**A simple technique investigating baseline heterogeneity may help to eliminate potential bias in meta-analyses**

(Amy Hicks, Caroline Fairhurst, David J Torgerson)

York Trials Unit

Department of Health Sciences

University of York

YORK YO10 5DD

Dr Amy Hicks

Foundation Year 2 Doctor

Caroline Fairhurst, MSc

Research Fellow

Professor David Torgerson, PhD

Director

Correspondence to: Dr A Hicks

**Summary box:**

* Bias in randomised controlled trials can impact on the validity of outcomes and conclusions reached in meta-analyses which include them.
* True randomisation produces treatment groups that differ only by chance. The true difference in baseline variables between the two groups is zero; therefore, when several studies are included in a meta-analysis of a baseline variable, the heterogeneity should be zero.
* This straightforward approach involves performing a meta-analysis of baseline variables and systematically removing trials, starting with those with the largest t-statistic, until the I2 measure of heterogeneity becomes 0%, then repeating the outcome meta-analysis with only the remaining trials as a sensitivity check.
* We recommend routine use of this technique, using age and a second baseline variable predictive of outcome for the particular study chosen, to help eliminate potential bias in meta-analyses.

**Introduction:**

Systematic reviews and meta-analyses of randomised controlled trials (RCTs) provide an important summary of the best available evidence to help guide clinical practice. However, not every RCT is conducted as rigorously as we might expect and inclusion of such trials might introduce bias into meta-analyses. Selection bias resulting from inadequate allocation concealment is one key source1. Investigating differences in baseline variables between randomised groups helps to assess the robustness of the allocation process. True randomisation should ensure that baseline variables only differ between randomised groups by chance; therefore, a meta-analysis of a baseline measurement should produce no overall difference and zero heterogeneity. In this paper we describe a method that could be used to identify and remove potentially biased trials from meta-analyses.

**Investigating baseline variables: a background**

This approach stems from work by Trowman et al.2 who undertook meta-analyses of outcome and baseline body weight, and noted a significant imbalance at baseline which explained almost all of the differences between groups at follow up. Clark et al.3 expanded on this work by investigating baseline heterogeneity, as well as baseline imbalances. They argue that heterogeneity is more important than baseline imbalance because misallocation can favour either the treatment or control group, which might result in baseline balance of observable variables as these could cancel each other out. In contrast, heterogeneity will be elevated as individual trial’s baseline imbalances will contribute to this irrespective of the direction of the difference. Clark et al.3 noted a baseline imbalance in age in two out of twelve meta-analyses published in major medical journals, with significant heterogeneity in eight. They also identified that other variables more predictive of outcome than age demonstrated greater heterogeneity, and consequently suggested using another prognostic variable alongside age in baseline meta-analyses to assess the validity of outcome results.4

For individually randomised trials many statisticians argue that testing of baseline variables is misleading and should not be done5. The argument against baseline testing in trials is based on the fact that if randomisation is conducted properly, then any difference between the two groups will solely be due to chance. Baseline testing, therefore, is a fruitless activity and can result in misleading findings if it influences the outcome analysis. For example, if the decision is made to include a variable seen to be statistically significantly imbalanced at baseline in the analysis model but this variable has no relationship to the outcome then the analysis will ‘correct out’ some of the randomisation (as it correlates with this) and result in a biased estimate. On the other hand, Berger argues that baseline testing in individual trials may indicate where there have been allocation problems6. In this paper we are concerned with baseline testing across randomised trials to identify potentially biased studies rather than baseline testing within a trial.

A recent study by Ker et al.7 demonstrated serious flaws in a number of trials investigating the effects of Tranexamic acid on the prevention of postpartum haemorrhage, and a meta-analysis of baseline variables suggested an inadequate randomisation process in many cases. Whilst existing evidence suggests that many meta-analyses contain flawed trials, this does not mean that most trials are not fit for purpose. It is the aim of this paper to describe a straightforward statistical technique that can be incorporated into meta-analyses to identify and remove potentially biased trials.

