Journal of Physiotherapy 60 (2014) 170–173

Journal of PHYSIOTHERAPY

journal homepage: www.elsevier.com/locate/jphys

Appraisal | Research Note

Forest plots

Introduction

A systematic literature review can provide a robust answer to a clinical question by identifying individual studies that provide evidence relevant to the question and summarising their results. In the field of physiotherapy, systematic reviews commonly summarise the results of randomised trials that test the effect of an intervention.[1] However, systematic reviews can also summarise studies of the accuracy of diagnostic tests,[2] studies of the prevalence of a clinical condition,[3] studies of prognostic factors,[4] or other study types. If the included studies are sufficiently similar and the results can be obtained in the same format, a meta-analysis may be performed.[5,6] Meta-analysis is a statistical method used to summarise numerical data from the individual studies into one overall estimate. The results of the individual studies and the overall estimate from the meta-analysis are usually presented in a graph called a forest plot.

What is a forest plot?

Although forest plots have been used since the 1970s,[7] the name 'forest plot' was first used in 2001.[8] The name refers to the forest of lines produced when the results of multiple individual studies are plotted against the same axis. The Cochrane Collaboration's official definition[9] states: 'A forest plot is a graphical representation of the individual results of each study included in a meta-analysis together with the combined meta-analysis result. The plot also allows readers to see the heterogeneity among the results of the studies.' The forest plot provides a quick visual representation of overall effect estimates and study heterogeneity and is therefore considered to be a very powerful tool in meta-analysis.[5,6]

Forest plots of randomised trials

Continuous measures of treatment outcome

The forest plot presented in Figure 1 was generated using data from a systematic review of randomised controlled trials of rhythmic cueing to improve walking speed in people with Parkinson's disease.[10] The first column (on the left) lists the studies included in the meta-analysis. Next are three columns of data for the experimental group in each study: the mean walking speed (in cm/s), the standard deviation (to indicate how much walking speed varied among the participants) and the number of participants. The same three columns of data are then presented for the control group. These data are then used to generate a mean difference in walking speed (still in cm/s) between the groups for each study. In the first study, for example, walking speed improved by a mean of 6 cm/s in the experimental group and 1 cm/s in the control group. The mean difference is therefore 6 minus 1 = 5 cm/s. This is presented numerically and also graphically by plotting a blue square over the horizontal line. Blue squares to the right of the vertical line indicate that the study favoured cueing, whilst those to left favour control. Note that a 95% CI is presented in both the numerical presentation (by two numbers in parentheses) and the graphical display (by a horizontal black line). Confidence intervals have been discussed previously in this journal.[11,12] Briefly, each study provides an estimate of the true effect of cueing on gait speed in people with Parkinson's disease. If any of the studies were repeated, a slightly different result would be expected. Loosely speaking, the confidence interval indicates the range within which the true effect of cueing probably lies. The estimate from each study is plotted, with the vertical line presenting the line of no effect. The size of the squares denotes the weight given to the study, with larger squares reflecting

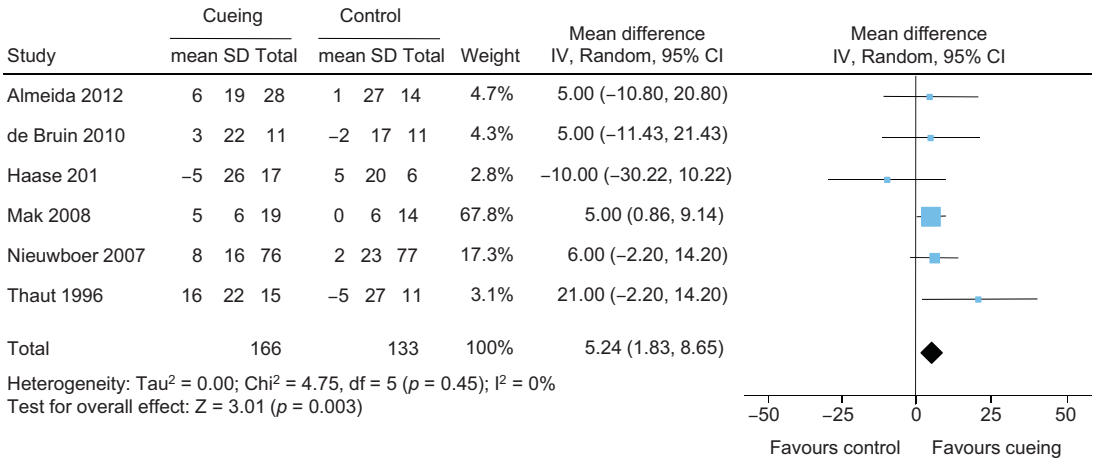

Figure 1. Meta-analysis of trials of the effect of cueing versus no cueing on gait speed (cm/s) in people with Parkinson's disease, using a random-effect model. A negative mean gait speed means a slower overall gait speed, SD = standard deviation, total = number of participants. Test for heterogeneity: $chi^2 = 4.75$, $I^2 = 0\%$, $p = 0.45$. Modified from the systematic review by Tomlinson and colleagues.[10]

