Published in final edited form as:

Int J Biostat. ; 8(1): 18. doi:10.1515/1557-4679.1423.

Statistical Issues and Limitations in Personalized Medicine Research with Clinical Trials

Daniel B. Rubin and
Food and Drug Administration

Mark J. van der Laan
University of California - Berkeley

Abstract

We discuss using clinical trial data to construct and evaluate rules that use baseline covariates to assign different treatments to different patients. Given such a candidate personalization rule, we first note that its performance can often be evaluated without actually applying the rule to subjects, and a class of estimators is characterized from a statistical efficiency standpoint. We also point out a recently noted reduction of the rule construction problem to a classification task and extend results in this direction. Together these facts suggest a natural form of cross-validation in which a personalized medicine rule can be constructed from clinical trial data using standard classification tools and then evaluated in a replicated trial. Because replication is often required by the FDA to provide evidence of safety and efficacy before pharmaceutical drugs can be marketed, there are abundant data with which to explore the potential benefits of more tailored therapy. We constructed and evaluated personalized medicine rules using simulations based on two active-controlled randomized clinical trials of antibacterial drugs for the treatment of skin and skin structure infections. Unfortunately we present negative results that did not suggest benefit from personalization. We discuss the implications of this finding and why statistical approaches to personalized medicine problems will often face difficult challenges.

Keywords

personalized medicine; classification; clinical trials

1 Introduction

A great deal of hope has been placed on the promise of personalized medicine, or the ability to tailor therapy to individual patients. One overview of the current state and future prospects of this research is provided by Rugnetta and Kramer (2009), who state that “Americans today remain guinea pigs in a ‘one size fits all’ approach to medicine” and that “a personalized medicine approach may well allow (perhaps in the not too distant future) every individual patient to receive the best in tailor-made, evidence-based pharmacogenomic medicine.” In fact there have already been several apparent successes in this direction, such

as the development of maraviroc for HIV patients with a certain viral mutation or ivacaftor for the treatment of a small subset of cystic fibrosis patients. This hope for personalized medicine has a great deal of face value validity, as it should be obvious that different patients will often respond to therapy in different ways. One definition that has been used for the art of medicine is that it is the process of applying what is known about groups of patients to the individual patient seeking treatment. Personalized medicine attempts to make this art more scientific, and there has recently been a great deal of statistical research regarding the development of individualized treatment rules, for example the work of Polley and van der Laan (2009) or Cai et al. (2011).

We here consider the problems of constructing and evaluating personalization rules. After introducing our notation and assumptions in Section 2 we review in Section 3 how given a candidate rule, one can often form an unbiased estimator of its performance using independent validation data without conducting a trial that actually applies the rule to patients, and we characterize the class of asymptotically unbiased estimators. In Section 4 we extend previous work to show very generally how the problem of constructing a personalization rule can be reduced to a classification problem, for which there is already a large statistical literature. These facts suggest an experimental design in which a personalized medicine rule can be constructed using standard classification methods and then evaluated in a replicated trial. After presenting simulation results in Section 5 we illustrate this approach using data from two clinical trials, after introducing noise to disguise the data, in Section 6. Unfortunately our results do not provide evidence of gains from more tailored therapy. We conclude in Section 7 by discussing some of the difficulties confronting statistical research in personalized medicine, and ideas for how our experimental setup can be used for further exploration.

2 Setup and Notation

We suppose that for each clinical trial subject a vector X of pre-randomization covariates is collected, the subject is then randomly assigned to one of two treatment arms such that $A = 1$ and $A = 0$ represent assignment to treatment or control, and then a binary response Y is measured such that $Y = 1$ represents success and $Y = 0$ represents failure. The observed data for a subject are therefore

$$\{X, A, Y\}.$$

We work in the Neyman (1923) model of causal inference in which each subject has potential outcomes $Y^{(1)}$ and $Y^{(0)}$ corresponding to the response that would have been seen, possibly hypothetically, if the subject had been assigned to treatment or control. Hence, we assume the observed response can be written as

$$Y = AY^{(1)} + (1 - A)Y^{(0)}.$$

As we consider estimating and evaluating personalized medicine rules in randomized trials we also make the randomization assumption that treatment assignments are made independently of baseline covariates or counterfactual outcomes, in that

$$A \perp (X, Y^{(0)}, Y^{(1)}).$$

For a trial with n total subjects assume counterfactual data $\{(X_i, Y_i^{(0)}, Y_i^{(1)})\}_{i=1}^n$ are n independent and identically distributed draws from a population of subjects. However, we suppose the randomization is constrained such that a fixed number of subjects are assigned to the treatment arm, and thus the treatment assignment variables $A_1, \dots, A_n$ are identically distributed but not independent. As notation, the probability of assignment to the treatment arm is

$$\pi = P(A = 1), 0 < \pi < 1.$$

We also define the success rates that would be expected if all subjects in the population were to be assigned to either the treatment arm or the control arm, where the equalities below follow immediately from the randomization assumption:

$$\begin{aligned} \mu^{(1)} &= E[Y^{(1)}] = E[Y|A = 1] = \text{success rate under treatment,} \\ \mu^{(0)} &= E[Y^{(0)}] = E[Y|A = 0] = \text{success rate under the control.} \end{aligned}$$