**Dealing with baseline heterogeneity: a worked example**

We took a single meta-analysis from the *British Medical Journal* (BMJ) to demonstrate our approach. The systematic review ‘Effects of interventions in pregnancy on maternal weight and obstetric outcomes: meta-analysis of randomised evidence’ by Thangaratinam et al. 20128, containing 44 RCTs9-52, was chosen because Clark et al. had identified this as a review with significant heterogeneity (I2 = 50%) and imbalance in baseline age3. The heterogeneity was identified through the use of a fixed effects meta-analysis of age. A fixed effects approach was used because we *know* that the null is true and all of the trials will be estimating the same difference: that is there is no difference in age. Consequently, the only legitimate source of between study variation is due to chance, which is accounted for in a fixed effects model.

The following information was extracted for each component RCT where available: number of participants, a summary of all reported baseline variables (mean or median) and their dispersion (standard deviation (SD), standard error (SE), range, interquartile range), outcome data (reported value, dispersion and significance level) and allocation concealment information. When RCTs contained three arms we grouped into two arms as per Thangaratinam et al. We did not contact the authors of the RCTs for further information when data was not reported in the published articles.

The most commonly reported baseline variables across the component trials were age, weight of mother (kg) and BMI. Standard approximation formulas were applied when these variables were not reported as mean and SD (e.g. if median and range were presented instead53), and conversion of SE to SD was applied where necessary. For each trial, the t-statistic for the difference in baseline variables between treatment arms was calculated54, and studies ranked by its absolute value. The t-statistic, calculated by dividing the difference in means by its standard error, is used to test the null hypothesis that there is no difference between the means of two samples. The larger the t-statistic, the less likely the difference has occurred by chance. For each of baseline age, weight and BMI, a fixed-effect meta-analysis was conducted and the heterogeneity, as quantified by the I2 statistic, recorded. The meta-analysis was repeated after removing the component RCT with the largest t-statistic. This process was continued until zero heterogeneity was observed, i.e. I2=0%. For validation the procedure was conducted in reverse, starting with the RCT with the smallest t-statistic, trials were added in until heterogeneity was observed (i.e., I2>0%) and the results compared.

The meta-analyses of the outcomes conducted in the original study were replicated, and then repeated once the studies contributing to heterogeneity in each baseline variable had been identified and removed. All meta-analyses were performed in Meta-Light55. Figure 1 provides an illustrative summary of the method used. Note that the baseline variables used in the meta-analysis were identified *a priori*.

Figure 1

An illustrative summary of the method used to identify and remove ‘suspect’ trials.

**Results**

We obtained full text versions for 43 studies through the University of York Library service. One study (Gomez15) was only available in abstract format; however, Thangaratinam et al. only obtained the abstract of this study too. Three studies15,33,37 were translated into English using Google Translate56.

*Replication of primary outcome data meta-analyses*

Exact replication of the original meta-analyses of outcome data proved difficult. A number of studies did not report the required data or provided figures that were unsuitable for inclusion. Furthermore, in some cases it appears that the authors wrongly used baseline data as the outcome variable (e.g. for Baciuk23) – that is instead of using follow-up measures of weight they accidentally used baseline weight in the meta-analysis. We suspect that Thangaratinam et al. obtained extraneous data by contacting authors of component trials, since they included data we could not find. Consequently, the results of our replicated primary outcome data meta-analysis differ slightly from the original meta-analysis.

*Analyses using baseline data*

A number of studies were excluded from the meta-analyses because baseline data was not reported: 17 excluded for age, 18 for baseline BMI, and 22 excluded for baseline weight. Figure 2 shows the flow of component RCTs included in the weight gain meta-analysis.

Figure 2

**Primary outcome: weight gain in pregnancy**

Involved in the systematic review:

44 RCTs

10 excluded:

* Did not use the outcome measure ‘weight gain in pregnancy’

Included by Thangaratinam et al:

34 RCTs

9 excluded:

* 4 = Presented ‘final weight’ data rather than ‘weight gain in pregnancy’
* 2 = No dispersion value reported
* 3 = Did not report outcome data

Included in our original meta-analysis:

25 RCTs

10 excluded:

* 4 = No baseline demographic data reported
* 5 = Did not report baseline age
* 1 = Did not report baseline weight

Reported baseline data for age **and** weight:

15 RCTs

3 excluded: contributed to baseline heterogeneity

* Crowtherw6 and Huiw38 for baseline age
* Barakatw26 for baseline weight in kg

Included in final meta-analysis:

12 RCTs

Comment: this flowchart highlights a limitation to the approach. For the primary outcome ‘weight gain in pregnancy’ the number of trials included in the final meta-analysis is halved compared to the trials reporting useable outcome data.