http://dx.doi.org/10.1016/j.jphys.2014.06.021

more weight. The weight given to each study is automatically calculated based on the precision of the study's estimate (precise estimates receive more weight). All the individual estimates are then statistically pooled using meta-analysis to produce an overall estimate. This is presented graphically as a black diamond, where the centre of the diamond is the overall estimate and the width of the diamond is the overall confidence interval. The pooled estimate is also presented numerically.

The format of the forest plot presented in Figure 1 would be suitable for other continuous outcomes, such as activities of daily living or quality of life, where higher values are better. For continuous outcomes where lower values are better (such as pain intensity), values to the left of the vertical line would favour the experimental group. Also, the studies in Figure 1 all reported gait speed in more or less the same way and reported the results in cm/s, so these units can be retained throughout the analysis. If instead the group of studies had measured an outcome using a variety of tests (such as different scales for depression), the data could still be pooled, but each result would be reported as a standardised mean difference (SMD). To calculate the SMD, the data in the original units are divided by the standard deviation. This presents the size of the treatment effect independent of the scale used. Commonly, effect sizes below 0.3 are considered to be small, above 0.5 are considered to be moderate, and above 0.8 are considered to be large. However, these thresholds can be misleading if they are not interpreted in relation to the standard deviation of the study population.

Dichotomous measures of treatment outcome

Although forest plots for dichotomous outcomes are similar to those for continuous outcomes, some differences in format are required. The forest plot presented in Figure 2 is from a systematic review of trials of surgical techniques in people with chronic neck pain.[13] The outcome assessed in the studies is recovery, so the data columns now present the number of events (ie, people who had recovered during the follow-up period) and the total number of participants. The contrast between groups is now calculated as a risk difference (ie, the difference in the chance or 'risk' of recovery between the groups) and this is again presented numerically and graphically. Although Figure 2 shows risk difference, other statistics can be calculated and meta-analysed for dichotomous outcomes. One is relative risk, which is the ratio of the probability of an event in the treated and control groups. Another is the odds ratio, which is the ratio of the odds of recovery in the treatment group to the odds of recovery in the control group.

Interpreting forest plots

Statistical significance

When a confidence interval includes the vertical line of no effect, the result is statistically non-significant. This is evident graphically when the horizontal black line (for an individual study) or the black diamond (for the pooled estimate) crosses the vertical line of the forest plot. Therefore, the pooled estimate in Figure 1 is statistically significant but several of the individual studies are not. This illustrates the ability of meta-analyses to harness the statistical power of multiple studies to produce a more precise overall estimate, as shown by the narrower confidence interval of the pooled estimate than the individual studies.

Analysis of subgroups

Individual studies in a forest plot can be arranged in groups according to a characteristic of the patient population, intervention or follow-up. The estimate of the treatment effect in each subgroup can be compared to the overall effect estimate. Figure 3 presents a hypothetical subgroup analysis according to the age of participants. Subgroup differences can occur by chance, so a test for subgroup heterogeneity is presented. Subgroup differences can occur even when the test for heterogeneity does not indicate heterogeneity.

Another way that forest plots can reveal the relationship between the treatment effect and a characteristic of the participants, interventions or assessment is to place the studies on a vertical axis that shows the range of that characteristic. A forest plot with an extra vertical axis locating each study by a characteristic on a continuous scale (eg, the gender ratio of the participants, the duration of the intervention, or the time of assessment) is called a *modified forest plot*.[14]

Clinical relevance

Sometimes, an additional vertical line is added to the forest plot, indicating the threshold for clinical relevance; that is, the effect that is large enough to justify the cost, risks and inconvenience of the intervention. A hypothetical example of this is shown by the red dotted line in Figure 3. The location of the confidence interval in relation to this line and the line of no effect shows how it should be interpreted. The effect in children is statistically non-significant, whereas the other subgroups all show a statistically significant effect because their confidence intervals are all to the right of the line of no effect. Among the subgroups with a statistically