A personalized medicine rule h is a mapping from a subject's vector X of covariates to one of the treatment arms, such that

$$h: \mathcal{X} \rightarrow \{0, 1\}.$$

That is,

$$\begin{aligned} h(X_i) = 1 &\text{ means rule } h \text{ assigns subject } i \text{ to treatment,} \\ h(X_i) = 0 &\text{ means rule } h \text{ assigns subject } i \text{ to control.} \end{aligned}$$

If the rule were applied to subject i , then the clinical outcome would be $Y_i^{(1)}$ if the rule assigned the subject to the treatment arm and the outcome would be $Y_i^{(0)}$ if the rule recommended the control. Hence, the clinical outcome for subject i under the personalized rule would be

$$h(X_i)Y_i^{(1)} + [1 - h(X_i)]Y_i^{(0)} = \text{response for subject } i \text{ under rule } h.$$

The quality of the personalized medicine rule can be defined through the success rate that would be expected if all subjects in the population had their treatment assignment made through the rule, or

$$\mu(h) = E\left\{h(X)Y^{(1)} + [1 - h(X)]Y^{(0)}\right\} = \text{success rate under rule } h.$$

Consequently, there is promise in the personalized medicine rule if this success rate is substantially larger than the success rates that would be obtained under assignment of all subjects in the population to a fixed treatment arm.

There is a distinction between presenting a personalized medicine rule and presenting subgroup analysis through forest plots when summarizing clinical trial results. Subgroups considered typically are not mutually exclusive, so if a treatment appears beneficial relative to control for women and the elderly and harmful for males and younger subjects it may be unclear what a treatment guideline should recommend for young women or elderly men. The rules we consider are more ambitious than subgroup analysis because they attempt to make a treatment recommendation for each subject using all relevant baseline covariate information.

3 Evaluating Personalized Medicine Rules

Suppose that a proposed personalized medicine rule has been obtained, and we seek to evaluate its performance on a new dataset. These new data could be from a holdout set from the data used to construct the rule, but we considered simulations based on datasets from drug development programs in which two phase III trials were conducted with identical protocols. In many disease areas such replicated evidence is a standard regulatory requirement from the FDA needed to obtain marketing approval. Integrated datasets containing results from the two trials are requested from pharmaceutical industry sponsors and allow a natural form of cross-validation in which a personalized medicine rule can be estimated and evaluated in separate trials.

In the following we therefore consider rule h to be a given function. We begin by defining the following random variable, which will be needed soon to estimate the performance of the personalized medicine rule:

$$D(h; X, A, Y; m) = \frac{A}{\pi}h(X)Y + \frac{\pi - A}{\pi}h(X)m(1, X) \\ + \frac{1 - A}{1 - \pi}[1 - h(X)]Y + \frac{A - \pi}{1 - \pi}[1 - h(X)]m(0, X).$$

This is indexed by a function

$$m: \{0, 1\} \times \mathfrak{X} \rightarrow [0, 1],$$

where $\mathfrak{X}$ denotes the sample space of the vector of baseline covariates. We will call this function a nuisance function because it must be chosen before our rule performance estimator can be applied, and a specific choice is proposed in Section 3.2. After the nuisance function is selected the performance of the personalized medicine rule can be estimated through the empirical mean

$$\mu_n(h; m) = \frac{1}{n} \sum_{i=1}^n D(h; X_i, A_i, Y_i; m). \quad (1)$$

The intuition behind the estimator is easiest to grasp by considering the special case when $m(A, X) = 0$, in which the estimated success rate is simply a weighted average success rate among subjects who by chance were randomized to the arm consistent with the assignment of the personalized medicine rule under consideration. Other choices of nuisance function can increase efficiency through utilizing baseline covariate information, as will be discussed in more detail in the following subsection. Our first result is that the estimator in (1) is unbiased.

Lemma 1

The estimator $\mu_n(h; m)$ is unbiased in that $E[\mu_n(h; m)] = \mu(h)$.

