Informing the Individualization of Treatment Using Pragmatic Trial Data

Dan Lizotte, Eric Laber and Susan A. Murphy
{danjl, laber, samurphy}@umich.edu

Abstract

Purpose: Practical trials which employ randomization produce rich data that allow comparison of the effectiveness of several different treatments on heterogeneous participants. They also enable investigation of how the relative effectiveness of different treatments varies with patient or illness features like age or symptom severity. We present exploratory data analysis tools that can be used to reveal relationships between prescriptive variables and the best choice of treatment, in order to suggest interesting hypotheses for future investigation.

Method: We use a bootstrap aggregation procedure to estimate the "selection probability" for each treatment—the probability that the treatment would be found optimal in a repeated trial and analysis. This quantity has a natural extension to the case of comparing more than two treatments simultaneously. We apply the method to data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial, where we use the most recent self-reported Quick Inventory of Depressive Symptoms (QIDS) score to predict the best subsequent treatment for individual patients.

Results: The selection probability, estimated using a bootstrap aggregation procedure, can be used to indicate possible interactions between patient variables and optimal treatment while mitigating the problem of spurious "false positives." The example using the STAR*D data shows how the results can be presented in a way that is intuitively interpretable by a wide audience.

Conclusion: Exploratory analyses of practical trial data as presented here can yield insights that suggest avenues for future research and inform clinical practice.

Quantifying Uncertainty

A measure of confidence for exploratory data analyses serves a different purpose than a measure of confidence for primary analyses.

- In primary analyses, a measure of confidence, e.g. the p-value, serves to indicate when there is strong evidence in the data for a true underlying effect.
- In exploratory analyses, a measure of confidence speaks to plausibility of hypotheses, which may be proven or disproven by future investigations.

In this setting, we allow ourselves to examine the data with less stringent significance criteria in order to recover potential hypotheses that are not supported to the degree required for verification, but that are still of interest. Because our focus is on generating hypotheses that will be verified experimentally in future trials, we ask, "What is the probability we would select this treatment as the best in a future study?"

If we are comparing more than two treatments, we ask, "What is the probability this treatment would be found best in each one of a series of independent, two-armed trials?"

We use a bootstrap aggregation procedure to estimate the pairwise selection probability for each pair of treatments, and use these to compute the probability that a treatment would be found superior to all of the others in a series of independent two-armed trials.

STAR*D

The Sequenced Treatment Alternatives to Relieve Depression trial was designed to evaluate the effectiveness of several different depression treatments. These results show how time to remission is related to a patient's depressive symptoms and to choices of treatment. We show the estimated optimal treatment, the estimated optimal value function, along with our measure of confidence, the selection probability.

These probabilities, in general, do not sum to 100% because for any complete set of pair-wise trials, there will sometimes be no overall winner (e.g., SER beats VEN, VEN beats BUP, but BUP beats SER). We call this type of result "discordant."

The hypothesis most strongly supported by this analysis is "For patients with high QIDS-SR, treating with SER is less effective than treating with VEN or BUP."

Why might this be? All patients had a course of citalopram, a selective serotonin reuptake inhibitor, for up to 12 weeks starting at the start of the study. SER also affects serotonin, VEN affects serotonin and noradrenaline, and BUP affects noradrenaline and dopamine. Therefore, one possible explanation for our findings is "Patients with low QIDS-SR have partially responded to serotonin-targeted therapy, and will improve further with serotonergic effects of SER or VEN. Patients with high QIDS-SR have not responded to serotonin-targeted therapy, and thus will respond better a drug that affects different neurotransmitters—norepinephrine for VEN and noradrenaline/dopamine for BUP."

Analysis Procedure

Confidence Intervals

In addition to providing a point estimate for the probability of selection, we can also compute bootstrap confidence intervals for the probability of selection of each treatment. This graphs show 95% confidence intervals for the selection probability of each of SER, VEN, and BUP.

In the 3-treatment case, if all treatments were equally effective, each would have a selection probability of 25%, with 25% left over for discordant results.

We acknowledge support for this project from National Institutes of Health grants R01 MH088015 and P30 DA010073. We also thank the STAR*D research team (http://www.star-d.org).