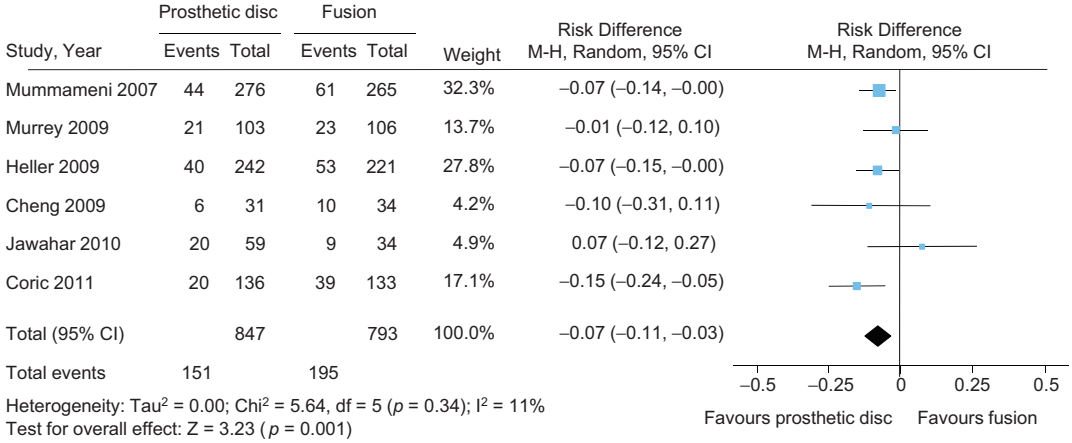

Figure 2. Meta-analysis of trials of the effect of prosthetic disc versus cervical fusion on recovery in people with chronic disabling neck pain, using a random-effects model and the Mantel-Haenszel method. Test for heterogeneity: $chi^2 = 5.64$, $I^2 = 11\%$, $p = 0.34$.
Modified from the systematic review by Verhagen and colleagues.[13]

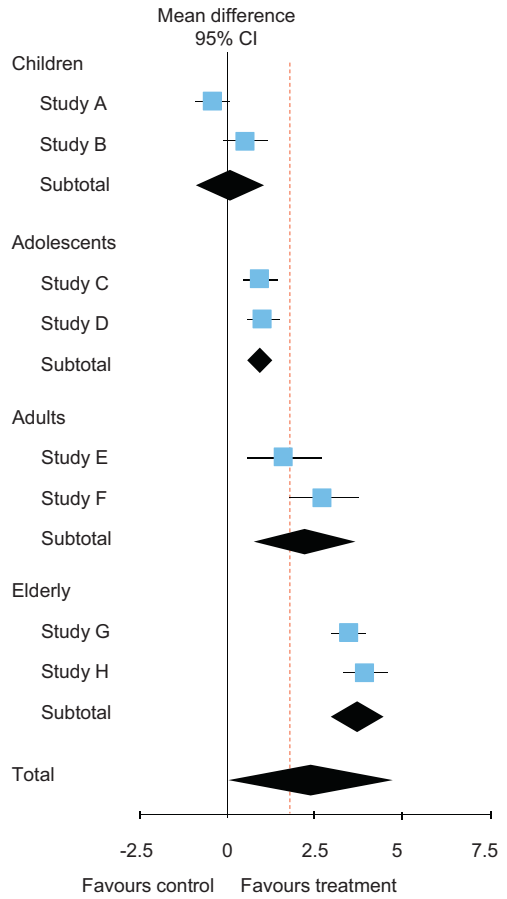

Figure 3. Hypothetical meta-analysis of trials, using a random-effects model, with subgrouping by age categories. Dotted vertical line represents the threshold for clinical relevance. Test for heterogeneity: $I^2 = 67\%$, $p = 0.04$.

significant effect, the effect in adolescents is clearly not worthwhile because the confidence interval is below the threshold for clinical relevance. The effect in adults may or may not be clinically worthwhile because the confidence interval crosses the threshold. The effect in the elderly is clearly worthwhile.