Proof

We first note that because $AY = AY^{(1)}$ and $(1 - A)Y = (1 - A)Y^{(0)}$ that we can rewrite

$$\begin{aligned} D(h; X, A, Y; m) &= \frac{A}{\pi} h(X) Y^{(1)} + \frac{\pi - A}{\mu} h(X) m(1, X) \\ &+ \frac{1 - A}{1 - \pi} [1 - h(X)] Y^{(0)} + \frac{A - \pi}{1 - \pi} [1 - h(X)] m(0, X). \end{aligned}$$

Now noting $A \perp (X, Y^{(1)}, Y^{(0)})$ by the randomization assumption, and recalling $E[A] = \pi$, we find the expected values of the four terms sum to

$$E\{h(X)Y^{(1)}\} + 0 + E\{[1 - h(X)]Y^{(0)}\} + 0 = \mu(h).$$

□

A practical implication is that the performance of a personalized medicine rule can be evaluated without actually conducting a trial in which the rule is applied to subjects. It is sufficient to use data from a standard clinical trial in which subjects are only randomly assigned to fixed treatment arms, provided that the rule has been generated independently of this trial.

3.1 Nuisance Function Choice

The choice of nuisance function determines estimator variance, and we now describe the choice used in our experiments. This subsection is more technical than the rest of the article and some readers may wish to skip ahead, but for reference our nuisance function choice is defined herein in formula (2).

To provide background for our discussion of nuisance functions, note that we can think of the observed data for a subject as

$$\{X, A, (Y, \tilde{Y})\}$$

for

$$\begin{aligned}\tilde{Y} &= \{Ah(X) - (1-A)[1-h(X)]\}Y = A\tilde{Y}^{(1)} + (1-A)\tilde{Y}^{(0)}, \\ \tilde{Y}^{(1)} &= h(X)Y^{(1)}, \tilde{Y}^{(0)} = -[1-h(X)]Y^{(0)}\end{aligned}$$

The observable $\tilde{Y}$ is not of direct clinical interest, but we introduce it to mathematically reframe the problem as estimation of a more typical treatment effect. The randomization assumption implies the independence

$$A \perp (X, \tilde{Y}^{(1)}, \tilde{Y}^{(0)}).$$

Also, the success rate under the personalized medicine rule can be written as

$$\mu(h) = E[\tilde{Y}^{(1)}] - E[\tilde{Y}^{(0)}].$$

With this setup we are back in the familiar problem of estimating a difference in expected responses under treatment and control in a randomized trial, where efficiency theory is well-studied (van der Laan and Robins, 2003; Tsiatis, 2006; Tsiatis et al., 2008). *We sketch several implications, referring to these references for technical details.*

In the language of efficiency theory (Bickel et al., 1993), as the sample size increases and the treatment assignment proportion converges to a fixed value in $(0,1)$ any regular asymptotically linear estimator sequence can be written as

$$\mu_n = \frac{1}{n} \sum_{i=1}^n \left(\frac{A_i}{\pi} - \frac{1-A_i}{1-\pi} \right) \left[\tilde{Y}_i - \eta(X_i) \right] + o_P(n^{-1/2}).$$

Up to the $o_P(n^{-1/2})$ remainder this is an unbiased estimator for any integrable choice of nuisance function $\eta(\cdot)$. Further, the choice minimizing the asymptotic variance of $\sqrt{n}(\mu_n - \mu(h))$ is given by

$$\begin{aligned}\eta^*(X) &= (1-\pi)E[\tilde{Y}|A=1, X] + \pi E[\tilde{Y}|A=0, X] \\ &= (1-\pi)h(X)E[Y|A=1, X] - \pi(1-h(X))E[Y|A=0, X].\end{aligned}$$

The link with our proposed estimator (1) is that it is simple algebra to check

$$\mu_n(h; m) = \frac{1}{n} \sum_{i=1}^n \left(\frac{A_i}{\pi} - \frac{1-A_i}{1-\pi} \right) \left[\tilde{Y}_i - \eta(X_i) \right]$$

for

$$\eta(X) = (1 - \pi)h(X)m(1, X) - \pi[1 - h(X)]m(0, X)$$

Our estimator is thus asymptotically efficient, meaning it achieves the smallest asymptotic variance among regular asymptotically linear estimators, if

$$m(A, X) = m^*(A, X) \equiv E[Y|A, X].$$

Hence, we have characterized the class of possible estimators for this problem. Different nuisance function choices will lead to different variances, with the optimal choice being the regression function of the response on treatment assignment and baseline covariates, which of course is unknown.

