TP, FP, FN, TN Calculations Instructions & Troubleshooting

Document: how to calculate, collect, or derive TP, FP, FN, TN values from articles for a confusion matrix/for entropy calculations.

What are TP, FP, FN, and TN values?

TP: Stands for True Positive. FP: False Positive. FN: False Negative. TN: True Negative.

What am I looking for in the articles?

Some articles have the TP, FP, etc values all laid out for you, so if they are available, no need to calculate anything; just use those.

For articles that don't provide those values, there are four things you need to look for in order to calculate the values: **N (total # of patients), Sensitivity, Specificity, and Prevalence**. Almost always, if the study provides them, they will be in the figures section (so look for them in the tables/lists).

What if N, Sensitivity, Specificity, and/or Prevalence are missing?

Check first if they can be derived. I have noticed that prevalence is often (but not always) the only one missing when everything else is there. Check for mentions of prevalence in the article, as sometimes they "assume a prevalence." If they assume a prevalence of, for example, 14%, then you can use that number in your calculations. **Prevalence refers to the proportion of the population studied that actually has the condition being studied; it might not always be listed under the specific term "prevalence".** Other times, it may be listed as "XCondition %" (ex. AHF %). Just check to make sure.

If one of these is missing, then you will be unable to calculate the values for that study. Mention in the limitations as needed.

How do I calculate TP, FP, FN, and TN?

Here are the formulas:

True Positives (TP): Prevalence × Sensitivity × Total Patients
False Positives (FP): (1− Prevalence) × (1− Specificity) × Total Patients
False Negatives (FN): Prevalence × (1− Specificity) × Total Patients
True Negatives (TN): (1− Prevalence) × Specificity × Total Patients

Put #s into Excel and don’t calculate by hand for best accuracy. **Make sure the Prevalence, Sensitivity, and Specificity are in decimal form when computing.**

How do I put the calculations into Excel?

List the Study name in the 1st column, N (total # of patients) in the 2nd column, Prevalence in the 3rd, then Sensitivity in the 4th, and Specificity in the 5th.

Now make individual columns for TP, FP, FN, and TN.

For TP, the formula from above was Prevalence*Sensitivity*Total Patients, so you would input the formula **=ROUND(Cell name1*Cell name2*Cell name3)**.

(ex. If Prevalence is in C3, Sensitivity is in D3, and N (Total Patients) is in B3, the formula would be =ROUND(C3*D3*B3). Put “ROUND” so that you end up with a whole number. Drag the formula down to the other cells of that column.

For FP, the formula was (1-Prevalence)*(1-Specificity)*Total Patients so you would input the formula **=ROUND((1-Cell name1)*(1-Cell name2)*Cell name3).**

(ex. If Prevalence is in C3, Specificity is in E3, and N (Total Patients) is in B3, the formula would be =ROUND((1-C3)*(1-E3)*B3). Drag the formula down to the other cells of that column.

Repeat the process with formula adjustments for the columns of FN and TN.

FN formula would be in the format of **=ROUND((1-Cell name1)*Cell name2*Cell name3)**

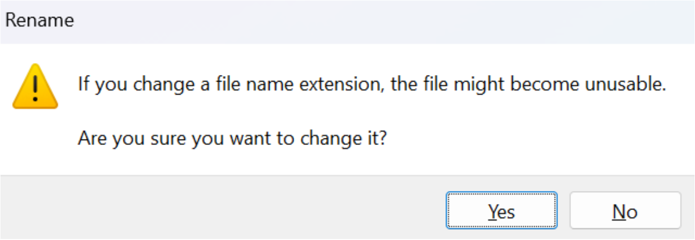
TN formula would be in the format of **=ROUND(Cell name1*(1-Cell name2)*Cell name3)**

Note that prevalence, sensitivity, and specificity are often given in % form, so you will need to change them into decimal form for the calculations. You can do this by adding another column next to in, and inputting the formula =Cell name/100 into the cells of that column (ex. If Specificity is in cell C3, the formula would be =C3/100). Drag the formula down to the other cells of that column.

What if the article is in PDF form and I cannot copy and paste the necessary information (N, Prevalence, etc…) without my cursor selecting a bunch of unnecessary information that won’t input into Excel in the correct column/row format?

Open your Notepad app on your computer. Select all the information and copy & paste it into Notepad (including the unnecessary parts). Go through it and delete all the information you *don’t* need, and separate all the information you *do* need with commas.

(ex. Study name, Prevalence, Sensitivity, Specificity)
For every different study, make sure you press Enter/start a new line so they are all individually on their own line.
(ex. AHF, 1278, 27.8, 45, 40
BIVA, 1500, 37, 20.5, 10)
You don't need a comma at the end of each line. Once you have done this for all your information, go to the upper left corner of Notepad and press "File" then "Save As". Since you can't save it as a csv from Notepad, just save it as a txt. Then, open your Folder or "File Explorer" and go to the download, click to rename it, and remove the txt part, replacing it with csv.
(ex. Turn "AHF, 1278.txt" to "AHF, 1278.csv")
This notification may pop up:

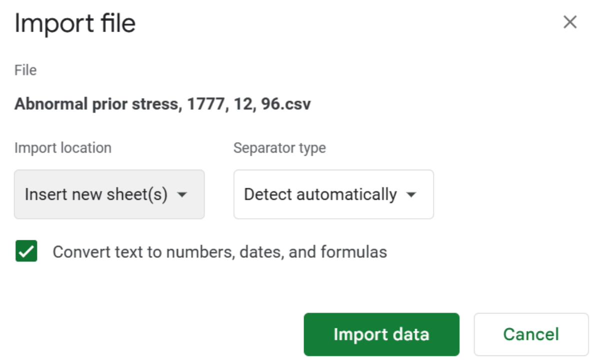


Rename

If you change a file name extension, the file might become unusable.

