# Obtaining estimates of outcome validity from a small set of parameters: the *component strategy* from the ADVANCE project

Rosa Gini[1], Caitlin Dodd[2,3], Kaatje Bollaerts[4], Miriam Sturkenboom[3,4]

(1) Osservatorio di Epidemiologia, Agenzia regionale di sanità della Toscana. Florence, Italy.
(2) Department of Medical Informatics, Erasmus Medical Center. Rotterdam, The Netherlands
(3) Utrech Medical Center, Utrecht, The Netherlands
(4) P95, Leuven, Belgium

## Background

The fundamental problem of epidemiology is assessing causal relationships between an exposure $E$ and an outcome $Y$

$E$ ——— $Y$

Database studies are observational, so associations may be caused by **confounding**. Moreover, in database studies the outcome $Y$ is not observed: existing data is processed to obtained a proxy $M_Y$, so associations, and heterogeneiy thereof, may be caused by **measurement error**.





We have methodologies to analytically address confounding specific to database studies (e.g.: propensity scores), but we don't have comparable tools to address measurement error

## Objective

Validity indices of $M_Y$ can be used to adjust effect estimates by misclassification errors. Conducting validation studies to estimate validity indices is often unfeasible, due to resource limitations or privacy issues.
We show that the complete set of validity indices can be analytically derived from a small set of input parameters.

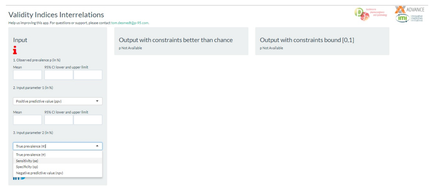
## One algorithm

It is easy to prove from definitions that the following system of 3 equations with 6 parameters hold

$$\begin{cases} P\ PPV = SE\pi \\ NPV(1-P) = SP(1-\pi) \\ P = SE\pi + (1-SP)(1-\pi) \end{cases}$$

$\pi$ true frequency
$P$ observed frequency
$PPV$ positive predictive value
$SE$ sensitivity

Since observed prevalence is a parameter that is always known, this implies that from knowledge of any other two parameters **the other 3 can be analytically derived** by solving the system. We developed a freely available tool that allows for computation of the derived indices from any given triplets, as well as uncertainty intervals




## Composition of two algorithms

Similarly, it can be proven that the validity of the composition of two algorithms $A$ and $B$ is interrelated with the validity of the components. This allows to **compute all the indices starting from any combination of 3 parameters** between validity indices of the components or of the composite, or true prevalence.

For instance, if $\pi$, $PPV_A$ and $PPV_B$ are known, then

$$\begin{cases} SE_{A\ \textbf{OR}\ B} = \dfrac{P_A\ PPV_A}{\pi} + \dfrac{P_B\ PPV_B}{\pi} - \dfrac{P_{A\ \textbf{AND}\ B} \max(PPV_A, PPV_B)}{\pi} \\ PPV_{A\ \textbf{OR}\ B} = \dfrac{SE\pi}{P} \end{cases}$$

$\pi$ true frequency
$P$ observed frequency
$PPV$ positive predictive value
$SE$ sensitivity

Or, if $SE_{A\ \textbf{OR}\ B}$, $PPV_A$ and $PPV_B$ are known, then

$$\begin{cases} \pi = \dfrac{P_A\ PPV_A}{SE} + \dfrac{P_B\ PPV_B}{SE} - \dfrac{P_{A\ \textbf{AND}\ B} \max(PPV_A, PPV_B)}{SE} \\ PPV_{A\ \textbf{OR}\ B} = \dfrac{SE\pi}{P} \end{cases}$$

Each database of the network participating in a multi-database study may define its study outcome as the composition (via OR logical connectors) of a particular set of components.

## Component algorithms

In the European EMIF project a component algorithms strategy was defined, where case-finding algorithms are split in simpler algorithms, each defined by a quadruple.

**Data domain** involved, among diagnosis, drug utilization, laboratory tests, ...

**Semantics** what is the meaning of the information that is searched?

**Setting** where was the information collected, among primary care, outpatient specialist care, inpatient care, emergency care, death

**Pattern** temporal rules

The rationale is that **the 4 dimensions of a component algorithm partially explain the validity**

This strategy is used and extended in the European ADVANCE Project.

Validity of the necessary input parameters can be estimated from ad-hoc validation studies, or obtained by assuming transportability of parameters found in the literature, or by developing scenarios.

**Application**
The problem of assessing validity of case-finding algorithms can be reduced to a small set of input parameters. The rest of the information is obtained empirically from observing the prevalence of the component algorithms and of their intersections.

**Conclusion**
This set of formulas may be implemented in the OHDSI set of tools and support exploration of the validity of the case-finding algorithms used to define study outcomes, based on information that can be found in the literature, and on empirical observation.

**Disclosure** This research received support from the Innovative Medicines Joint Undertaking under ADVANCE grant agreement Nr: 115557