Aggressiveness in fitting this regression relates to willingness to leverage covariate information (Tsiatis et al., 2008; Moore and van der Laan, 2009). For instance, Rubin and van der Laan (2008) discuss empirical efficiency maximization, or fitting a working regression model to optimize asymptotic variance for the resulting estimator of the parameter of interest, which can gain efficiency relative to maximum likelihood fitting when the working model is misspecified. However, to reduce machinery we use the (data-determined) nuisance function

$$m_n(A, X) = A\mu_n^{(1)} + (1 - A)\mu_n^{(0)}, \quad (2)$$

where

$$\begin{aligned} \mu_n^{(1)} &= \frac{1}{n} \sum_{i=1}^n \frac{A_i}{\pi} Y_i, \\ \mu_n^{(0)} &= \frac{1}{n} \sum_{i=1}^n \frac{1 - A_i}{1 - \pi} Y_i \end{aligned}$$

are empirical response rates under treatment and control, corresponding to fitting optimal $m^*(A, X)$ with separate constants in the two arms. In Section 4 it will be convenient that this choice has no dependence on rule h .

Unfortunately, with a data-determined nuisance function we may no longer have exact finite sample unbiasedness as in Lemma 1. However, this issue disappears with sample size: if $m(A, X) = A\mu^{(1)} + (1 - A)\mu^{(0)}$ then $\sqrt{n}[\mu_n(h; m_n) - \mu(h)]$ and $\sqrt{n}[\mu_n(h; m) - \mu(h)]$ will converge in law to normal distributions with mean zero and equal variances, as it is simple to verify that $|\mu_n(h; m_n) - \mu_n(h; m)| = O_P(n^{-1})$. In fact if a collection $\mathcal{H}$ of personalized rules is considered as in Section 4 then convergence is uniform in that $\sup_{h \in \mathcal{H}} |\mu_n(h; m_n) - \mu_n(h; m)|$ will at worst be $O_P(n^{-1/2})$. Thus, complications from a data-determined nuisance function will not be further discussed.

3.2 Confidence Intervals for Success Rates Differences

Interest may lie in the difference in expected success rates under the personalized medicine rule and one of the two treatment arms, defined through

$$\begin{aligned}d^{(1)}(h) &= \mu(h) - \mu^{(1)}, \\d^{(0)}(h) &= \mu(h) - \mu^{(0)}.\end{aligned}$$

A general class of unbiased estimators for the first difference is given by

$$d_n^{(1)}(h; m) = \mu_n(h; m) - \frac{1}{n} \sum_{i=1}^n \left\{ \frac{A_i}{\pi} [Y_i - m(1, X_i)] + m(1, X_i) \right\}.$$

It is also a simple combinatorial exercise to compute the variance as

$$\begin{aligned}\text{Var}(d_n^{(1)}(h; m)) &= \frac{\pi}{n} \text{Var} \left\{ \left[\frac{Y - (1 - \pi)m(1, X)}{\pi} - m(0, X) \right] [1 - h(X)] \middle| A = 1 \right\} \\&+ \frac{1 - \pi}{n} \text{Var} \left\{ \left[m(1, X) - \frac{Y - \pi m(0, X)}{1 - \pi} \right] [1 - h(X)] \middle| A = 0 \right\},\end{aligned}$$

leading to a standard error estimate $\sigma_n^{(1)}(h; m)$ when the two conditional variances on the right side are estimated with the sample variances of

$$\begin{aligned}&\left\{ \frac{Y_i - (1 - \pi)m(1, X_i)}{\pi} - m(0, X_i) [1 - h(X_i)] : A_i = 1 \right\}, \\&\left\{ m(1, X_i) - \frac{Y_i - \pi m(0, X_i)}{1 - \pi} [1 - h(X_i)] : A_i = 0 \right\}.\end{aligned}$$

With the normal approximation one can thus evaluate the performance by forming a 95% confidence interval

$$d_n^{(1)}(h; m) \pm 1.96 \sigma_n^{(1)}(h; m).$$

For reasons we have previously discussed, the confidence interval will be asymptotically valid when using our data-determined nuisance function choice in (2). By symmetry the same formulas of course hold for $d^{(0)}(h)$, or the difference in success rates between the personalized rule and the control arm, after reversing treatment assignment labels.

4 Constructing Personalized Medicine Rules

We now consider construction of a personalized medicine rule using the results of a clinical trial, and extend a recently shown reduction of this problem to binary classification. The benefit of this reduction is that there is already a large literature on classification and many algorithms available in standard software, which can thus be directed toward finding personalized therapies.

Regarding notation, we now consider $\{(X, A, Y)\}_{i=1}^n$ and π to be the observable data and treatment assignment rate in the trial used for rule construction.

