# About the effect of doubling observed cases on the resulting true risk

Rosa Gini

Let $N$ be the number of people in a population, consider a condition that come people in the population suffer, and consider an algorithm to detect the condition. This identifies two subsets of the population, those who truly have the condition and those retrieved by the algorithm (see Figure 1). Let's define the following parameters

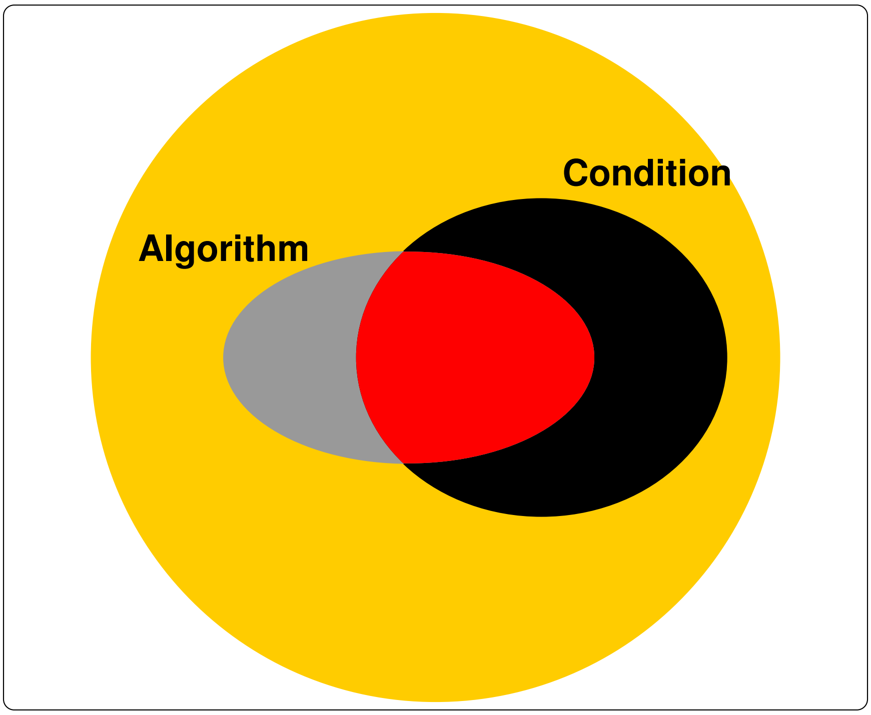
| | |
|---|---|
| **TP** | number of true positives: correctly retrieved by the algorithm |
| **FP** | number of false positives: retrieved by the algorithm but without the condition |
| **FN** | number of false negatives: with the condition, but not retrieved by the algorithm |
| $C$ | number of persons with the condition in the population |
| $n$ | number of people retrieved by the algorithm |
| $p$ | true frequency of the condition in the population |
| $P$ | observed frequency of the algorithm in the population |
| $PPV$ | positive predictive value of the algorithm |
| $SE$ | sensitivity of the algorithm |

By definition, the following formulas hold

$$
\begin{aligned}
C &= \mathbf{TP} + \mathbf{FN} \\
n &= \mathbf{TP} + \mathbf{FP} \\
p &= \frac{C}{N} && = \frac{\mathbf{TP} + \mathbf{FN}}{N} \\
P &= \frac{n}{N} && = \frac{\mathbf{TP} + \mathbf{FP}}{N} \\
SE &= \frac{\mathbf{TP}}{n} && = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}} \\
PPV &= \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}}
\end{aligned}
$$

From the definitions, the following equations hold

**Figure 1. A population with a condition and an algorithm that attempts to retrieve its cases**



$$\begin{aligned}
\mathbf{TP} &= \mathbf{TP} && \times \frac{\mathbf{TP}+\mathbf{FP}}{\mathbf{TP}+\mathbf{FP}} \\
&= (\mathbf{TP}+\mathbf{FP}) && \times \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}} \\
&= n && \times PPV
\end{aligned}$$

and

$$\begin{aligned}
C &= \mathbf{TP}+\mathbf{FN} \\
&= (\mathbf{TP}+\mathbf{FN}) && \times \frac{\mathbf{TP}+\mathbf{FP}}{\mathbf{TP}} && \times \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}} \\
&= (\mathbf{TP}+\mathbf{FP}) && \times \frac{\mathbf{TP}+\mathbf{FN}}{\mathbf{TP}} && \times \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}} \\
&= n && \times \frac{1}{SE} && \times PPV
\end{aligned}$$