*Age*

In Table 1 we show the mean differences in age between the randomised groups for each study included in the meta-analysis and the t-statistic for that difference. We rank the studies by the absolute value of the t-statistics (highest to lowest) and show the resulting heterogeneity for the meta-analysis as each study is removed. The total heterogeneity for the 27 studies in the meta-analysis of baseline age was 35.4%. (Clark et al. were only able to obtain 20 trials for their meta-analysis, hence our total heterogeneity is lower - 50.1% vs. 35.4%. Five studies were systematically removed based on their t-statistic until no heterogeneity existed14,32,35,38,46. The same trials were rejected when adding in trials starting with the smallest t-statistic. A statistically significant baseline imbalance was present for age (raw mean difference 0.271, p=0.04), which was eliminated when trials with baseline heterogeneity were removed (mean difference 0.128, p=0.53).

**Table 1: Studies included in the meta-analysis of baseline age ranked by t-statistic for difference in age between randomised groups**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Study** | **Mean Difference (kg)** | **t-statistic** | **absolute value of t-statistic** | **Heterogeneitya I2 (%)**  (35.4% total) |
| Hopkins | 2 | 2.618964 | 2.618964 | 29.3 |
| Crowther | 0.8 | 2.319949 | 2.319949 | 25.5 |
| Santos | -2.6 | -2.306597 | 2.306597 | 12.8 |
| Marquez-Sterling | 3.5 | 2.142188 | 2.142188 | 1.07 |
| Hui | 1.4 | 1.738516 | 1.738516 | 0.0 |
| Clapp | 1 | 1.732051 | 1.732051 | 0.0 |
| Barakat | 0.9 | 1.615753 | 1.615753 | 0.0 |
| Guelinckx | -1.4 | -1.603391 | 1.603391 | 0.0 |
| Wolff | -2 | -1.542686 | 1.542686 | 0.0 |
| Barakat | 1 | 1.294297 | 1.294297 | 0.0 |
| Quinlivan | -1.2 | -1.265797 | 1.265797 | 0.0 |
| Jackson | -0.8 | -1.193001 | 1.193001 | 0.0 |
| Haakstad | 0.9 | 1.133327 | 1.133327 | 0.0 |
| Baciuk | 1.4 | 1.120274 | 1.120274 | 0.0 |
| Erkkola | 0.4 | 1.025268 | 1.025268 | 0.0 |
| Landon | 0.3 | 0.821535 | 0.821535 | 0.0 |
| Bung | -1 | -0.625135 | 0.625135 | 0.0 |
| Khoury | -0.2 | -0.479653 | 0.479653 | 0.0 |
| Phelan | -0.2 | -0.385095 | 0.385095 | 0.0 |
| Erkkola + Makela | 0.2 | 0.368133 | 0.368133 | 0.0 |
| Garshasbi | -0.21 | -0.328250 | 0.328250 | 0.0 |
| Asbee | 0.3 | 0.265534 | 0.265534 | 0.0 |
| Huang | 0.22 | 0.262568 | 0.262568 | 0.0 |
| Khaledan | 0.15 | 0.093955 | 0.093955 | 0.0 |
| Sedaghati | 0.02 | 0.026337 | 0.026337 | 0.0 |
| Huib | 0 | 0 | 0 | 0.0 |
| Vinterb | 0 | 0 | 0 | 0.0 |
| aheterogeneity observed in meta-analysis of baseline age when this study (and those with higher t-statistic) removed  bstudies with same t-statistic ranked according to sample size (largest first) | | | | |

*BMI*

Substantial heterogeneity (I2=67.2%) was observed in the meta-analysis of 26 studies reporting baseline BMI. The heterogeneity was reduced to zero after just one study was removed (Barakat34, t-statistic 10.72). Statistically significant mean differences in BMI were present (mean difference 0.44, p<0.001), but were eliminated once heterogeneity was removed (mean difference -0.029, p=0.91).