Ideally the expected success rate under a personalization rule h should be as large as possible, meaning we would like to maximize

$$h \rightarrow \mu(h).$$

However, this cannot be done directly because the success rate under any given rule will not be known. Instead, a general strategy known in different contexts as empirical risk minimization or minimum contrast estimation is to solve such optimization problems by replacing unknown population expectations by unbiased empirical counterparts. The idea is thus that we could attempt to optimize

$$h \rightarrow \mu_n(h; m)$$

over a candidate class $\mathcal{H}_n$, which could grow with sample size to determine the tradeoff between underfitting and overfitting. In fact even this optimization problem will turn out to be too computationally difficult, but the discussion of empirical risk minimization will be very helpful for framing our reduction to classification. As mentioned in the previous section the nuisance function choice would enter the problem by impacting the stochastic distance between the population success rate $\mu(h)$ and the unbiased empirically estimated success rate $\mu_n(h; m)$ of a personalization rule.

In preparation for our proposed rule construction method define the binary-valued variable

$$Z = Z(A, Y) = AY + (1 - A)(1 - Y) \quad (3)$$

and also define

$$\begin{aligned} K = K(X, A, Y; m) &= (1 - A)(1 - Y) \left[m(1, X) + \frac{\pi}{1 - \pi} m(0, X) \right] \\ &+ A(1 - Y) \left[\frac{1 - \pi}{\pi} m(1, X) + m(0, X) \right] \\ &+ (1 - A)Y \left[\frac{1}{1 - \pi} - m(1, X) - \frac{\pi}{1 - \pi} m(0, X) \right] \\ &+ AY \left[\frac{1}{\pi} - \frac{1 - \pi}{\pi} m(1, X) - m(0, X) \right]. \end{aligned} \quad (4)$$

The following lemma demonstrates how our problem reduces to weighted classification. A similar result has been independently obtained by Kosorok (2011) for the special case when $m(A, X) = 0$ is used as the nuisance function, in work to be expanded upon in Zhao et al. (in press). With a binary outcome and 1:1 randomization this would reduce to examining subjects with a successful clinical response and attempting to classify which treatment they received. As mentioned, our experiments instead used the nuisance function in (2).

Alternatively, using $m(A, X) = 1$ would have meant restricting to nonresponders and trying to classify which treatment they did not receive.

Lemma 2

Consider a given nuisance function $m(\cdot, \cdot)$ and define the weighted misclassification loss

$$L(h; X, A, Y; m) = K(X, A, Y; m) \cdot I[h(X) \neq Z(A, Y)],$$

for $Z \in \{0, 1\}$ and K defined in (3) and (4), where $K = 0$ if the nuisance function satisfies $0 \leq m(A, X) \leq 1$. Also define the random variable

$$\begin{aligned} C(X, A, Y; m) &= (1 - A)(1 - Y)m(1, X) + A(1 - Y)m(0, X) \\ &+ (1 - A)Y \left[\frac{1}{1 - \pi} - \frac{\pi}{1 - \pi} m(0, X) \right] + AY \left[\frac{1}{\pi} - \frac{1 - \pi}{\pi} m(1, X) \right], \end{aligned}$$

which has no dependence on any rule h . Then the personalized medicine rule success rate $\mu(h)$ is decreasing in misclassification risk, and the estimated success rate $\mu_n(h; m)$ is decreasing in empirical misclassification risk, in that

$$\begin{aligned} \mu(h) &= E[C(X, A, Y; m)] - E[L(h; X, A, Y; m)], \\ \mu_n(h; m) &= \frac{1}{n} \sum_{i=1}^n C(X_i, A_i, Y_i; m) - \frac{1}{n} \sum_{i=1}^n L(h; X_i, A_i, Y_i; m). \end{aligned}$$

Proof

We first recall from the definition (1) that the unbiased estimator of the success rate of rule h is $\mu_n(h; m) = n^{-1} \sum_{i=1}^n D(h; X_i, A_i, Y_i; m)$. By checking the eight possible values of $D(h; X, A, Y) \in \{0, 1\}^3$ one can verify

$$D(h; X, A, Y; m) = C(X, A, Y; m) - L(h; X, A, Y; m),$$

which immediately implies the lemma.

The lemma clearly suggests forming a personalized rule as follows. After selecting nuisance function $m(\cdot, \cdot)$, form observation weights $\{K_i\}_{i=1}^n$ as in (4) and consider the weighted classification problem in which the predictors are baseline covariate vectors $\{X_i\}_{i=1}^n$ and class labels are $\{Z_i\}_{i=1}^n$ as in (3). One can then feed data into standard classification software.

In contrast, with a non-binary clinical outcome Zhang et al. (2012) considered a related problem of optimizing the empirically estimated success rate of a personalized medicine

rule, but needed to apply genetic algorithms and grid searches, which are less well-developed.

There is a large body of work in the statistical literature on classification. Following Fisher's linear discriminant analysis in the 1930s a variety of well-known methods have been proposed (Hastie et al., 2009) such as naive Bayes methods, logistic regression classifiers, nearest neighbors, decision trees, neural networks, and ensemble techniques such as bagging, boosting, and random forests.

