Exploratory Analysis of Practical Trial Data for Informing the Individualization Of Treatment

Lizotte, D. J., Laber, E., Rush, A. J., & Murphy, S. A.

The Pennsylvania State University

Technical Report Series
#09-96

College of Health and Human Development
The Pennsylvania State University
Exploratory Analysis of Practical Trial Data for Informing the Individualization Of Treatment

Daniel J. Lizotte†, Eric Laber†, A. John Rush‡, and Susan A. Murphy†

†University of Michigan
Ann Arbor, MI, United States
‡Duke-NUS Graduate Medical School Singapore
Singapore
Abstract

Objective: Practical randomized trials produce rich data that allow comparison of the effectiveness of several different treatments on heterogeneous participants. They also enable investigation of how relative effectiveness of different treatments varies with patient variables like age or symptom severity. The authors 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: Characteristics of practical clinical trial data are reviewed, and the use of such data for informing individualization of treatment is motivated. A method for suggesting potential strategies for treatment individualization is proposed, and its use is illustrated using data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study. The authors present an analysis that uses the self-reported Quick Inventory of Depressive Symptomatology (QIDS-SR) score to predict the best subsequent treatment.

Results: The proposed exploratory data analysis method provides a measure of confidence that estimates, for a given value of a prescriptive variable, the probabilities that each treatment would be selected to be the best if the trial and analyses were repeated. The authors relate the meaning of this quantity to confidence intervals for the standardized effect size, making the method a useful alternative interpretation of more traditional analyses.

Conclusion: Exploratory analyses of practical trial data as presented here can yield insights that suggest avenues for future research and eventually inform clinical practice.
Exploratory Analysis of Practical Trial Data for Informing
the Individualization Of Treatment

Historically, the majority of randomized clinical trials have been designed to establish that a new treatment is efficacious. Numerous treatments have been vetted in this way, most commonly by comparison against standard care and against placebo (Dawson & Lavori, 2004). This practice has provided clinicians with an extensive menu of options for treating the patients in their care, but has seldom offered guidance on how to select a treatment for an individual, in large part because such trials typically neglect the consideration of external validity (Rothwell, 2005a). This lack of guidance has persisted despite a clear desire on the part of clinicians to tailor treatments to individuals; for example, clinical experts have come together to develop guidelines for predicting which class of antidepressant would most likely benefit a particular patient (Stern, Rush, & Mendels, 1980). Since the discovery and approval of modern neurotransmitter reuptake inhibitors, the opportunity for tailored treatment of depression has increased even further.

Currently, there is significant interest in implementing large practical clinical trials designed to simultaneously investigate the comparative effectiveness of a set of treatments. For example, the STAR*D and CATIE (Clinical Antipsychotic Trials in Intervention Effectiveness) (Stroup et al., 2003) trials were designed to study the effect of a variety of appropriate treatments on a population similar to the population of patients a clinician encounters in practice (Tunis, Stryer, & Clancy, 2003). The usefulness of these trials stems from two important characteristics: First, they assess the comparative effectiveness of a variety of standard treatments. Second, the trials use broad inclusion criteria and thus acquire data on a heterogeneous group of patients who exhibit a range of treatment responses. The data produced by these trials are therefore useful for investigating the effect of treatments that are clinically relevant on the patients who are are likely to present in practice.
Primary analysis of these trials commonly consists of two-way comparisons between treatment groups. These comparisons use standard statistical analyses, e.g. hypothesis tests and confidence intervals, to convey the statistical significance of the observed difference in the outcome of each group. Note that these comparisons average over outcomes of all patients who fall in each group—populations that, as we have mentioned, can be quite diverse in practical trials. In this work, we focus on analyses that take advantage of this diversity to look inside treatment groups and investigate relationships between patient variables and outcomes that can inform the tailoring of treatment to individual patients. Our goal is to augment established sources of clinical knowledge using exploratory data analysis methods in a way that is useful for directing future scientific investigation in the service of informing clinical decisions.

Exploratory analysis of the data from a randomized trial is useful way of generating hypotheses about how treatment should be tailored; however, any conclusions drawn should not be attributed the same validity as the conclusion of the primary analysis for which the trial was designed. Rothwell (Rothwell, 2005b) states, “The best test of validity of subgroup-treatment effect interactions is their reproducibility in other trials.” We are in agreement. We therefore produce exploratory analyses with a measure of confidence that is relevant to future investigations by asking: “How much evidence is there in our existing data that an observed effect will replicate in a future trial?”