*Weight in kg*

The heterogeneity for the meta-analysis of baseline weight in kg, containing 22 studies, was initially 64.7%. We identified an apparent error in the Garshasbi study30 as the mean difference in baseline weight did not correlate with BMI. Fortunately, height was provided which allowed us to correct the error (it read 55kg for one group rather than 65kg). When this was corrected, total heterogeneity was 0% without removing any studies. The mean difference in weight was not significant between groups (mean difference 0.208, p=0.60).

Of the 44 studies in the review, 35 did not provide sufficient information to determine their methods of allocation concealment. Only five trials were considered to have an allocation concealment method judged as ‘low risk’ by the Cochrane Collaboration’s tool for assessing risk of bias57. The remaining four trials were considered to be at high risk of bias.

*Meta-analyses of primary outcomes incorporating baseline data*

The outcomes considered were the continuous outcomes of weight gain in pregnancy (kg) and baby’s birthweight (g), and the categorical outcomes of small for gestational age and large for gestational age. Nineteen studies reported baseline variables for both age and weight (BMI or weight in kg) and did not contribute to heterogeneity in any of the baseline meta-analyses 9,13-14,16-17,22-23,25,28,30-31,33,39,41-42,45-47,50,52.

The results of the outcome meta-analyses are displayed in Table 2. For birthweight and large for gestational age babies, a statistically significant difference was noted which was not present in the original meta-analysis by Thangaratinam et al. This difference was apparent when we originally replicated their meta-analysis, and persisted for the birthweight outcome when only the studies contributing to 0% heterogeneity were included. For weight gain in pregnancy, the results were unchanged compared to our replication of the full meta-analysis. Similarly, the relative risk of babies born small for gestational age remained non-significant.

Table 2 also shows the heterogeneity of the outcome measures. For two of the four outcome measures the heterogeneity actually increased with the removal of the studies that contributed to baseline heterogeneity. This difference was not large and could be due to chance; however, removal of studies that produce baseline heterogeneity may not affect outcome heterogeneity if there are other reasons for this, such as differences in trial populations or treatment quality or dose.

**Table 2: Results of the outcome meta-analyses**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Outcome variable** | **Meta-analysis** | **No. of studies** | **Effect size (95% CI)d** | **p-value** | **Heterogeneity I2 (%)** |
| Weight gain in pregnancy (kg) | Originala | 34 | -1.42 (-1.89, -0.95) | <0.001 | 80 |
| Replicationb | 25 | -1.69 (-1.74, -1.65) | <0.001 | 88.7 |
| After identified trials removedc | 12 | -1.67 (-1.85, -1.49) | <0.001 | 80.5 |
| Baby’s birthweight (g) | Originala | 31 | -50.0 (-100.0, 0.0) | 0.08 | 57 |
| Replicationb | 27 | -59.3 (-90.3, -28.4) | <0.001 | 69.2 |
| After identified trials removedc | 13 | -66.6 (-110.0, -22.9) | <0.001 | 70.9 |
| Small for gestational age | Originala | 11 | 1.00 (0.78, 1.28) | 0.99 | 0 |
| Replicationb | 10 | 0.99 (0.77, 1.27) | 0.998 | 0 |
| After identified trials removedc | 4 | 1.11 (0.77, 1.60) | 0.976 | 0 |
| Large for gestational age | Originala | 18 | 0.85 (0.66, 1.09) | 0.21 | 38 |
| Replicationb | 18 | 0.76 (0.65, 0.90) | 0.0435 | 39.6 |
| After identified trials removedc | 8 | 0.76 (0.58, 0.98) | 0.0731 | 46 |
| aAs reported in the published meta-analysis by Thangaratinam et al.  bOur replication of the meta-analysis  cThe meta-analysis repeated when the studies identified as contributing heterogeneity to the meta-analysis of baseline age and BMI are removed  dMean difference for weight gain and baby’s birthweight, and relative risk for small- and large for gestational age | | | | | |

**Discussion**

Important clinical decisions are made on the basis of conclusions from systematic reviews and meta-analyses, yet they can produce flawed results when bias is introduced58. It is important, therefore, that we attempt to identify ‘suspect’ RCTs within meta-analyses and at the very least undertake a sensitivity analysis to assess whether results change when we remove papers demonstrating baseline imbalances. Previous work on the impact of allocation issues tend to compare papers that have reported poor allocation concealment or those who have not. Some have found a strong relationship with reported allocation concealmentwhilst others have not59,60. However, this approach relies upon the allocation method being reported accurately, which is not always the case, and this will mask any issues with poor allocation methods.