There is also much theory relating the misclassification risk of particular classifiers to the optimal risk achievable (i.e., the Bayes risk) and showing consistency and rates of convergence (Devroye et al., 1996). One major type of method in our setting would attempt to estimate conditional probability function $P(Z = 1|X)$ and classify accordingly. This could also include generative techniques that model $P(X|Z)$ and apply Bayes rule. It can often be precisely quantified how estimation of the conditional probability function translates into misclassification risk (Audibert and Tsybakov, 2007). A second general type of method, exemplified by boosting and support vector machines, would relate to minimization of $f \rightarrow \sum_{i=1}^n K_i \phi([2Z_i - 1]f(X_i))$ over a candidate set $\mathcal{F}_n$ of real valued functions, for a convex "surrogate loss" function ϕ , where $f \in \mathcal{F}_n$ corresponds to the classification rule $h(x) = I(f(x) > 0)$.

This can be a computational shortcut for the intractable problem of directly minimizing empirical misclassification risk over a large candidate set, but the classification performance can still often be quantified (Bartlett et al., 2006).

In our simulations and clinical trial experiments we constructed personalized medicine rules with bagged decision trees (Breiman, 1996), which are considered to be competitive with other well-known classification techniques. Tree construction used the rpart package in the R language (Therneau and Atkinson, 2012) with default settings and the bagging ensemble combined 100 trees. Classification trees were convenient base classifiers because they were able to naturally handle the messy aspects of our datasets without manual preprocessing or data manipulation, such as missing covariate values, collinearity, and a mixture of continuous and categorical predictors.

5 Initial Simulations

We first conducted simulation studies to examine personalized rules. The data generating distribution was as follows. Each element of a subject's baseline covariate vector was independently drawn from a standard normal distribution. Only the first five covariates were relevant to the outcome, and remaining covariates were noise variables. Letting W denote the sum of the first five covariates, success rates were $1/[1 + \exp(-W/3)]$ and $1/[1 + \exp(W/3)]$ under the treatment and control arms.

This was a logistic model for the response rate as a function of treatment and covariates in which there were treatment-covariate interactions determining which arm was superior. The optimal personalized medicine rule assigning therapy based on covariate values would attain an expected success rate of approximately 64%, a large improvement over the 50% success rate achieved by either the treatment group or the control group.

We simulated clinical development programs in which a rule was first constructed in a training trial through the bagged classification tree approach outlined in the previous section and was then evaluated in a separate validation trial. In our simulations both trials used 1:1 randomization. The validation trial was made large with 20,000 total subjects, and in each simulated development program we computed the success rate under assignment to the rule formed in the training trial. As this was a simulation both counterfactual outcomes were available for every subject, so the response under the rule was known for all validation subjects. With 100 Monte Carlo simulations of training and validation clinical development programs we computed the expected validation trial success rate of a personalized medicine rule.

We began with a training trial of 600 total subjects and a 39-dimensional covariate vector, matching the settings of our data analysis in the next section. The expected validation trial success rate was $59.3\% \pm 0.1\%$, where the standard error represented negligible simulation error. Thus, with strong treatment-covariate interactions one can gain by estimating personalized rules.

The results differed when the configuration was changed. We next considered a training trial with 100 total subjects and 100 baseline covariates. Such classification problems where number of observations n does not exceed the dimensionality p and there are few useful predictors have generated significant interest and new methodology in the statistical community over the past decade. The expected success rate from estimating a personalized rule was $52.8\% \pm 0.2\%$, so there was minimal gain over the expected 50% success rate achieved by both the treatment and control arms.

The message from the simulation studies was thus that the benefit from estimating a personalized medicine rule can be sensitive to the data generating distribution, and that one should not count on large gains with small sample sizes or too many irrelevant predictors.

6 Simulations from Clinical Trial Datasets

We illustrate the personalized medicine rules using data from two randomized phase III clinical trials, after adding noise to disguise the data. A form of cross-validation was used in which one trial was first designated as the training set used to construct a personalized medicine rule with the classification approach described in Section 4. The other was used as a validation set, and the performance of the constructed personalized medicine rule was evaluated as described in Section 3. We then repeated this entire process after switching the roles of the training and validation trials.

The two trials had identical protocols, and examined two active antibacterial drugs for the treatment of skin and skin structure infections. Each study was conducted in a double-blind manner, used 1:1 randomization, and was conducted at multiple sites in multiple countries. In both trials a primary analysis population consisted of all randomized subjects who received any dose of study drug, and this analysis population contained between 300 and 400 subjects in each arm of each trial. The primary efficacy endpoint was a binary success/

fail outcome variable corresponding to whether signs and symptoms of infection had sufficiently resolved at a test-of-cure visit in the opinion of a clinical investigator.