To begin, we review the concepts of qualitative and non-qualitative interactions and illustrate how their presence or absence can be estimated using regression. We then present our measure of confidence, which we term “selection probability,” and present two analysis examples, one using synthetic data and one using a subset of the STAR*D trial data. We describe how selection probability is related to the standardized effect size, and how a confidence interval for one relates to a confidence interval for the other.
The usefulness of a patient variable for tailoring treatment is determined by whether or not there is an “interaction” between that variable and treatment outcomes. For example, the patient variable could be a baseline variable like age or gender, or it could measure a patient’s response to a previous treatment. We say a “qualitative interaction” is present when the best treatment differs depending on the value of a patient variable; such a variable is termed “prescriptive.” The “best” treatment among those under consideration is the one that minimizes\(^1\) a chosen outcome. One example of a qualitative interaction between QIDS-SR and best treatment might be “Patients with lower QIDS-SR scores respond better to treatment A, while patients with higher QIDS-SR scores respond better to treatment B.” We say that a “non-qualitative interaction” is present when the best treatment is the same for all values of the variable, but the magnitude of its advantage over other treatments changes depending on the specific value of the variable. An example of a non-qualitative interaction would be “Patients respond better to B than they do to A. Patients with high QIDS-SR scores respond much better to B than to A, while patients with low QIDS-SR scores respond moderately better.” In the most extreme instance of a non-qualitative interaction, a treatment may be clearly superior for some values of a prescriptive variable, but essentially equivalent to its competitors for other values.

Knowledge of a qualitative interaction between a prescriptive variable and best treatment is useful because it can inform the choice of treatment for future patients. For example, if trial data indicate an interaction between baseline QIDS-SR score and best treatment, we may decide that in future states of clinical equipoise we should measure the baseline QIDS-SR of an incoming patient and, all else being equal, select the treatment that is expected to result in the best response for the measured value of QIDS-SR.

Knowledge of a non-qualitative interaction can be useful as well; suppose for example that
overall, patients respond better to B, but patients with low QIDS-SR scores respond equally well to B and A. Then for patients with low QIDS-SR, we may be more inclined to select treatment based on other desiderata of the patient and physician, such as avoiding certain side effects, despite the fact that B appears to be superior on average. For patients with higher QIDS-SR, we would eliminate treatment A from consideration.

Interactions can also reveal scientific insights and generate hypotheses about the relationship between patient characteristics, a treatment’s mechanism of action, and disease progression. For example, suppose we offer a first-line treatment T to a patient, and monitor their disease progression. After a time, if results are unsatisfactory, we consider two follow-up treatments A and B. Suppose A and T are considered similar, and B and T are considered dissimilar in terms of mechanism of action. We might therefore expect B to be better overall for patients who have failed T, but qualitative interactions could reveal prescriptive variables that describe the circumstances in which A is in fact a better follow-up choice. Discovery of such interactions can be a first step toward identifying the populations in which A and T act differently, and spur future scientific investigation into the reason for this difference.

**Interactions and Confidence**

We now introduce methodology for investigating the presence of qualitative interactions of prescriptive variables with treatment outcomes using data generated from practical clinical trials. For simplicity of exposition, we describe the method for the case where both the prescriptive variable and the patient outcome are continuous variables. This setting allows the use of an intuitive, regression-based approach to assess the presence of a qualitative interaction (“Advances in Clinical Trial Biostatistics”, 2004).

For each patient, the data includes the prescriptive variable of interest, the treatment provided, and the outcome of interest. First, we use the data to build a
regression model that can predict the expected patient outcome for any value of the prescriptive variable and any choice of treatment. Then, to predict the best treatment for patients with a particular value of the prescriptive variable, we query the regression model for the expected outcome given each different treatment at that fixed value of the prescriptive variable, and we note which treatment results in the smallest predicted outcome. If this procedure finds different best treatments given different values of the prescriptive variable, we have some evidence for a qualitative interaction, and hence some evidence that the variable in question is useful for tailoring treatment. Figure 1 shows a schematic representation of this process in the simplest case of comparing two treatments. In this example, the resulting regression indicates a qualitative interaction where treatment A is preferable for low values of the prescriptive variable, and treatment B is preferable for high values of the prescriptive variable. The difference between the predicted outcome for A and the predicted outcome for B is the estimated effect size for A vs. B. Since we prefer smaller outcomes, a negative estimated effect size indicates a preference for A and a positive estimated effect size reflects a preference for B.