In this article we have demonstrated a relatively simple method to investigate the potential effect of biased randomisation. We argue that heterogeneity in a meta-analysis of baseline variables should not exist, and therefore removing trials which contribute to heterogeneity from a meta-analysis will produce a more valid result. We confirmed the validity of this method by both removing and adding studies whilst observing their impact on overall heterogeneity, and confirmed that the same results are observed in this example. In this instance the results of the outcome meta-analyses, as shown in Table 2, did not greatly alter the findings, however we believe it is important to remove trials with baseline imbalances to check that the results do not change.

Our results demonstrate that, as well as eliminating baseline heterogeneity, baseline imbalance is likewise removed. Analysis of both baseline imbalances and heterogeneity is advised, as a single method might not show the effects alone. For example, although the meta-analysis of baseline weight by Trowman et al. had 0% baseline heterogeneity it still demonstrated statistically significant baseline imbalances. It is possible that this occurred because a number of the included trials have allocation subversion all favouring the same treatment arm. Overall a review of baseline data should expect to see no heterogeneity and no statistically significant differences between groups.

Of the six excluded studies (five contributing to heterogeneity for baseline age, and one contributing for baseline weight in kg), three did not report their method of allocation concealment and two used envelopes, a technique which is generally not recommended61. Although one of the two used the sealed opaque numbered envelope (SNOSE) system, which is generally seen as a robust method, it does not prevent advanced opening of envelopes which can result in biased allocation62. This supports the possibility that such trials may be flawed and were appropriately excluded on the basis of baseline heterogeneity between treatment groups. Crowther14 contributed to baseline heterogeneity, yet this was a relatively large study (total sample size 1000) using a central randomisation process. This does not exclude the possibility of allocation subversion; alternatively it may be a chance finding.

It is important to distinguish between allocation misconduct and data fabrication. Both are forms of fraud; however, with misallocation there are actually ‘real’ trial participants, whereas with data fabrication ‘trials’ may include fictitious participants. The statistical methods of dealing with these two problems are different. In our study we focus on identifying possible misallocation rather than data fabrication. For the latter other statistical methods have been described by Carlisle63.

**Limitations**

The main limitation to this approach is the consequent exclusion of studies that do not report baseline data. In our example, a number of studies were excluded on this basis. This could impact on the minimum number of trials required to make a meta-analysis valid. Nevertheless, the reporting of baseline demographics is a feature of a good quality study so one may argue that exclusion of these studies is appropriate. As with an outcome meta-analysis, the method works best with a greater numbers of studies.

**Recommendations for authors**

We propose that analysis of baseline variables should be incorporated into meta-analyses routinely. The approach does not require in depth details of randomisation processes which are often not accurately reported, yet it can help provide assurances that the outcomes reached are more likely to be valid.

Ideally the chosen baseline variables should be pre-specified in the protocol to avoid the risk of bias. We propose that age, an important predictor of outcome in most instances, should always be used. Additionally, this is a variable which can easily, consciously or unconsciously, be used to subvert allocation3. The inclusion of a second variable strengthens the process, and this is recommended to be a variable that strongly predicts outcome4. If the primary outcome is also measured at baseline (e.g. body weight, pain severity) then this should be used. If the primary outcome is not measured at baseline (e.g. cancer recurrence) then another key predictor should be chosen and pre-specified.

The advantages of this method, as opposed to a sensitivity analysis based on the identification of studies that report a poor allocation method, is that it is not dependent on accurate methodological reporting of the allocation method, or the somewhat subjective judgement of its risk of bias. On the other hand, more recent trials do tend to better report the allocation method – so this approach could be used in conjunction with methods based on reporting of allocation concealment.

Further work on a greater number of meta-analyses is advised to investigate the efficacy of this technique in more detail.