The development program was anonymized by calling the two treatment arms Drug A and Drug B and computing results on a bootstrap resample in which 300 subjects in each arm of each trial were sampled with replacement from the original patient set. The results presented below were qualitatively similar to results with the original datasets. Results were also fairly consistent when repeating our noise-corrupting procedure different times.

Baseline covariates used for constructing the rule were a mixture of demographic and clinical variables. As classification performance was our criterion for rule construction we did not place emphasis on avoiding collinearity or restricting to a small handful of covariates (Breiman, 2001). In total we included 39 variables that were a mix of continuous and categorical predictors.

Among these patient characteristics were demographic and background factors such as age, sex, race, ethnicity, country, region, height, weight, body mass index, creatinine clearance, and diabetes status.

Also included were skin clinical parameters such as a severity score, abscess presence and dimension, fever, hypothermia, body temperature, white blood cell counts, type of infection, location of infection, prior treatment failure status, prior antibiotics use, and lesion size (length, width, and area).

Finally, we included clinical signs and symptoms such as measurements of bullae, the depth of involvement of the lesion, discharge, erythema, fluctuance, lymphangitic spread, necrosis, sanguineous status, and swelling, warmth, and tenderness of the infection site.

Baseline variables not included were measurements of pathogens isolated and their susceptibility to different study drugs. While such measurements can be pre-randomization variables for the purposes of subgroup inference they are usually not available at the time when a treatment decision must be made, as it can take several days for a culture taken at baseline to grow. Consequently, although microbiological variables may in principle be useful for personalization if better rapid diagnostics could be developed and there are many instances where organisms are susceptible to one study drug and resistant to the other, we did not consider this information available for personalized medicine rules.

Table 1 below presents our results, which were disappointing. No matter which trial was used for rule construction and which for rule validation our personalized medicine rule performed worse than Drug A, meaning there was no estimated benefit over just assigning all subjects to this arm. The 95% confidence intervals for differences in success rates between personalized medicine rules and fixed treatments seemed to rule out the possibility of large gains from the rules we constructed. In addition, in Section 4 we discussed how the rule construction problem could be reduced to a weighted classification problem, and our personalized rules had empirical weighted misclassification rates of 54% and 50% in the first and second trials, which were worse than expected from chance guessing. To

summarize, in this clinical trial program we were unable to form a beneficial personalized medicine rule.

7 Discussion of Results

While we view our results as having several limited negative implications that we discuss below, there are many possible objections that could be made if one were to draw the conclusion from our experiments that there is no potential in utilizing clinical trial data to construct and evaluate personalized medicine rules.

It is possible that personalized medicine rules might have had better performance with different baseline covariates. Our answer is that we used the major demographic and clinical variables that were readily available for analysis in the integrated datasets to which we had access.

Another point we have so far overlooked is that our process of constructing and evaluating personalized medicine rules on separate trials could be misleading if there is major heterogeneity between the trials. In our experiment the two multicenter trials were conducted at different sites but had identical protocols and there was no noticeable heterogeneity in patient characteristics or outcomes. If unexplained trial-level heterogeneity contributed to the poor performance then that is a testament to the fragility of the approach.

We showed in Section 4 that the performance of a personalized rule depends on a certain misclassification risk, and it is possible that our classifier was suboptimal. We are uncertain whether the misclassification risk of our method could have been greatly improved with an alternative approach or ensemble of different approaches (Polley and van der Laan, 2010), but we note that bagged decision trees have been fairly competitive with state of the art methods in a variety of benchmark classification problems, although they can perform worse than boosting methods when there is little measurement error in class labels (Dietterich, 2000).

Performance can also be related to the choice of nuisance function $m(\cdot, \cdot)$ needed to define the classification problem. However, preliminary exploration of our clinical trial datasets with nuisance function choices other than our method in (2) has so far given similar results.

Our results were of course limited to one clinical setting, and performance may have been stronger if we had considered other disease areas. Indeed, we noted in our introduction that there have already been several major success stories in the personalized medicine field.

Nevertheless, given the hope invested in personalized medicine research with clinical trials our results may provide a better perspective on the difficulty of this problem.

One reason for the challenge relates to sample size. Clinical trials are typically not powered to examine subgroup effects or interaction effects, which are closely related to personalization. Our simulations in Section 5 illustrated that even if an optimal personalized medicine rule can provide substantial gains it may be difficult to estimate this rule with few subjects.

The methods we have discussed for personalized medicine rule construction are also a simplification of how treatment guidelines are made in practice. We considered constructing a rule from a single trial. Also, we only dealt with efficacy and did not consider adverse events or subpopulation-specific risk-benefit profiles of different therapies.