A Measure of Confidence: Selection Probability

We use the estimated effect size to select a best treatment because it reflects our best estimate of the true effect size for A vs. B. However, if the true effect size is near zero and the individual patient outcomes are highly variable relative to the size of our dataset, even our best estimate of effect size will not be reliable, and therefore we will not be able to reliably select the true best treatment. Suppose we were to run many identical experiments with the same population and sample size. If we have a small sample size or large outcome variation (or both), we would observe large variations in estimated effect sizes from one trial to the next. If in addition the true effect size for A vs. B is near zero, these estimated effect sizes would frequently have different signs in different
trials—sometimes we would estimate a negative effect and select A, and sometimes we would estimate a positive effect and select B. Thus, in this setting, any particular selection of best treatment based on a single dataset is largely arbitrary. We therefore propose a procedure that provides a measure of our certainty that each treatment is best at each value of the prescriptive variable, rather than simply reporting our treatment selections based on the effect size estimated from the single dataset we have available.

To convey our confidence, we ask for each value of the prescriptive variable and each treatment T, “What is the probability that we would select treatment T to be best if we ran a new study using the same treatments, obtained a new dataset of the same size, and ran a new regression analysis using the same design?” We call this the “selection probability” for treatment T. At each value of the prescriptive variable, a hypothetical future dataset casts a “vote” for its selection of best treatment. If we were able to run many studies to obtain many datasets, we could estimate with arbitrary precision the selection probability for T by computing the proportion of votes received by T over all the datasets.

Suppose that our choice of regression model permits a good fit of the effect size for A vs. B. If this true effect size were negative—e.g. treatment A resulted in shorter time to remission than treatment B, on average—then the selection probability for A will be greater than 0.5. Three factors influence the selection probability for A: The effect size, the variance in observed outcomes, and the dataset size. For more negative effect sizes, the selection probability for A will be larger because we are more likely to correctly estimate the sign of the effect of A vs. B. For smaller variance in observed outcomes, the selection probability for A will be larger for the same reason. Finally, for larger datasets, the selection probability for A will be larger, again because we are better able to correctly identify the sign of the effect size.

If, on the other hand, the effect size of A vs. B were exactly zero, their selection
probabilities will be exactly equal (0.5 and 0.5) because any nonzero estimated effect will be due to chance variations in the patient outcomes recorded in the current dataset. If the effect size is near zero, outcome variability is large, or our dataset is small, the selection probability will be near 0.5, indicating that selecting A to be the best treatment using a single dataset gives only weak evidence that treatment A is in fact better than treatment B.

**Point Estimate and Confidence Interval for Selection Probability**

Obviously we do not mean to suggest that one should start running trials repeatedly just to estimate the selection probability for different treatments. We will instead estimate this quantity from the data we have available. Suppose we are comparing treatment A versus treatment B and estimating the selection probability for A. First, we generate some number of datasets (say 100) using the bootstrap resampling procedure (Efron, 1979; Efron & Tibshirani, 1993). We call these “outer bootstraps.” Each one is constructed by sampling patient records from the original dataset uniformly randomly with replacement. From each of the 100 outer bootstraps, we generate 100 more datasets using the bootstrap resampling procedure, which we call “inner bootstraps.” We then run our regression analysis on each of these inner bootstraps, and record the proportion of these analyses that selected treatment A. Thus each outer bootstrap is associated with a proportion of votes from its 100 inner bootstraps. We then take the average of these proportions over the 100 outer bootstraps to compute our estimate of the probability of selection. (Note that this estimate is produced using a total of 100 \times 100 = 10000 resampled datasets.)

This approach, known as “bootstrap aggregation” or “bagging” (Breiman, 1996), reduces the variance of our estimator.

