Combining and converting groups when extracting data for meta-analysis

Kathryn S Taylor, Kamal R Mahtani, Jeffrey K Aronson

Nuffield Department of Primary Care Health Sciences, University of Oxford, OX2 6GG, United Kingdom

Correspondence to: K Taylor kathryn.taylor@phc.ox.ac.uk

Word count: 1355

It is important that as much evidence as possible is included in the pooled analyses of systematic reviews. Therefore, systematic reviewers may want to analyse outcomes that are not completely specified in the reports of studies that they would like to include. However, using equations and other methods it is sometimes possible to fill in missing data and also check the validity of the reported data. Equations that are particularly useful are those for changing the grouping of patients from that which is reported. The desire to change groups often arises in systematic reviews and meta-analyses of interventions. For example, you may wish to combine the baseline data for two or more groups. Consider a clinical question as an example:

What is the effect of inhibiting the renin-angiotensin aldosterone (RAS) system, compared with placebo, on urinary albumin concentrations in patients with type 2 diabetes?

Patients with type 2 diabetes may be subgrouped into those with diabetic nephropathy (microalbuminuric) and those without diabetic nephropathy (normoalbuminuric), such as in a study by Sawaki *et al*,[1] which we shall use as an example. They reported the effects of losartan on urinary albumin, measured by the urinary albumin:creatinine ratio (UACR), expressed as mg of albumin per gram of creatinine (mg/g). Baseline characteristics were reported by treatment group, and some of the data needed for meta-analysis were reported for the subgroups within each treatment group.

When compiling a table of baseline characteristics of the study populations of the included studies in a systematic review, you may find that some studies report baseline data for one or more groups, split by treatment, or by another group variable, such as a measure of disease severity, while other studies only report baseline data for the whole trial population. A baseline characteristics table for the total trial populations for all studies presents more consistent information and also provides a convenient way to observe the variation of the population characteristics across the studies. We shall refer to this way of changing groups as *combining groups.* Another way of changing groups is when we want to carry out a subgroup analysis and a problem arises because of incomplete information about the subgroups. We shall refer to this as *converting groups for subgroup analysis*.

Baseline characteristics will include continuous variables, such as age, which will be presented as mean and SD, and categorical variables, such as diabetic status, which will be reported as the number and percentage of patients with diabetes. For meta-analysis of continuous outcomes, the required data from each study and for each treatment group are the mean and standard deviation (SD) of the outcome measure, and the number of participants. Note that this implies normally distributed data.

For skewed data that have been logarithmically transformed before analysis, results will generally be back-transformed (using the exponential function) and reported in terms of geometric means, with a measure of variability such as a confidence interval. Geometric means and arithmetic means (normally simply referred to as means) should not be pooled. There are methods for converting geometric means to arithmetic means[2] but these are beyond the scope of this article.

Here we present equations for deriving data required for meta-analysis using the reported summary data for both cases of changing the grouping, as shown in Figure 1. Shading indicates reported summary statistics and patterned circles indicates that summary statistics have been derived.

Figure 1. Using reported data when changing the groups

I – intervention; C – control/comparator.

The equations in Table 1 show how to pool the reported summary statistics of two treatment groups into those for the group of all patients combined (Figure 1: Combining groups).

Table 1. Equations for calculating the summary statistics of a combined group

Summary statistic	Group 1	Group 2	Combined group
N	n_1	n_2	$n_1 + n_2$
Mean	m_1	m_2	$\frac{n_1 m_1 + n_2 m_2}{n_1 + n_2}$
SD	SD_1	SD_2	$\sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2 + \frac{n_1 n_2}{n_1 + n_2}(m_1^2 + m_2^2 - 2m_1 m_2)}{n_1 + n_2 - 1}}$
Percentage	p_1	p_2	$\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

a × b is shown as ab. $\sqrt{\ }$ is the square root sign (for example $\sqrt{25} = 5$ because 5×5=25). n_1, n_2, n_c – number of patients in group 1, group 2 and the combined group; m_1, m_2, m_c – means; SD_1,SD_2,SD_c – standard deviations; p_1, p_2, p_c – percentages.

The pooled number, mean, and SD equations are shown in the Cochrane Handbook.[3] The pooled number is a sum, and the pooled mean and pooled percentage are weighted averages. The pooled SD is derived by expanding the equation for the SD of the outcome measure for the combined group, using the equations for the means and SDs of the subgroups and rearranging the equations.[4]

(Example of rearranging equations: $x = b + c$ and $y = b - c$ rearrange to $b = \frac{x+y}{2}$ and $c = \frac{x-y}{2}$)

If there are more than two groups, the equations can be used sequentially after combining data for groups 1 and 2, adding data from another group to the combined group at each stage.

