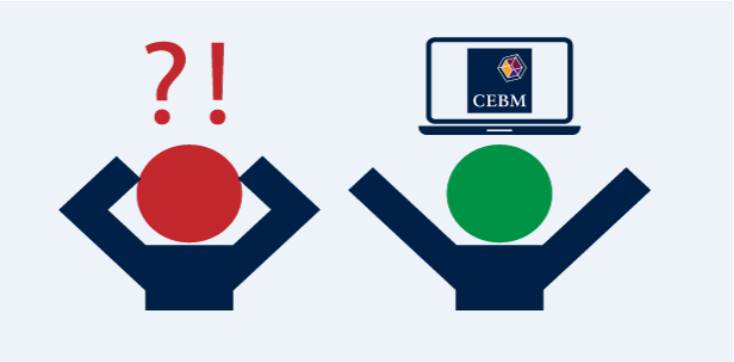
Tip for data extraction for meta-analysis – C7



What if neither the summary statistic I want, nor a similar statistic is reported?

Kathy Taylor

Previously, in post C1, I highlighted a list of ways where, when extracting data for meta-analysis of continuous outcomes, you might find that a summary statistic that you want is missing. In this post I'll focus on the 4th way – **neither the summary statistic you want, nor a similar statistic is reported.** This can arise in several different ways.

Sample sizes are not reported

Sometimes studies report the total number of patients and not numbers for each treatment group. The group equations that I showed before in post C3 can't be used as insufficient information is reported. However, these studies can be included in a meta-analysis using the generic inverse method (see section 10.3 in the Cochrane Handbook), where data are entered in the form of the appropriate effect estimate (for example, the mean difference) and its standard error (SE). For the study with missing sample sizes, the SE will be missing but this can be imputed. Imputation involves 'filling in' with a sensible value, such as the average SEs of the same treatment arms of other studies.

Missing mean and no other average measure

If a study has missing mean and the median also is not reported, the methods of Hozo, Bland and Wan that I mentioned previously in post C5 cannot be used. A study that doesn't report a mean may only report an effect estimate. This situation will be covered in my next blog post.

You may find that, instead of a mean, a study has reported a percentage change from baseline. If baseline values are also reported, you can calculate the mean final value:

$$Percentage\ reduction = \frac{Baseline - Final}{Baseline}$$

$$\Rightarrow Final = Baseline - Baseline \times Percentage\ change$$

and

$$Percentage\ increase = \frac{Final - Baseline}{Baseline}$$
$$\Rightarrow Final = Baseline + Baseline \times Percentage\ change$$

The symbol ⇒ means 'therefore'

You will need to impute the SD.

Parving 2001 reported that the urinary albumen excretion rate (UAER) reduced by 38% (32% to 40%) in the 300mg irbesartan treatment group, 24% (19% to 29%) in the 150mg irbesartan treatment group, and by 2% (-7% to 5%) in the placebo group. Baseline UAER values reported as 53.4 (2.2), 58.3(2.7) and 54.8 (2.5) µg/min respectively. We estimate the mean final urinary albumen excretion rates as:
53.4 - 53.4 x 0.38 =53.4 x 0.67=35.8
58.3 - 58.3 x 0.24 =58.3 x 0.76=44.3
54.8 - 54.8 x 0.02 =54.8 x 0.98=53.7

In a future post I'll look at the case where you want to pool final values but a study reports a percentage change and does not report baseline values.

Missing standard deviation and no other measure of variability

The Cochrane Handbook (6.5.2.3) shows that within group SDs may be calculated from summary statistics of a mean difference (MD). The MD, for which the more correct term is the difference of means, is the absolute difference between the mean values of a particular variable of the two groups in a randomised clinical trial.

Calculating a within-group SD from a SE of a MD:

$$SD = \frac{SE}{\sqrt{\frac{1}{n_{Intervention}} + \frac{1}{n_{Control}}}}$$

Note that this SD is the average of the SDs of the two groups and so it this same SD should be inputted into the meta-analysis for both groups.

Calculating a within-group SD from a CI of a MD:

A SE of a MD can be calculated from CI of the MD, as shown previously in post C6,

$$SE = \frac{(upper\ CI - lower\ CI)}{D}$$

For large samples (The Cochrane Handbook recommend this to be at least 60 in each group), the denominator (D) for MDs will be 3.92 for 95% CIs, 3.29 for 90% CIs and for 99% CIs. The denominators are the Z values from standard normal tables, which I showed before (see 'Where did the equations come from?'). For small samples, CIs for MDs should have been calculated from t-distributions and the denominators should therefore be the t-values from a t-distribution table which I used before.

Then having calculated the SE of the MD, the within group SD can be calculated from the SE, as shown above.

Calculating a within-group SD from p-value for a MD:

A SE of a MD may be calculated from a p-value by finding the associated t-value, taken from a t-distribution table.