In addition to computing this point estimate, we can also construct confidence intervals for the true selection probability. To do this, we first note an important
connection between selection probability and \textit{standardized effect size}—the effect size divided by the standard deviation of the outcomes (Cohen, 1988). We denote the true standardized effect size of A versus B by $\delta_{AB}$. The selection probability for B is equal to the probability that our estimate of $\delta_{AB}$ is positive. Therefore, if our estimate of $\delta_{AB}$ is approximately normally distributed with mean $\delta_{AB}$ and standard deviation $1/\sqrt{n}$, which will hold for large enough sample sizes, then we can approximate the selection probability for B with the quantity $\Phi(\sqrt{n} \times \delta_{AB})$. Here, $\Phi$ is the standard normal cumulative distribution function, and $n$ is the data set size. Using this relationship, we can form a confidence interval for selection probability by first constructing a confidence interval for the standardized effect size $\delta_{AB}$, and then applying $\Phi$ to both ends of the interval. Because of this relationship between standardized effect size and selection probability, a confidence interval for the standardized effect size that contains zero corresponds to a confidence interval for selection probability that contains 0.5—they are simply different views of the evidence in the data. In fact, a confidence interval for selection probability contains 0.5 only if the confidence interval for the standardized effect size contains zero. Therefore if we find that the standardized effect size for A versus B is significantly greater than zero at the 95% level, then we can conclude that the selection probability for B is higher than the selection probability for A at the 95% confidence level.

Selection probability is a generalization of the quantity known as the “probability of replication” (Killeen, 2005a), which is the selection probability of the treatment that was found to be best by the original analysis. This approach has been explored in other contexts, and has been praised for its intuitiveness but has also been met with skepticism (Cumming, 2005; Doros & Geier, 2005; Macdonald, 2005; Killeen, 2005b) about its usefulness as a measure of confidence. Much of this discussion has been clouded by the absence of a clearly defined estimand. Our view is that careful estimation of the selection probability as defined here, combined with confidence intervals, can provide a useful
We now give examples showing the point estimate and confidence interval for selection probability, first on a synthetic dataset and then on a dataset from STAR*D. The synthetic dataset allows us to illustrate how our analyses relate to a known underlying model, and the STAR*D analyses illustrate how a substantive hypothesis can be generated using our procedure.

**Analysis of Synthetic Data**

Figure 2 illustrates the use of our procedure using synthetic data. The scatter plot shows the outcomes for all patients, of which there are 100 in treatment group A, and 100 in treatment group B. The regression lines show the predicted outcome for a patient in each group over the domain of the prescriptive variable, based on a synthetic dataset. In the true underlying model for this example, treatment A is best when the prescriptive variable is less than 0.4, and treatment B is best when the prescriptive variable is greater than 0.4. The two treatments are equivalent when it is equal to 0.4. The stacked bar graph shows how our estimates of the selection probability of each of the treatments changes over the range of the prescriptive variable. The height of each coloured bar represents our estimate of the selection probability. One can see that our estimate of the selection probability of A increases as we move left, and decreases as we move right. Conversely, the selection probability for B decreases as we move left, and increases as we move right. These trends are consistent with the underlying model used to generate the data. When the prescriptive variable is equal to 0.4, there is no difference in the effect of A and B. In truth, since both treatments have the same effect at this point, their selection probabilities are equal; therefore the particular dataset generated for this example—just by
chance—makes it appear that A has a higher selection probability than B. Still, this graph illustrates our best estimate of the true selection probabilities given the available data.

For a more refined view of the evidence provided by the data, we can construct confidence intervals for the selection probability of each treatment. Figure 2 also shows error bars corresponding to a 95% confidence interval for the selection probability of treatment A. If we now examine the analysis where the prescriptive variable is equal to 0.4, we see that although our best estimate indicates that treatment A has the a higher selection probability than B, the error bars indicate that it may in fact have a selection probability anywhere between 10% and 99%, implying that the data are not conclusive about which treatment would be selected more often.

