meta-analysis

Of the 12 included systematic reviews, we found that five had heterogeneity ($I^2 > 0.30$) [15,19,21,25,26], including one with substantial heterogeneity (0.50 < $I^2$ ≤ 0.75) [25] and one with considerable heterogeneity ($I^2 > 0.75$) [15]. Two systematic reviews [15,20] had a statistically significant difference in age between the control and intervention groups overall. In each instance, the mean age in the intervention group was greater than that of the control. Table 1 presents the results for age of the meta-analysis ranked in order of heterogeneity and forest plots for each systematic review can be found in Appendix B at www.jclinepi.com. A forest plot for one particular review can be found in Fig. 2 as an example. When sensitivity analyses were performed omitting the trials where approximation formulas had been used to convert median age to mean age, there was no difference in results observed; that is, the results remained the same whether we included trials where we had to use approximations to the mean and SD.

3.2. P-value comparison and histograms of standardized control group means of baseline age

There was evidence of a statistically significant discrepancy between the distributions of $P$-values observed compared with that expected in two of the reviews [24,25].

In six of the reviews, the distribution exhibited a statistically significant deviation from that expected ($P < 0.01$) [15,21,23–26] when histograms of standardized control groups means at baseline were investigated. Fig. 2 shows an example of these two plots for one particular review.

3.3. Metaregression results

Metaregression analysis suggested that allocation concealment explained the heterogeneity observed in one review ($P = 0.03$; 15). See Appendix C at www.jclinepi.com for the metaregression statistics and bubble plots for each review.

In Table 2, we summarize the conclusions made in the included 12 systematic reviews and the findings from our analysis. The dots indicate the presence of evidence for the marker of potential bias in the RCT, of which 8 of the 12 reviews had at least one suggesting the conclusions drawn may not be reliable.

4. Discussion

Systematic reviews of randomized controlled trials are considered the highest form of evidence that underpins evidence-based medicine. There is a wealth of research that shows that some randomized trials have had their random allocation subverted. Much of this evidence examines the relationship between allocation description and effect sizes. Broadly, such evidence shows that descriptions of rigorous methods of allocation tend to be associated with smaller effect sizes. In this article, we have examined the problem using a different approach. We hypothesized that a systematic review that contains a significant proportion of trials with biased randomization will exhibit significant heterogeneity in a baseline variable, in this instance age. Such an approach does not rely on detailed descriptions of the randomization process, which is often missing or may be false. We argue that our approach is easier and more sensitive than qualitatively making a judgment about the randomization process, which is often missing or may be false. We argue that our approach is easier and more sensitive than qualitatively making a judgment about the randomization process.

In 12 recent systematic reviews, we found that only four reviews demonstrated the expected zero heterogeneity [16,18,22,24]. All other reviews had at least one “symptom” of the following: significant heterogeneity in baseline age; significant differences in baseline age; or an unlikely distribution of $P$-values or standardized means of baseline age.
Two of the reviews [15, 25] had three factors and three had two factors [21, 24, 26] that suggest that some of their component RCTs were unreliable. Because all the trials in these reviews were reported to have been randomized, we would expect that the null hypothesis of no age differences between groups except by chance would be true. Thus, we should have observed no heterogeneity in ages between groups; a meta-analysis of age should show equivalence and the distribution of P-values of baseline ages should be uniform. As we had used a baseline variable common to the trials and not an outcome variable, explanations for heterogeneity often found in meta-analyses, such as different populations or slightly different interventions leading to differing effect sizes, do not apply in this instance. The most plausible explanation for heterogeneity is therefore poor randomization practice. Although we did explore this in a metaregression, it was only statistically significant as an explanation in one study. This may be because allocation concealment was so poorly reported in most randomized trials that it is masked as a source of the heterogeneity. Although some of the “statistically significant” differences in age are relatively minor, the fact that they exist is important because they act as a marker for poor allocation practice. Because a systematic review contains some trials that have misallocation, the entire review is weak and should not be used to drive major changes in clinical practice.

4.1. Limitations

We were unable to examine all the RCTs that were included in the original systematic reviews, as a number
Table 2. Conclusions made in each systematic review with the results of each test applied in this current review

