1.Objective

We aimed at developing a statistical test for the benefit of personalization, as well as innovative approaches to identify ITRs, and to apply these methods to real data.

2.Methods Used

The analysis considered a counterfactual outcomes framework, where we posit that for each patient, there exists two potential outcomes, Y(1) and Y(0), representing the outcome that the patient would experience should s/he receive the studied treatment (indexed by 1) or its comparator (indexed by 0), respectively. The counterfactual individual treatment effect D = Y(1) – Y(0) cannot be observed but it is possible to estimate the expected value of D given a set of covariates X, the conditional average treatment effect (CATE) d(X), and use it to select upon treatment (individualized treatment strategy). The original protocol planned to use a method we have developed to develop a treatment strategy based on d(X) while testing its benefit. In the meantime, Chernozhukov (Working Paper 24678, National Bureau of Economic Research, 2018) proposed methods for heterogeneity hypothesis testing, and in particular the Sorted Group ATE (GATES). The capacity of detecting heterogeneity and building subgroups of “good" and “bad" responders depends directly on the CATE predictions accuracy, so that these methods allow to evaluate the performance of individualized treatment strategies based upon the CATE. They also provide p-values. We therefore switched to these approaches.

We applied 12 models combining 3 meta-algorithms (T, S and X learners) with 3 regressors (linear regression, Lasso and Random Forests) base models. The hyper-parameters optimization was performed with a grid search evaluated by a 3-fold inside the training set. Each trained model discriminated the patients into 4 subgroups with the CATE predictions quantiles. The predictions of the trained model on the other dataset allowed us to compute p-values and confidence intervals.

3.Results

Overall, the best meta-algorithm in terms of R2 score on the outcomes was the S-learner. Regardless of the base model and the number of arms included in the training set, the S-learner model obtain equal or better scores in the outcome prediction than the T and X-learners. Using the three arms in the training set, the S-learner succeeded in explaining more than 10% of the control outcome variation, more than 20% of the 300mg arm outcome variations on the test set using the Lasso or Random Forest as base models.

Despite this, none of the setting described previously yielded significant results in the detection of the treatment effect heterogeneity.

4.Conclusions

On this example, the S-learner had better performance than more complex T and X-learners to predict outcomes. However, using an approach such as GATES, may limit the risks of claiming for a treatment effect heterogeneity, and therefore deriving spurious individualized treatment strategies.