Let’s denote $PPV$ by $v$ and $\frac{1}{SE}$ by $k$, then the equations become, respectively

$$\mathbf{TP} = vn$$

and

$$C = nkv$$

As a consequence, the number of false negatives is

$$\mathbf{FN} = C - \mathbf{TP} = kvn - vn = (k-1)vn$$

## Increasing the observed risk

Let's now choose a random set of $n$ subjects not retrieved by the algorithm, and let's define a new condition as the union of the previous condition and the new cases. We want to estimate the risk of the new condition relative to the previous condition, as the ratio between the cases of the new condition and the cases in the old, that is

$$RR = \frac{\text{previous cases } \textbf{OR} \text{ additional cases}}{\text{previous cases}}$$

To count the numerator we need to discount from the sum of the two sets their intersection, using the formula

$$|A \cup B| = |A| + |B| - |A \cap B|$$

The intersection is composed by those, among the $n$ additional cases, who are false negatives fro the algorithm. The proportion of false negatives among the additional cases is the same as the proportion of false negatives among all the $N - n$ negatives, that is $\frac{(k-1)vn}{N-n}$. This number can also be expressed in terms of $p$ as follows

$$\begin{aligned}\frac{(k-1)vn}{N-n} &= \frac{kvn - vn}{N-n} \\ &= \frac{\frac{kvn}{N} - \frac{vn}{N}}{1 - \frac{n}{N}} \\ &= \frac{p - \frac{p}{k}}{1 - \frac{p}{kv}}\end{aligned}$$

Among the $n$ subjects those who were not cases are $n$ minus the above proportion of $n$, that is

$$\begin{aligned}n - n\frac{p - \frac{p}{k}}{1 - \frac{p}{kv}} &= \frac{n - \frac{np}{kv} - np + \frac{np}{kv}}{1 - \frac{p}{kv}} \\ &= n\frac{1-p}{1 - \frac{p}{kv}}\end{aligned}$$

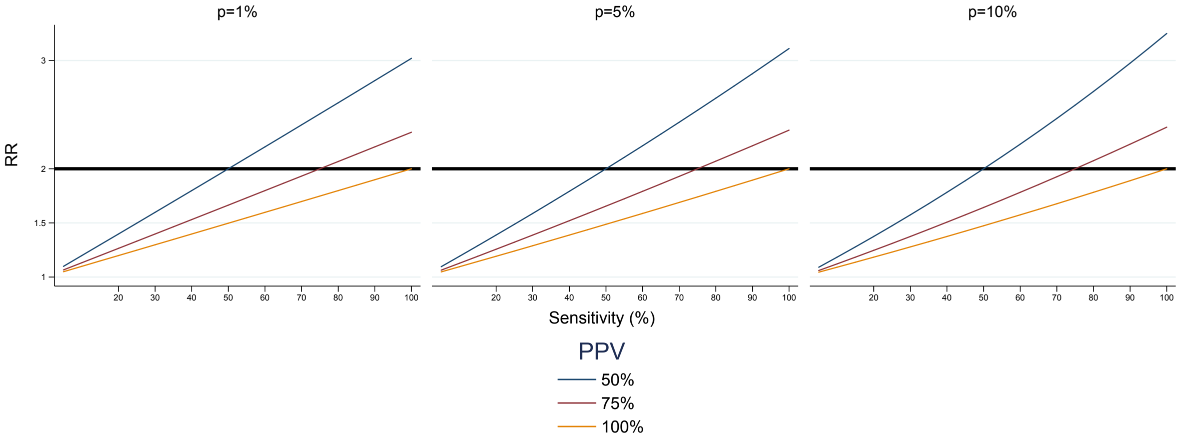
The relative risk of the new condition with respect to the previous is therefore

$$\begin{aligned}RR &= \frac{nkv + n\frac{1-p}{1-\frac{p}{kv}}}{nkv} \\ &= 1 + \frac{\frac{1-p}{1-\frac{p}{kv}}}{kv} \\ &= 1 + \frac{1}{kv}\frac{1-p}{1-\frac{p}{kv}}\end{aligned}$$

In Figure 2 this formula is represented for three values of $p$ (1%, 5% and 10%), for 3 values of PPV (50%, 75% and 100%), as a function of sensitivity.

The formula is 2 if $vk = 1$, that is, if sensitivity is equal to PPV, and, in particular, if they are both 100%. Otherwise the formula takes values that range from 1.5 to 3 for combinations of PPV and sensitivity that may easily occur, like low sensitivity and high PPV. For instance a sensitiviy of 60% and a PPV of 90%, would give RR from 1.66 to 1.64 in the three scenarios of prevalence.

**Figure 2. Risk of the new condition relative to the previous. The reference level is 2.**


p=1%
p=5%
p=10%
RR
Sensitivity (%)
PPV