**References**

1. Torgerson DJ and Torgerson CJ (2008). Chapter 3: Bias in randomised controlled trials. *Designing Randomised Trials*. Hampshire: Palgrave Macmillan; 2008.
2. Trowman R, Dumville JC, Torgerson DJ, Cranny G. The impact of trial baseline imbalances should be considered in systematic reviews: a methodological case study. *J Clin Epidemiol* 2007;60:1229-33.
3. Clark L, Fairhurst C, Hewitt C, Birks Y, Brabyn S, Cockayne S et al. A methodological review of recent meta-analyses has found significant heterogeneity in age between randomised groups. *J Clin Epidemiol* 2014; Sep;67(9): 1016-24.
4. Clark L, Fairhurst C, Cook E, Torgerson DJ. Important outcome predictors showed greater baseline heterogeneity than age in two systematic reviews. *J Clin Epidemiol* 2015 Feb;68(2):175-81.
5. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials *Lancet* 2000;355:1064-69.
6. Berger VW. *Selection Bias and Covariate Imbalances in randomized clinical trials.* (Chichester, England: Wiley, 2005).
7. Ker K, Shakur H, Roberts I. Does tranexamic acid prevent postpartum haemorrhage? A systematic review of randomised controlled trials. *BJOG* 2016 Oct;123(11):1745-52.
8. Thangaratinam S, Rogozinska E, Jolly K, Glinkowski S, Roseboom T, Tomlinson JW et al. Effects of interventions in pregnancy on maternal weight and obstetric outcomes: meta-analysis of randomised evidence *BMJ* 2012; 344:e2088.
9. Jackson RA, Stotland NE, Caughey AB, Gerbert B. Improving diet and exercise in pregnancy with Video Doctor counseling: a randomized trial. *Patient Educ Couns* 2011;83:203-9.
10. Badrawi H, Hassanein MK, Badraoui MHH, Wafa YA, Shawky HA, Badrawi N. Pregnancy outcome in obese pregnant mothers. *JPerinatMed* 1992;20:203.
11. Bechtel-Blackwell DA. Computer-assisted self-interview and nutrition education in pregnant teens. *ClinNursRes* 2002;11:450-62.
12. Briley C, Flanagan NL, Lewis N. In-home prenatal nutrition intervention increased dietary iron intakes and reduced low birthweight in low-income African-American women. *JAm DietAssoc* 2002;102:984-7.
13. Clapp IJF. Diet, exercise, and fete-placental growth. *Arch Gynecol Obstet* 1997;260:101-8.
14. Crowther CA, Hiller JE, Moss JR, McPhee AJ, Jeffries WS, Robinson JS. Effect of treatment of gestational diabetes mellitus on pregnancy outcomes. *N Engl J Med* 2005;352:2477-86.
15. Gomez TG, Delgado JG, Agudelo AA, Hurtado H. Diet effects on the perinatal result of obese pregnant patient. [Spanish]. *Rev Colomb Obstet Ginecol* 1994;45:313-6.
16. Khoury J, Henriksen T, Christophersen B, Tonstad S. Effect of a cholesterol-lowering diet on maternal, cord, and neonatal lipids, and pregnancy outcome: a randomized clinical trial. *Am J Obstet Gynecol* 2005;193:1292-301.
17. Landon MB, Spong CY, Thom E, Carpenter MW, Ramin SM, Casey B, et al. A multicenter, randomized trial of treatment for mild gestational diabetes. *N Engl J Med* 2009;361:1339-48.
18. Ney D, Hollingsworth DR, Cousins L. Decreased insulin requirement and improved control of diabetes in pregnant women given a high-carbohydrate, high-fiber, low-fat diet. *Diabetes Care* 1982;5:529-33.
19. Quinlivan JA, Lam LT, Fisher J. A randomised trial of a four-step multidisciplinary approach to the antenatal care of obese pregnant women. *Aust NZJ Obstet Gynaecol* 2011;51:141-6.
20. Rae A, Bond D, Evans S, North F, Roberman B, Walters B. A randomised controlled trial of dietary energy restriction in the management of obese women with gestational diabetes. *Aust NZJ Obstet Gynaecol* 2000;40:416-22.
21. Thornton YS, Smarkola C, Kopacz SM, Ishoof SB. Perinatal outcomes in nutritionally monitored obese pregnant women: a randomized clinical trial. *J Natl Med Assoc* 2009;101:569-77.