In our example study there were 14 patients in the losartan group with baseline mean (SD) UACR of 61.7 (79.9) mg/g and 8 (57%) were women. In the control group there were 15 patients, the baseline mean UACR was 19.3 (31.2) mg/g, and 7 (47%) were women. Using the equations in Table 1 we can calculate:

number in the trial $= 14 + 15 = 29$

percentage who were women $= \frac{(14 \times 0.57 + 15 \times 0.47)}{29} = \frac{8+7}{29} = 52\%$

mean baseline UACR $= \frac{(14 \times 61.7 + 15 \times 19.3)}{29} = 39.8$ mg/g

SD baseline UACR

$$= \sqrt{\frac{(14-1) \times 79.9^2 + (15-1) \times 31.2^2 + \frac{14 \times 15}{29}(61.7^2 + 19.3^2 - 2 \times 61.7 \times 19.3)}{14 + 15 - 1}} = 62.6$$

The summary statistics of a particular subgroup (e.g. n_1, m_1, SD_1) may be calculated from the summary statistics of the complementary subgroup (n_2, m_2, SD_2) and the combined group (n_c, m_c, SD_c), using equations in Table 2 (Figure 1: Converting groups for subgroup analysis). These equations arise from rearranging the equations shown in Table 1.[4]

Table 2. Equations for calculating the summary statistics of a subgroup

Summary statistic	Combined group	Group 2	Group 1
N	n_c	n_2	$n_1 = n_c - n_2$
mean	m_c	m_2	$m_1 = \frac{n_c m_c - n_2 m_2}{n_c - n_2}$
SD	SD_c	SD_2	$\sqrt{\frac{(n_c-1)SD_c^2 - (n_2-1)SD_2^2 - \frac{n_c n_2}{n_c - n_2}(m_c^2 + m_2^2 - 2m_c m_2)}{n_c - n_2 - 1}}$
Percentage	p_c	p_2	$\frac{n_c p_c - n_2 p_2}{n_c - n_2}$

a × b is shown as ab. $\sqrt{\ }$ is the square root sign. n1, n2, nc – number of patients in group 1, group 2 and the combined group; m1, m2, mc – means; SD1,SD2,SDc – standard deviations; p1, p2, pc – percentages.

Returning to the example study, there were 6 patients in the losartan group with microalbumuria and their mean baseline UACR was 130.3(81.5) mg/g. The sex split was not reported for the subgroups.

Using these data and the reported mean baseline UACR for all the 14 patients in the losartan group, 61.7(79.9) mg/g, we can calculate the number of patients and their mean baseline UCAR for the losartan patients with normoalbuminuria by applying the equations in Table 2:

number of patients in the losartan group with normoalbuminuria $= 14 - 6 = 8$

mean baseline UACR $= \frac{(61.7 \times 14 - 130.3 \times 6)}{8} = 10.3$ mg/g

SD baseline UACR

$$= \sqrt{\frac{(14-1)\ \times 79.9^2 - (6-1)\ \times 81.5^2 - \frac{14 \times 6}{8}(61.7^2 + 130.3^2\ - 2\ \times 61.7\ \times 130.3)}{14-6-1}} = 7.3$$

In the control group in the example study there were 12 patients with normoalbuminuria with mean baseline UACR of 6.7(3.2) mg/g. Using these data and the reported mean baseline UACR for all 15 patients in the control group, 19.3 (31.2) mg/g, we can calculate the number and mean baseline UACR in the control patients with microalbuminuria using the equations in Table 2:

number of patients in the control group with microalbuminuria$= 15 - 12 = 3$

mean baseline UACR $= \frac{(19.3 \times 15 - 6.7 \times 12)}{3} = 69.7$ mg/g

SD baseline UACR

$$= \sqrt{\frac{(15-1)\ \times 31.2^2 - (12-1)\ \times 3.2^2 - \frac{15 \times 12}{3}(19.3^2 + 6.7^2\ - 2\ \times 19.3\ \times 6.7)}{15-12-1}} = 44.7$$

The equations in Tables 1 and 2 can be applied to other data, such as the values at the end of the trial or changes from baseline. These equations can also be used to check the validity of the reported data, which can be illustrated by considering the example study. This study reported standard errors for the microalbuminuria losartan subgroup and normoalbuminuria control subgroup. Based on these data, applying the equations produced a negative value for the baseline UACR SD for the losartan normoalbuminuria subgroup. This study was included in a systematic review[5] where other equations, not shown here, were used to derive the data that were pooled. The reported data were used in the main meta-analysis, and for the main subgroup meta-analysis the reported standard errors were assumed to be SDs. Also, because the reported subgroup summary data were questionable, this study featured in the sensitivity analysis. In one separate sensitivity analysis, this study was excluded from the subgroup analysis, and in another sensitivity analysis the SDs were imputed using the average of the baseline UACR SDs of the intervention groups from the other included studies.

Acknowledgements

This research was supported by the National Institute for Health Research Applied Research Collaboration Oxford and Thames Valley at Oxford Health NHS Foundation Trust. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Contributions

KT and KM conceived the idea of the series of which this is one part. KT wrote the first draft of the manuscript. All authors revised the manuscript and agreed the final version.

Competing interests

Dr Mahtani and Dr Aronson were Associate Editors of BMJ Evidence Medicine at the time of submission.

References

1. Sawaki H, Terasaki J, Fujita A, *et al*. A renoprotective effect of low dose losartan in patients with type 2 diabetes. *Diabetes Res Clin Pract.* 2008;79(1):86-90. doi:10.1016/j.diabres.2007.08.004

2. Higgins JP, White IR, Anzures-Cabrera J. Meta-analysis of skewed data: combining results reported on log-transformed or raw scales. *Stat Med*. 2008;27(29):6072-6092.

3. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook. [accessed 14 Oct 2020]

4. Taylor K. What if the summary statistics are given for the wrong group? https://www.cebm.ox.ac.uk/resources/data-extraction-tips-meta-analysis/[accessed 14 Oct 2020]

5. Hirst JA, Taylor KS, S+tevens RJ, *et al*. The impact of renin-angiotensin-aldosterone system inhibitors on Type 1 and Type 2 diabetic patients with and without early diabetic nephropathy. *Kidney Int.* 2012;81(7):674-683.