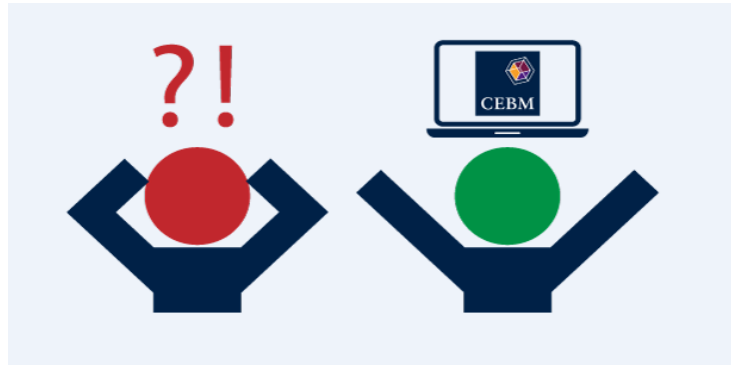


## Tip for data extraction in meta-analysis - 2



### What if the prevalence is also not reported?

Kathy Taylor

In my [last post](#), I showed some equations that you can use to construct the 2x2 diagnostic accuracy classification table using the reported prevalence, study size, sensitivity and specificity. Sometimes the prevalence is only reported for the whole study and not for subgroups. Recall that the prevalence is the proportion of those in the group of interest (whole study population or a subgroup) who have the disease. If the size of the subgroup and the number in the subgroup who have the disease is known:

$$\text{Prevalence} = \frac{\text{Number with the disease in the subgroup}}{\text{Total in the subgroup}}$$

If the size of the subgroup and the number in the subgroup who have the disease are not known, the prevalence may still be calculated if the positive predictive value (PPV) or the negative predictive value (NPV) are provided with the sensitivity and specificity. The PPV is the proportion of those who test disease-positive (1st row in Table given in my last post) who have the disease. The NPV is the proportion of those who test disease-negative (2nd row) who do not have the disease. The PPV and NPV are often reported with the sensitivity and specificity.

A bit of maths (see below if you're interested) shows us

$$\text{Prevalence} = \frac{PPV \times (1 - \text{Specificity})}{\text{Sensitivity} \times (1 - PPV) + PPV \times (1 - \text{Specificity})}$$

Or

$$\text{Prevalence} = \frac{\text{Specificity} \times (1 - NPV)}{NPV \times (1 - \text{Sensitivity}) + \text{Specificity} \times (1 - NPV)}$$

You have to be very careful when doing calculations that use brackets as it's easy to overlook a bracket or put it in the wrong place when you type up these equations in a calculator, spreadsheet

or computer program. Also note that statistics, such as sensitivity, can be reported as percentages or as decimal fractions. We're assuming that all these statistics are reported as fractions.

Let me show you an example from a review that I'm currently working on. In a [study](#) of the detection of asymptomatic left ventricular dysfunction, using N-terminal pro-brain natriuretic peptide, measured with the Elecsys Modular E device, at a threshold of 125 pg/mL and compared to echocardiography, for males aged under 67 years, the sensitivity is reported at 85.7%, specificity at 92.9%, NPV at 99.5% and the PPV at 30.0%. Prevalence is not reported for this subgroup but it is reported that, in this subgroup, 7 patients have diastolic dysfunction and 196 have normal left ventricular function. Therefore,

$$\text{Total in the subgroup} = 7 + 196 = 203$$

$$\text{Prevalence} = \frac{7}{203} = 0.034 \text{ or } 3.4\%$$

We may also use the other equations to calculate the prevalence, perhaps to check that the all the data makes sense. First note that this study reports percentage inputs, so we first need to convert these to decimal fraction inputs by dividing the percentages by 100 (e.g. sensitivity of 85.7% becomes 0.857).

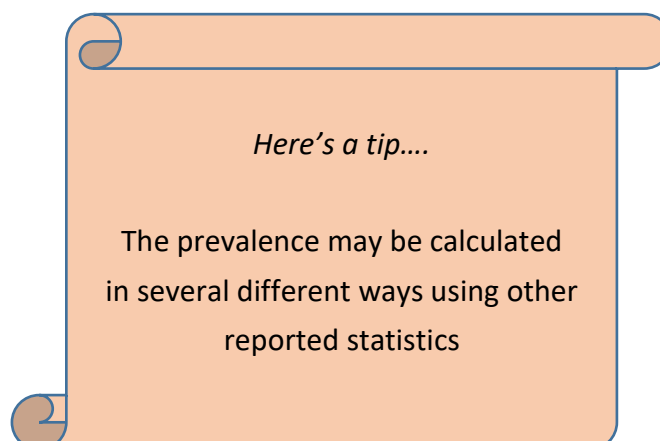
Using the PPV

$$\text{Prevalence} = \frac{0.3 \times (1 - 0.929)}{0.857 \times (1 - 0.3) + 0.3 \times (1 - 0.929)} = \frac{0.0213}{0.6338} = 0.034 \text{ or } 3.4\%$$

Using the NPV

$$\text{Prevalence} = \frac{0.929 \times (1 - 0.995)}{0.995 \times (1 - 0.857) + 0.929 \times (1 - 0.995)} = \frac{0.004645}{0.14693} = 0.032 \text{ or } 3.2\%$$

The slight differences arise from rounding errors.



In the next post I'll explain what you might do if a sensitivity or specificity is not reported.

Where did the equations come from?

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

[Previously](#), I showed the following equations

$$\begin{aligned}TP &= \text{Sensitivity} \times \text{Prevalence} \times \text{Total} \\FN &= \text{Prevalence} \times \text{Total} \times (1 - \text{Sensitivity}) \\TN &= \text{Specificity} \times \text{Total} \times (1 - \text{Prevalence}) \\FP &= \text{Total} \times (1 - \text{Prevalence}) \times (1 - \text{Specificity})\end{aligned}$$

I stated above that the PPV is the proportion of those who test disease-positive (1st row in Table given previously) who have the disease i.e.

$$PPV = \frac{TP}{TP + FP} \quad (\text{equation 1})$$

and the NPV is the proportion of those who test disease-negative (2nd row) who do not have the disease.

$$NPV = \frac{TN}{TN + FN} \quad (\text{equation 2})$$

To calculate prevalence using the PPV

Substitute for TP and FP in equation 1

$$PPV = \frac{\text{Sensitivity} \times \text{Prevalence} \times \text{Total}}{\text{Sensitivity} \times \text{Prevalence} \times \text{Total} + \text{Total} \times (1 - \text{Prevalence}) \times (1 - \text{Specificity})}$$

Cancel out Total and rearrange

$$\text{Prevalence} = \frac{PPV \times (1 - \text{Specificity})}{\text{Sensitivity} \times (1 - PPV) + PPV \times (1 - \text{Specificity})}$$

To calculate prevalence using the NPV

Substitute for TN and FN in equation 2

$$NPV = \frac{\text{Specificity} \times \text{Total} \times (1 - \text{Prevalence})}{\text{Sensitivity} \times \text{Total} \times (1 - \text{Prevalence}) + \text{Prevalence} \times \text{Total} \times (1 - \text{Specificity})}$$

Cancel out Total and rearrange

$$\text{Prevalence} = \frac{\text{Specificity} \times (1 - NPV)}{NPV \times (1 - \text{Sensitivity}) + \text{Specificity} \times (1 - NPV)}$$

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](#)