22. Wolff S, Legarth J, Vangsgaard K, Toubro S, Astrup A. A randomized trial of the effects of dietary counselling on gestational weight gain and glucose metabolism in obese pregnant women. *Int J Obes* 2008;32:495-501.
23. Baciuk EP, Pereira RI, Cecatti JG, Braga AF, Cavalcante SR. Water aerobics in pregnancy: cardiovascular response, labor and neonatal outcomes. *Reprod Health* 2008;5:10.
24. Barakat R, Lucia A, Ruiz JR. Resistance exercise training during pregnancy and newborn’s birthsize: a randomised controlled trial. *Int J Obes* 2009;33:1048-57.
25. Barakat R, Cordero Y, Coteron J, Luaces M, Montejo R. Exercise during pregnancy improves maternal glucose screen at 24-28 weeks: a randomised controlled trial. *Br J Sports Med* 2011; Sep26,epubaheadofprint. Doi:10.1136.
26. Bell RJ, Palma SM. Antenatal exercise and birth-weight. *Aust NZJ Obstet Gynaecol* 2000;40:70-3.
27. Clapp JF, III, Kim H, Burciu B, Lopez B. Beginning regular exercise in early pregnancy: effect on fetoplacental growth. *Am J Obstet Gynecol* 2000;183:1484-8.
28. Erkkola R. The influence of physical exercise during pregnancy upon physical work capacity and circulatory parameters. *Scand J Clin Lab Invest* 1976;6:747-9.
29. Erkkola R, Makela M. Heart volume and physical fitness of parturients. *Ann Clin Res* 1976;8:15-21.
30. Garshasbi A, Faghih ZS. The effect of exercise on the intensity of low back pain in pregnant women. *Int J Gynaecol Obstet* 2005;88:271-5.
31. Haakstad L, BoK. Exercise in pregnant women and birthweight: a randomized controlled trial. *BMC Preg Childbirth* 2011;11:66.
32. Hopkins SA, Baldi JC, Cutfield WS, McCowan L, Hofman PL. Exercise training in pregnancy reduces offspring size without changes in maternal insulin sensitivity. *J Clin Endocrinol Metab* 2010;95:2080-8.
33. Khaledan A, Sh, Motahari Tabari NS, Ahmad Shirvani M. Effect of anaerobic exercise program on fetal growth in pregnant women. *HAYAT: J Faculty Nurs Midwifery* 2010;16:78.
34. Lee G, Challenger S, McNabb M, Sheridan M. Exercise in pregnancy. *Mod Midwife* 1996;6:28-33.
35. Marquez-Sterling S, Perry AC, Kaplan TA, Halberstein RA, Signorile JF. Physical and psychological changes with vigorous exercise in sedentary primigravidae. *Med Sci Sports Exerc* 2000;32:58-62.
36. Ong MJ, Guelfi KJ, Hunter T, Wallman KE, Fournier PA, Newnham JP. Supervised home-based exercise may attenuate the decline of glucose tolerance in obese pregnant women. *Diabetes Metab* 2009;35:418-21.
37. Prevedel T, Calderon I, DM, Adami H-O, RM. Maternal and perinatal effects of hydrotherapy in pregnancy. *Rev Bras Ginecol Obstet* 2003;25:53-9.
38. Santos IA, Stein R, Fuchs SC, Duncan BB, Ribeiro JP, Kroeff LR, et al. Aerobic exercise and submaximal functional capacity in overweight pregnant women: a randomized trial. *Obstet Gynecol* 2005;106:243-9.
39. Sedaghati P, Ziaee V, Ardjmand A. The effect of an ergometric training program on pregnants weight gain and low back pain. *Gazzetta Medica Italiana Archivioperle Scienze Mediche* 2007;166:209-13.
40. Yeo S, Steele NM, Chang MC, Leclaire SM, Ronis DL, Hayashi R. Effect of exercise on blood pressure in pregnant women with a high risk of gestational hypertensive disorders. *J Reprod Med* 2000;45:293-8.
41. Asbee SM, Jenkins TR, Butler JR, White J, Elliot M, Rutledge A. Preventing excessive weight gain during pregnancy through dietary and lifestyle counseling: a randomized controlled trial. *Obstet Gynecol* 2009;113:305-12.
42. Bung P, Artal R, Khodiguian N, Kjos S. Exercise in gestational diabetes. An optional therapeutic approach? *Diabetes* 1991;40(suppl2):182-5.