<table>
<thead>
<tr>
<th>Systematic review</th>
<th>Evidence of heterogeneity (I² &gt; 0.3)</th>
<th>Statistically significant difference in age between control and intervention groups</th>
<th>Distribution of P-values deviated from uniform distribution</th>
<th>Distribution of standardized control means deviates from standard normal</th>
<th>Conclusion of systematic review taken directly from the abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anothaisintawee et al. 2012</td>
<td>Yes</td>
<td>Yes, P &lt; 0.01</td>
<td>Yes, P &lt; 0.001</td>
<td>α-Blockers, antibiotics and combinations of these therapies appear to achieve the greatest improvement in clinical symptom scores compared with placebo. Anti-inflammatory therapies have a lesser but measurable benefit on selected outcomes. However, beneficial effects of α-blockers may be overestimated because of publication bias.</td>
<td></td>
</tr>
<tr>
<td>Rutjes et al. 2012</td>
<td></td>
<td>Yes, P &lt; 0.05</td>
<td>Yes, P &lt; 0.01</td>
<td>In patients with knee osteoarthritis, viscosupplementation is associated with a small and clinically irrelevant benefit and an increased risk for serious adverse events.</td>
<td></td>
</tr>
<tr>
<td>Hemmingsen et al. 2012</td>
<td></td>
<td></td>
<td></td>
<td>There was no evidence or even a trend toward improved all-cause mortality or cardiovascular mortality with metformin and insulin, compared with insulin alone in type 2 diabetes. Data were limited by the severe lack of data reported by trials for patient relevant outcomes and by poor bias control.</td>
<td></td>
</tr>
<tr>
<td>Thangaratinam et al. 2012</td>
<td>Yes</td>
<td>Yes, P &lt; 0.05</td>
<td>Yes, P &lt; 0.01</td>
<td>Dietary and lifestyle interventions in pregnancy can reduce maternal gestational weight gain and improve outcomes for both mother and baby. Among the interventions, those based on diet are the most effective and are associated with reductions in maternal gestational weight gain and improved obstetric outcomes.</td>
<td></td>
</tr>
<tr>
<td>Umpierre et al. 2011</td>
<td>Yes</td>
<td></td>
<td>Yes, P &lt; 0.001</td>
<td>Structured exercise training that consists of aerobic exercise, resistance training, or both combined is associated with HbA1c reduction in patients with type 2 diabetes. Structured exercise training of more than 150 minutes per week is associated with greater HbA1c declines than that of 150 minutes or less per week. Physical activity advice is associated with lower HbA1c, but only when combined with dietary advice.</td>
<td></td>
</tr>
<tr>
<td>Neumann et al. 2012</td>
<td>Yes</td>
<td></td>
<td>Yes, P &lt; 0.001</td>
<td>Compared with LMWH, lower doses of oral factor Xa inhibitors can achieve a small absolute risk reduction in symptomatic deep venous thrombosis without increasing bleeding.</td>
<td></td>
</tr>
<tr>
<td>Heneghan et al. 2011</td>
<td>Yes</td>
<td></td>
<td></td>
<td>Our analysis showed that self-monitoring and self-management of oral anticoagulation is a safe option for suitable patients of all ages. Patients should also be offered the option to self-manage their disease with suitable health-care support as back-up.</td>
<td></td>
</tr>
<tr>
<td>Palmer et al. 2012</td>
<td></td>
<td></td>
<td>Yes, P &lt; 0.01</td>
<td>Benefits for antiplatelet therapy among persons with CKD are uncertain and are potentially outweighed by bleeding hazards</td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
Some were confidential reports or were reported in another language than English. Research suggests that methodological quality is similar or higher in English language publications compared with those that are published in other languages [27–30]; therefore, any findings here are likely to at least replicate what may be found in other language articles.

We had relatively small sample sizes for some of our meta-analyses because of the exclusion of a number of RCTs where age data were not reported (n = 186). However, randomized trials that did not report the mean ages with their associated SDs and group numbers are likely to be poorer in quality than well-reported RCTs, so it is possible the heterogeneity we observed may have increased had we been able to include them. To perform the metaregression analysis, 10 or more trials are required for the test to be reliable. In some of our metaregression, we did not have 10 trials, and so the results should be interpreted with caution.

Although we found some evidence that poor allocation concealment was likely to be a driver for the heterogeneity, this finding was only statistically significant in one review. There are difficulties in the judgment of allocation concealment as different criteria and scales are used to judge the adequacy of methods [31]. We used only the Cochrane criteria, so it was possible that had we chosen to use a different criteria, we may have obtained different results. However as previously emphasized, a large proportion of the trial publications inadequately reported the method of allocation concealment so that its adequacy could not be judged by any criteria.