For example, consider a trial with 20 participants in the intervention group 22 in the control group and a p-value of 0.01. We assume that this is a 2-sided probability.

$$dof = n_{Intervention} + n_{Control} - 2$$

dof = 20+22-2=40

From the t distribution table (Figure 1), the t-value is 2.704

You can also find the t-value from typing into an EXCEL cell
=TINV.2T(0.01,40).

$$SE = \left|\frac{MD}{t\ value}\right|$$

Then having calculated the SE of the MD, the within group SD can be calculated from the SE, as shown above.

dof	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
70	0.254	0.678	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
130	0.254	0.676	1.288	1.657	1.978	2.355	2.614	2.856	3.154	3.367
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Figure 1. t-distribution table

Note that if only p-value<0.05 is reported, the Cochrane Handbook suggest a conservative approach by using the upper limit i.e. p value=0.05. However, if p-value=NS (not significant) is reported we assume p-value>0.05 and we cannot calculate a SE, so we have to use imputation.

Dealing with missing SDs with imputation

If a large number of studies have no measure of variability, pooling data is not recommended. If only a small proportion of studies have no variability measure, and these studies will only contribute a small proportion of the data, you can deal with missing SDs by imputation, either using those included in your review, or from other meta-analyses. All the 'lending SD's should be similar and so it might be more appropriate to use the same-treatment SDs from that which is missing.

You could substitute the missing SD with a weighted average of SD from other studies

$$SD = \sqrt{\frac{\sum_N^1 SD_i^2 (n_i - 1)}{\sum_N^1 (n_i - 1)}}$$

This makes use of (n-1) that features in the calculation of the SD. This is Bessel's correction which corrects for bias.

Alternatively you should impute a SD with an unweighted average

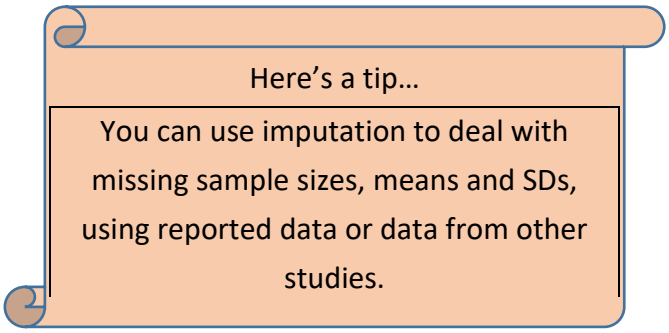
$$SD = \frac{\sum_N^1 SD_i}{N}$$

or take a conservative approach and substitute the missing SD with the highest valued available SD, as this will result on the lower weight given to the study.

More complicated imputation approaches include regressing the SDs of the same treatment from other studies onto other study covariates that are understood to be related to the missing SD. For example,

$$SD = \beta_0 + \beta_1 SD_{baseline} + \beta_2 X_2$$

The Cochrane Handbook highlights Marinho et al who, in their review of the preventative effect of fluoride toothpaste, dealt with missing data by predicting SDs from a linear regression of log(SD) on log(mean), citing the methods of the earlier review by van Rijkom et al to justify their use of a regression model.



In my next post, I'll focus on another example of the **4th way** of how a summary statistic that you want may be missing when dealing with continuous outcomes: **neither the summary statistic you want nor a similar statistic is reported.** I will pick up on what I mentioned above, that is, the case of a study only reporting an effect estimate.

Where did the equations come from?

(You can skip this if you are only interested in carrying out the calculations)

Calculating a within-group SD from a SE of a MD:

$$SD = \frac{SE}{\sqrt{\frac{1}{n_{Intervention}} + \frac{1}{n_{Control}}}}$$

In a previous proof I showed

$$Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$$

and when X and Y are independent, this becomes

$$Var(X - Y) = Var(X) + Var(Y)$$

If X and Y are independent, so are $\bar{X}$ and $\bar{Y}$. Therefore

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y})$$

For the proof of my last post I explained that the SE gives an estimate of the SD of its sampling distribution and that

$$Var(\bar{X}) = \frac{\sigma^2}{n} \approx \frac{s^2}{n}$$

Where s is the sample standard deviation and we assume that the two sample standard deviations are equal. Therefore,

$$SE(\bar{X} - \bar{Y}) = SD\ of\ the\ MD = \sqrt{Var(\bar{X} - \bar{Y})}$$

$$\Rightarrow SE(\bar{X} - \bar{Y})^2 = \frac{s^2}{n_1} + \frac{s^2}{n_2}$$

$$\Rightarrow SE(\bar{X} - \bar{Y}) = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\Rightarrow s = \frac{SE\ of\ MD}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Dr Kathy Taylor teaches data extraction in Meta-analysis. This is a short course that is also available as part of our MSc in Evidence-Based Health Care, MSc in EBHC Medical Statistics, and MSc in EBHC Systematic Reviews.

Follow updates on this blog, related news, and to find out about other examples of statistics being made more broadly accessible on Twitter @dataextips