Heterogeneity

Heterogeneity refers to variability between studies and can affect the ability to combine the data of the individual studies. There are two types of heterogeneity: clinical heterogeneity and statistical heterogeneity. *Clinical heterogeneity* refers to the variability caused by differences in clinical variables, such as the patient population, interventions, outcome measures or setting of the included studies. Clinicians determine clinical heterogeneity, which means that it will always be a rather subjective decision. Readers should also consider these differences and subjectively decide whether the clinical heterogeneity is small enough for meta-analysis to be appropriate. *Statistical heterogeneity* is the variability in effect estimates between the studies and can be quantified by various statistics. Forest plots only present the statistical heterogeneity. The simplest statistic is the I^2, which quantifies the heterogeneity from 0 to 100%. There is no clear cut-point beyond which there is too much heterogeneity. Some use a rule of thumb stating that around 25% is low heterogeneity, around 50% medium and around 75% high heterogeneity.[15] Although other statistics, such as the Tau^2 or chi^2 are sometimes used, the I^2 does not suffer from some of the drawbacks of these other tests.[15]

Because forest plots provide a visual representation of study estimates, another approach is to simply view the variation between studies and judge the presence of heterogeneity ('eyeball' analysis). This subjective assessment of heterogeneity has high reproducibility (intra-class correlation = 0.87) and has a significant association with the presence of heterogeneity, as assessed by a statistical test,[7] suggesting this is a reasonable approach.

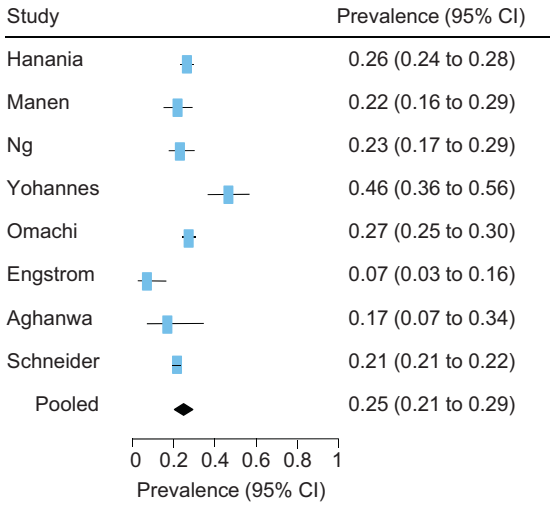

Figure 4. Forest plot of eight studies of the prevalence of depression among people with chronic obstructive pulmonary disease.
Modified from the systematic review by Zhang and colleagues.[3]

Forest plots of other study types

Prevalence studies

Estimates of the prevalence of a clinical condition from multiple observational studies can also be summarised in a forest plot. The individual and pooled estimates of prevalence and their confidence intervals are presented in a similar way to the forest plots discussed earlier. However, the estimates are plotted over the *x*-axis, which extends from 0 (no members of the population have the condition) to 1 (all members of the population have the condition). An example, which is summarising studies of the prevalence of depression among people with chronic obstructive pulmonary disease,[3] is presented in Figure 4. Note that the confidence intervals are symmetrical around prevalence estimates near 0.5, but they become increasingly asymmetrical as the estimates approach 0 or 1.

Reliability studies

Multiple reliability studies can also be presented on a forest plot. The example shown in Figure 5 is derived from a systematic review of reliability studies of the Berg Balance Scale.[2] The example summarises intra-rater reliability, with uniformly

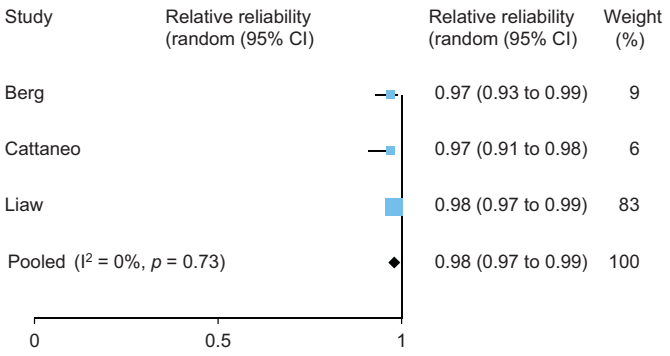

Figure 5. Forest plot of three studies of the intra-rater reliability of the Berg Balance Scale.
Modified from the systematic review by Downs and colleagues.[2]

Study	TP	FP	FN	TN	Sensitivity	Specificity
Firooznia 1984	97	4	8	7	0.92 (0.86 to 0.97)	0.64 (0.31 to 0.89)
Forristall 1988	20	2	4	5	0.83 (0.63 to 0.95)	0.71 (0.29 to 0.96)
Jackson 1989 I	89	25	36	81	0.71 (0.62 to 0.79)	0.76 (0.67 to 0.84)
Jackson 1989 II	35	8	24	53	0.59 (0.46 to 0.72)	0.87 (0.76 to 0.94)
Schaub 1989	13	6	5	5	0.72 (0.47 to 0.90)	0.45 (0.17 to 0.77)
Schipper 1987	140	7	57	31	0.71 (0.64 to 0.77)	0.82 (0.66 to 0.92)
Thornbury 1993	17	5	1	9	0.94 (0.73 to 1.00)	0.64 (0.35 to 0.87)