Randomization of participants to a trial arm should ensure that comparisons between randomized groups for

<table>
<thead>
<tr>
<th>Systematic review</th>
<th>Evidence of heterogeneity ((I^2 &gt; 0.3))</th>
<th>Statistically significant difference in age between control and intervention groups</th>
<th>Distribution of (P)-values deviated from uniform distribution</th>
<th>Distribution of standardized control means deviates from standard normal</th>
<th>Conclusion of systematic review taken directly from the abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orrow et al. 2012</td>
<td>Yes, (P &lt; 0.01)</td>
<td></td>
<td></td>
<td></td>
<td>Promotion of physical activity to sedentary adults recruited in primary care significantly increases physical activity levels at 12 months, as measured by self report. We found insufficient evidence to recommend exercise referral schemes over advice or counseling interventions. Primary care commissioners should consider these findings while awaiting further trial evaluation of exercise referral schemes and other primary care interventions, with longer follow-up and use of objective measures of outcome.</td>
</tr>
<tr>
<td>Coombes et al. 2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Despite the effectiveness of corticosteroid injections in the short term, non-corticosteroid injections might be of benefit for long-term treatment of lateral epicondylalgia. However, response to injection should not be generalised because of variation in effect between sites of tendinopathy.</td>
</tr>
<tr>
<td>Leucht et al. 2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Maintenance treatment with antipsychotic drugs benefits patients with schizophrenia. The advantages of these drugs must be weighed against their side-effects. Future studies should focus on outcomes of social participation and clarify the long-term morbidity and mortality of these drugs.</td>
</tr>
<tr>
<td>Hempel et al. 2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The pooled evidence suggests that probiotics are associated with a reduction in AAD. More research is needed to determine which probiotics are associated with the greatest efficacy and for which patients receiving which specific antibiotics.</td>
</tr>
</tbody>
</table>

Abbreviations: AAD, antibiotic-associated diarrhea; HbA₁c, hemoglobin A₁c; LMWH, low-molecular-weight heparin.
independent variables at baseline produce a uniform distribution of P-values. Deviations from this expected distribution could hint at a failure of randomization. We suggest that this method can be used to assess the reliability of randomization; however, caution must be taken, because the P-values for baseline tests will follow a uniform distribution only if the tests used to compute them are valid and independent [13]. The distribution of continuous variables being compared between the groups must be approximately normal for the t-test to be valid, which age seldom is within trial participant populations because for one, upper and/or lower age restrictions are often imposed by the inclusion criteria. This method could more reliably be applied to other prognostic factors known to be approximately normally distributed within trial groups. It is pertinent to acknowledge, however, that this review observed difficulties in accessing all necessary age data, and it may be the case that prognostic factors are not reported well or at all. It would be possible to check these assumptions explicitly if you were conducting a review and had access to individual participant data from each component trial, as opposed to just the reported aggregate summaries.

4.2. Recommendations

We recommend that meta-analysts use the techniques described in this article to check the validity of their analyses. We suggest that all four techniques should be used because a single approach may not be sufficient. For example, in the review by Trowman et al., the heterogeneity in baseline body weight was 0%; similarly, the review by Leucht et al. showed 0% heterogeneity yet both these reviews found statistically significant imbalances in age [9,20]. A reviewer should expect to see no heterogeneity and no statistically significant differences in the outcome between the groups. If a review showed a significant difference in baseline measure of outcome but no heterogeneity, the conclusion from this is that a proportion of the included trials have allocation subversion all favoring the same treatment arm. However, if there were simply significant heterogeneity but no overall difference in the baseline covariate, this does not imply that because subversion is operating in both directions that the review’s results are believable. It may mean that the trials are in significant imbalance in a particular direction that favors an unknown covariate. In truly randomized trials, this unknown or unmeasured covariate will be balanced across a group of trials, but when heterogeneity is present, we cannot be confident that this holds true. If a meta-analysis has significant numbers of randomized trials where the allocation is subverted, there may be no heterogeneity because the “true” difference is a difference in age so no heterogeneity is observed. Nevertheless, such trials are biased and the meta-analysis is unreliable. Current practice for estimating bias in meta-analyses is to grade component trials as being at high or low risk of bias using measures such as the Jadad scale [32] or the Cochrane guidance. Unfortunately, it is likely these scales will misclassify at least some trials as either being good, when they are poor, or vice versa. The approach we recommend here should be used as a complementary method to assess the rigor of a meta-analysis as it gives more information to the reader. We have chosen the prognostic factor of age; however, there are other potentially more important variables on which to do baseline testing. We would recommend that reviewers perform baseline testing on an important, prespecified, prognostic factor relative to their review (eg, if the review was in the field of obesity, body weight could be the baseline characteristic tested) where possible. Finally, we recommend that authors of systematic reviews should prespecify which baseline variables they choose to include in their analysis in their systematic review protocol. This is to avoid undertaking multiple baseline tests and only presenting those that have little or no heterogeneity.

In summary, there is significant unexplained heterogeneity of age in most of the sample of systematic reviews published in high-impact journals. Reviewers should adopt techniques to identify potential baseline imbalances in their trials and use this to drive sensitivity analyses.

Appendix

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2014.04.007.

References


Dechartres A, Charles P, Hopewell S, Ravaud P, Altman DG. Reviews assessing the quality or the reporting of randomized controlled trials are increasing over time but raised questions about how quality is assessed. J Clin Epidemiol 2011;64:136–44.