Are you sure you want to change it?

Click yes. Back to your Excel sheet, in the upper left, click "File" then "Import". Press "Upload" and upload the csv file into the sheet, choosing "Insert new sheet(s)" as Import location and "Detect automatically" as Separator type.



Import file

File

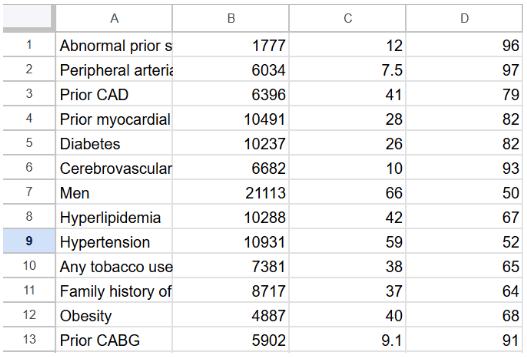
Abnormal prior stress, 1777, 12, 96.csv

Import location: Insert new sheet(s)

Separator type: Detect automatically

☑ Convert text to numbers, dates, and formulas

Once you click Import data, it will create another sheet with all your information neatly separated into cells like you want them to be, like this:



	A	B	C	D
1	Abnormal prior s	1777	12	96
2	Peripheral arteria	6034	7.5	97
3	Prior CAD	6396	41	79
4	Prior myocardial	10491	28	82
5	Diabetes	10237	26	82
6	Cerebrovascular	6682	10	93
7	Men	21113	66	50
8	Hyperlipidemia	10288	42	67
9	Hypertension	10931	59	52
10	Any tobacco use	7381	38	65
11	Family history of	8717	37	64
12	Obesity	4887	40	68
13	Prior CABG	5902	9.1	91

Formulas for Entropy Calculations:

The parent nodes refer to the total sample N, and the child nodes refer to the positive (TP, FP) and negative (FN, TN) tests. The representations for the number of positive and negative tests are n_positive and n_negative, respectively.

$$entropy_{parent\ node} = \left[\frac{FP+TN}{N}\times\left(\log_2(N)-\log_2(FP+TN)\right)\right] + \left[\frac{TP+FN}{N}\times\left(\log_2(N)-\log_2(TP+FN)\right)\right]$$

$$entropy_{child\ node\ 1} = \left[\frac{TP}{n_{positive}}\times\left(\log_2\left(n_{positive}\right)-\log_2(TP)\right)\right] + \left[\frac{FP}{n_{positive}}\times\left(\log_2\left(n_{positive}\right)-\log_2(FP)\right)\right]$$

$$entropy_{child\ node\ 2} = \left[\frac{FN}{n_{negative}}\times\left(\log_2\left(n_{negative}\right)-\log_2(FN)\right)\right] + \left[\frac{TN}{n_{negative}}\times\left(\log_2\left(n_{negative}\right)-\log_2(TN)\right)\right]$$

Entropy removal corresponds to the difference between the entropy of the parent node (the total entropy of the system) and the weighted average entropy of the children nodes (proportional to n_positive and n_negative, respectively).

$$entropy\ removal = \left[\left(\frac{FP+TN}{N}\times\left(\log_2(N)-\log_2(FP+TN)\right)\right)+\left(\frac{TP+FN}{N}\times\left(\log_2(N)-\log_2(TP+FN)\right)\right)\right]$$

$$-\left[\begin{array}{l}\left[\left(\frac{n_{positive}}{N}\right)\left(\frac{TP}{n_{positive}}\times\left(\log_2\left(n_{positive}\right)-\log_2(TP)\right)\right)+\left(\frac{FP}{n_{positive}}\times\left(\log_2\left(n_{positive}\right)-\log_2(FP)\right)\right)\right]\\ +\left[\left(\frac{n_{negative}}{N}\right)\left(\frac{FN}{n_{negative}}\times\left(\log_2\left(n_{negative}\right)-\log_2(FN)\right)\right)+\left(\frac{TN}{n_{negative}}\times\left(\log_2\left(n_{negative}\right)-\log_2(TN)\right)\right)\right]\end{array}\right]$$

What if some of the TP, FP, FN and TN values are 0, causing an error of #DIV/0! to show up when calculating removed entropy?

Essentially, if any of TP/FP/FN/TN have a value of zero, it practically makes that specific node of entropy calculation zero (there is no uncertainty if all variables are of one type and not the other), but it is mathematically undefined. So, to fix this, remove the nest of calculations containing the 0 value.
This resolves the mathematical error and gives the correct entropy calculations.

That's all!

Email me if you have questions.