*Analysis of STAR*D data*

We examine the group of STAR*D patients entering Level 3 who preferred to augment their current treatment with one of lithium (LIT) or triiodothyronine (THY). All of the patients in this group were given citalopram during Level 1 of the study for up to 12 weeks, and were then given one of sertraline, venlafaxine, bupropion, citalopram plus bupropion, or citalopram plus buspirone for up to an additional 12 weeks during Level 2 of the study. During this period, they did not achieve remission, defined as reporting a QIDS-SR score less than or equal to 5 during a clinic visit. At the beginning of the next level of the study, Level 3, the patients indicated a preference to augment their current treatment with one of LIT or THY. Consenting patients were randomized to each of these two treatments with equal probability, and their QIDS-SR score was recorded at randomization and at weeks 2, 4, 6, 9, and 12 after beginning treatment. The outcome we have chosen to examine is based on how long it takes, in weeks, for a patient to remit following entry into Level 2. Thus, treatments that produce smaller outcomes (i.e. faster remission) are preferable. Our analyses indicate that patients with low incoming QIDS-SR
may be better treated by augmenting with LIT, while those with high incoming QIDS-SR may be better treated by augmenting with THY. Furthermore, according to our best estimate of selection probability, a repeated analysis would choose LIT for patients with low QIDS-SR approximately 66% of the time, and would choose THY for patients with high QIDS-SR approximately 66% of the time.

Why might this be the case? All patients considered here have had a course of citalopram, a selective serotonin reuptake inhibitor, for up to 12 weeks starting at the beginning of the study. Subsequently, all but 26 were being treated with a seratonergic agent during Level 2. LIT is a seratonergic agent, whereas THY differentially increases norepinephrine levels. Therefore, one possible explanation for our findings is, “Patients with low QIDS-SR have partially responded to serotonin-targeted therapy, and improve further under the added seratonergic effects of LIT. Patients with a high QIDS-SR have not responded to serotonin-targeted therapy, and respond better to a drug that acts on a different neurotransmitter—norepinephrine in this case.”

The evidence in the data for this hypothesis is not conclusive, however. Confidence intervals for selection probability indicate that at no value of QIDS-SR can we exclude the possibility that the true selection probability is 0.5. Furthermore, the confidence intervals indicate that a true selection probability of LIT greater than about 77% is very unlikely even for low QIDS-SR, and a true selection probability of THY greater than 77% is very unlikely even for high QIDS-SR. This can be interpreted in terms of the sample size of the experiment relative to the estimated standardized effect sizes: Even if the true standardized effect size were at the extreme end of its confidence interval—the most favourable situation for correctly detecting an effect—this type of analysis would recover the correct sign of the effect size only 77% of the time.

Nonetheless, a researcher may decide that the level of evidence provided by the analysis, combined with a meaningful scientific explanation of why this trend might exist,
is sufficient to warrant further experimental investigation. This future investigation should be designed and powered appropriately using established statistical methods.

**The Role of Exploratory Analysis**

A measure of confidence for exploratory data analyses serves a different purpose than a measure of confidence for primary analyses: In primary analyses, measures of confidence like the confidence interval or the p-value serve to indicate when there is strong evidence in the data for a true underlying effect. These measures have become indispensable tools in the search for scientific truths, and their correct application and interpretation ensures that the scientific community can be confident in the findings reported by its members. Their main purpose is to use data to verify hypotheses. On the other hand, the purpose of exploratory analyses is to use data to generate 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.

Selection probability summarizes the effect size, the variance of the outcomes, and the data set size. Standard analysis tools such as confidence intervals and hypothesis tests are also functions of these three quantities; we have chosen to examine selection probability for exploratory data analysis because its interpretation is intuitive and appealing. It is especially appropriate for exploratory analyses because it expresses the degree of our confidence in an observed effect, as opposed to a hypothesis testing approach which gives a black-or-white, reject-or-fail-to-reject summary of the data. This allows us to clearly present the evidence in the data so that researchers can combine this evidence with their own expertise to formulate hypotheses for future verification. We are optimistic that this use of practical trial data will spur further research on individualization of treatment,
which will in turn have a positive impact on clinical practice and patient wellness.
References


Footnotes

1 We assume lower outcomes are better; for example, one might consider a measure of symptom severity, or time to remission.

2 More precisely, in scenarios where there is a low signal-to-noise ratio.

3 In the regression setting, $n$ will be the number of degrees of freedom.

4 The remaining 26 patients were treated with bupropion alone during Level 2.
Figure Captions

*Figure 1.* Estimating Qualitative Interactions

*Figure 2.* Selection Probabilities - Three Treatment Example

*Figure 3.* STAR*D Analysis of Level 3 Augment Patients
STAR*D Level 3 Analysis of Augment Patients
Outcome: Time to Remission [30], N = 177

Votes

Previous QIDS−SR

+LIT

+THY