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

Rosa Gini1, Caitlin Dodd2,3, Kaatje Bollaerts4, Miriam Strurkenboom2,4
(1) Osservatorio di Epidemiologia, Agenzia regionale di sanità della Toscana, Florence, Italy.
(2) Department of Medical Informatics, Erasmus Medical Center, Rotterdam, The Netherlands
(3)Utrecht 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 $Y$.

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 obtain a proxy $\pi$, so associations, and heterogeneity 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 $\pi$ 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. The rest of the information is obtained a proxy $\pi$, so associations, and heterogeneity thereof, may be caused by measurement error.

One algorithm

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

\[
\begin{align*}
P &= \frac{PPV}{SE} \\
SPV &= \frac{(1-P)}{SPV} \\
SE &= \frac{SPV(1-\pi)}{P}
\end{align*}
\]

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

\[
\begin{align*}
SE_{A,\&B} &= \frac{P_A PPV_A}{SE} \\
PPV_{A,\&B} &= \frac{SE_{A,\&B}}{SPV_{A,\&B}} \\
SPV_{A,\&B} &= \frac{PPV_A}{PPV_B} \\
\end{align*}
\]

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

\[
\begin{align*}
\pi &= \frac{PPV_A}{SE} \\
PPV_{A,\&B} &= \frac{PPV_A}{PPV_B} \\
PPV_{A,\&B} &= \frac{PPV_A}{PPV_B}
\end{align*}
\]

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