TITLE: **LEVERAGING HISTORICAL DATA FOR HIGH-DIMENSIONAL REGRESSION ADJUSTMENT, A MACHINE LEARNING APPROACH.**

Scientific Presentation made on June 2018 at the Promoting Statistical Insight Conference, London (UK).

Abstract

The amount of data collected from patients involved in clinical trials is continuously growing. All those patient's characteristics are potential covariates that could be used to improve study analysis and power. At the same time, the development of computerized systems simplifies the access to huge amount of historical data. However, it is still difficult to leverage those big data when dealing with small clinical trials, such as in Phases I and II. Their restricted number of patients limits the possible number of covariates included in the analysis.

The purpose of this talk is to present how machine learning can overcome this problem by taking advantage of historical data with larger sample sizes. Our approach is to pre-specify the combination of the baseline covariates by building a "meta-covariate". In small studies, using this meta-covariate alone will limit the loss of degrees of freedom while making the best uses of all generated data.

Two advantages of fitting the covariates on independent data are to free the modeling from the study constraints and to limit the risk of overfitting. Those are of particular interest with complex data, i.e non-normal distribution or in the presence of non-linearities. Our experiments show that the gain in power over standard approaches increases with the number of covariates or the decrease in the study sample size. Using simulated data, we also analyze the benefits of this methodology when the historical data are not representative of the study of interest. We also put the approach in perspective with the regulatory guideline on the use of adjustment for baseline covariates.

Authors

Samuel Branders, PhD; Guillaume Bernard, PhD; Alvaro Pereira, PhD