For instance, consider a placebo-controlled trial in which a test drug does not negatively impact efficacy. In our formulation the personalized rule optimizing the success rate would be to trivially assign all subjects to the treatment even if only a subgroup could metabolize the drug and obtain benefit, while the complementary subgroup would obtain no benefit but would be exposed to drug toxicities. This setting deserves further investigation.

There is also great interest in using active-controlled clinical trials for personalized medicine or comparative effectiveness research, and a further difficulty is that these often use noninferiority designs rather than superiority designs. In noninferiority trials the objective is to determine if a treatment is unacceptably worse than a comparator. In a typical placebo-controlled superiority trial the design attempts to minimize certain biases that tend to favor the null hypothesis of no treatment effect, but in a noninferiority trial these same biases can favor the alternative hypothesis. For instance, if a trial includes subjects with mild self-resolving disease, allows high rates effective concomitant therapy, and outcome assessment suffers from measurement error then it will be hard to detect differences between treatments or form beneficial personalization rules.

Going forward, one contribution of this article that we hope to see replicated by others is our experimental design for evaluating rule performance. Ideally a personalized medicine rule would be formed in advance of a clinical trial designed to evaluate it, but for exploratory purposes it may be useful to examine results in trials that have already been conducted. In particular, the abundance of replicated trials often required in drug development programs offer an opportunity to separate rule construction from rule validation.

While the results in this article suggest that this personalized medicine problem can be formidable, investigators who can meet such a statistical challenge on a validation set and demonstrate the added value of personalization will go a long way toward justifying the hope that has been placed on targeted therapy.

References

- Audibert JY, Tsybakov AB. Fast learning rates for plug-in classifiers. *The Annals of Statistics*. 2007; 35:608–633.
- Bartlett PL, Jordan MI, McAuliffe JD. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*. 2006; 101(473):138–156.
- Bickel, PJ., Klassen, CAJ., Ritov, Y., Wellner, JA. *Efficient and Adaptive Estimation for Semiparametric Models* Springer; New York: 1993
- Breiman L. Bagging predictors. *Machine Learning*. 1996; 24:123–140.
- Breiman L. Statistical modeling: the two cultures. *Statistical Science*. 2001; 16:199–215.
- Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*. 2011; 12:270–282. [PubMed: 20876663]
- Devroye, L., Györfi, L., Lugosi, G. *A Probabilistic Theory of Pattern Recognition* New York: Springer-Verlag; 1996

- Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*. 2000; 40:139–158.
- Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* Vol. 2009. Springer; New York: 2009 Springer series in statistics
- Kosorok, MR. Clinical trials and personalized medicine 2011 Presentation slides available at www.stat.sc.edu/~hitchcock/srcos2011. Accessed March 8, 2012
- Moore KM, van der Laan MJ. Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine*. 2009; 28:39–64. [PubMed: 18985634]
- Neyman, J. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (In Polish). In: Dabrowska, DM., Speed, TP., translators *Statistical Science* Vol. 5. 1923465-4801990
- Polley, EC., van der Laan, MJ. UC Berkeley Division of Biostatistics Working Paper Series 2009 Selecting optimal treatments based on predictive factors. Working Paper 244
- Polley, EC., van der Laan, MJ. UC Berkeley Division of Biostatistics Working Paper Series 2009 Super learner in prediction. Working Paper 266
- Rubin DB, van der Laan MJ. Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *International Journal of Biostatistics*. 2008 Volume 4, Article 5.
- Rugnetta M, Kramer W. Paving the way for personalized medicine. *Science Progress*. 2009 Sep. 14:2009.
- Therneau, TM., Atkinson, B. rpart: Recursive partitioning 2012 R package version 3.1-52. Available at <http://cran.r-project.org/web/packages/rpart/>
- Tsiatis, AA. *Semiparametric Theory and Missing Data* Springer; New York: 2006
- Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine*. 2008; 27:4658–4677. [PubMed: 17960577]
- van der Laan, MJ., Robins, JM. *Unified Methods for Censored Longitudinal Data and Causality* Springer-Verlag; New York: 2003
- Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. *Biometrics*. 2012 In press.
- Zhao, Y., Zeng, D., Rush, AJ., Kosorok, MR. *Journal of the American Statistical Association* 2012 Estimating individualized treatment rules using outcome weighted learning. In press

Table 1

Estimated Success Rates for Personalized Medicine Rules (PMRs)

Trial	Drug A	Drug B	PMR	PMR - Drug A
#1	261/300 (87.0)	256/300 (85.3)	84.3	-2.7 (-7.0, 1.5)
#2	259/300 (86.3)	247/300 (82.3)	84.4	-1.9 (-5.7, 1.8)