43. Ferrara A, Hedderson MM, Albright CL, Ehrlich SF, Quesenbery CP, Peng TP, et al. A pregnancy and postpartum lifestyle intervention in women with gestational diabetes mellitus reduces diabetes risk factors. A feasibility randomized control trial. *Diabetes Care* 2011;34:1519-25.
44. Guelinckx I, Devlieger R, Mullie P, Vansant G. Effect of lifestyle intervention on dietary habits, physical activity, and gestational weight gain in obese pregnant women: a randomized controlled trial. *Am J Clin Nutr* 2010;91:373-80.
45. Huang TT, Yeh CY, Tsai YC. A diet and physical activity intervention for preventing weight retention among Taiwanese child bearing women: a randomised controlled trial. *Midwifery* 2011;27:257-64.
46. Hui A, Back L, Ludwig S, Gardiner P, Sevenhuysen G, Dean H, et al. Lifestyle intervention on diet and exercise reduced excessive gestational weight gain in pregnant women under a randomised controlled trial. *BJOG* 2011;119:70-7.
47. Hui AL, Ludwig SM, Gardiner P, Sevenhuysen G, Murray R, Morris M, et al. Community-based exercise and dietary intervention during pregnancy: a pilot study. *Can J Diabetes* 2006;30:169-75.
48. Jeffries K, Shub A, Walker SP, Hiscock R, Permezel M. Reducing excessive weight gain in pregnancy: a randomised controlled trial. *Med J Aust* 2009;191:429-33.
49. Kulpa PJ, White BM, Visscher R. Aerobic exercise in pregnancy. *Am J Obstet Gynecol* 1987;156:1395-403.
50. Phelan S, Phipps MG, Abrams B, Darroch F, Schaffner A, Wing RR. Randomized trial of a behavioural intervention to prevent excessive gestational weight gain: the Fit for Delivery Study. *Am J Clin Nutr* 2011;93:772-9.
51. Polley BA, Wing RR, Sims CJ. Randomized controlled trial to prevent excessive weight gain in pregnant women. *Int J Obes* 2002;26:1494-502.
52. Vinter CA, Jensen DM, Ovesen P, Beck-Nielsen H, Jorgensen JS. The LiP (Lifestyle in Pregnancy) study: a randomized controlled trial of lifestyle intervention in 360 obese pregnant women. *Diabetes Care* 2011;34:2502-7.
53. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. *BMC Medical Research Methodology* 2005;5:13.
54. Swinscow TDV. Revised by Campbell MJ. Statistics at Square One [Online]. 9th ed. BMJ Publishing Group; 1997. Chapter 7: The t tests. [accessed Jan 2017]. Available from: http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/7-t-tests
55. Thomas J, Graziosi S, Higgins S, Coe R, Torgerson C, Newman M. MetaLight: software for learning and doing meta-analysis. 2011. London: EPPI-Centre, Social Science Research Unit, Institute of Education. Available at: <http://eppi.ioe.ac.uk/free-tools/meta-analysis/> [Accessed Jan 2017].
56. Google. *Google Translate.* Available from: <https://translate.google.co.uk/> [Accessed Jan 2017].
57. Higgins JPT, Altman DG, Sterne J. Chapter 8: assessing risk of bias in included studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). The Cochrane Collaboration; 2011. Available at: https://handbook.cochrane.org [Accessed Jan 2017]
58. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995 Feb 1;273(5):408-12.
59. Balk EM, Bonis PAL, Moskowitz H, Schmid CH, Ioannidis JPA, Wang C, Lau J. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* 2002;287:2973-82.
60. Unverzagt S, Prondzinsky R, Peinemann F. Single-centre trials tend to provide larger treatment effects than multicentre trials: a systematic review *J Clin Epidemiol* 2013;66:1271-80.
61. Torgerson DJ, Roberts C. Randomisation methods: concealment. *BMJ* 1999;319(7206):375-376.
62. Kennedy ADM, Torgerson DJ, Campbell AK, Grant AM. Subversion of allocation concealment in a randomised controlled trial: a historical case study. *Trials* 2017;18:204.
63. Carlisle JB. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. *Anaesthesia* 2017;doi:10.1111/anae.13938.