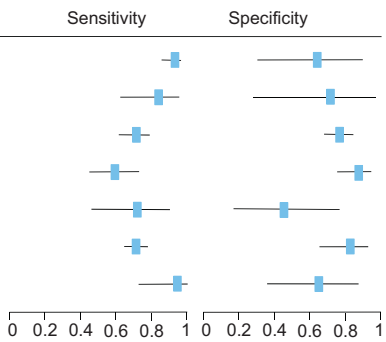

Figure 6. Forest plot of seven studies of the diagnostic accuracy of computer tomography (CT) scan compared to surgery as reference standard in people with low back pain. TP = true positive, FP = false positive, FN = false negative, TN = true negative.
Modified from the systematic review by van Rijn and colleagues.[16]

excellent results across all of the studies. Again the *x*-axis extends from 0 to 1 and the confidence intervals are asymmetrical. The same review uses exactly the same format for a forest plot of inter-rater reliability.[2]

Diagnostic accuracy studies

The forest plots of diagnostic accuracy studies differ from forest plots of treatment effectiveness because they show a double plot. The sensitivity and specificity estimates are presented together in one graph, as shown in Figure 6. In diagnostic test accuracy meta-analyses, the data presented entail true positive, false positive, false negative and true negative data for each study. Forest plots of diagnostic accuracy can provide the same information of pooled summary estimates and test heterogeneity. However, this is not recommended and therefore the program Review Manager of the Cochrane Collaboration does not provide this information together with the forest plots, but shows that in a receiver operating characteristic (ROC) graph.

Other variations

Forest plots may also report information about the statistical method used in the meta-analysis. For example, the meta-analysis in Figure 2 used the Mantel-Haenszel method, but other methods for dichotomous outcomes include the Inverse Variance method and the Peto method. Another possible variation in meta-analyses is whether a fixed-effect model or a random-effects model is used. Further information about these methods is available in the Cochrane handbook.[17]

Summary

Forest plots are frequently used in meta-analysis to present the results graphically. Without specific knowledge of statistics, a visual assessment of heterogeneity appears to be valid and reproducible. Possible causes of heterogeneity can be explored in modified forest plots. Forest plots in meta-analyses appear to be a valid and useful tool to quickly and efficiently scan and interpret the evidence. The expression 'a picture is worth a thousand words' certainly expresses the value of forest plots.

Arianne P Verhagen[a] and Manuela L Ferreira[b]
[a]*Department of General Practice, Erasmus Medical Centre University, Rotterdam, The Netherlands*
[b]*Musculoskeletal Division, The George Institute for Global Health, Sydney, Australia*

References

1. Moseley A, et al. *J Clin Epidemiol.* 2009;62:1021–1030.
2. Downs S, et al. *J Physiother.* 2013;59:93–99.
3. Zhang MWB, et al. *Gen Hosp Psychiatry.* 2011;33:217–223.
4. Fermont AJ, et al. *J Orthop Sports Phys Ther.* 2014;33:153–163.
5. Israel H, Richter RR. *J Orthop Sports Phys Ther.* 2011;41:496–504.
6. Callcut RA, Branson RD. *Respir Care.* 2009;54:1379–1385.
7. Bax L, et al. *Am J Epidemiol.* 2009;169:249–255.
8. Lewis D, Clarke M. *BMJ.* 2001;322:1479–1480.
9. Cochrane glossary. <www.cochrane.org/glossary> [Accessed May 4 2014].
10. Tomlinson CL, et al. *Cochrane Database Syst Rev.* 2013;9:CD002817.
11. Herbert RD. *Aust J Physiother.* 2000;46:229.
12. Herbert RD. *Aust J Physiother.* 2000;46:309.
13. Verhagen AP, et al. *Pain.* 2013;154:2388–2396.
14. Groenwold RHH, et al. *BMC Med Res Methodol.* 2010;65:253–261.
15. Higgins JP, et al. *BMJ.* 2003;327:557–560.
16. van Rijn RM, et al. *Eur Spine J.* 2012;21:228–239.
17. Cochrane handbook. <www.cochrane.org/handbook> [Accessed May 5 2014].