Published in final edited form as:

Stat Med. 2014 July 30; 33(17): 2881–2896. doi:10.1002/sim.6138.

Analyzing Longitudinal Data to Characterize the Accuracy of Markers Used to Select Treatment

Colleen M. Sitlani^{a,*} and Patrick J. Heagerty^b

^aDepartment of Medicine, University of Washington, Seattle, WA

^bDepartment of Biostatistics, University of Washington, Seattle, WA

Abstract

With the increasing availability of detailed clinical information there is optimism that treatment choices can be selectively directed to those individuals most likely to benefit. While standard clinical trials can establish whether a treatment appears to be effective on average, subsequent work is needed to determine whether there are identifiable subgroups of subjects for whom treatment is either particularly beneficial or harmful. Molecular assays and modern imaging technology now allow numerous candidate measures to be used as potential determinants of treatment choice. In this manuscript we focus on novel measures of decision accuracy that reflect the treatment marker objective. Specifically, we define longitudinal individual-level potential outcomes (principal strata) that characterize patient outcomes under treated and untreated states. We propose generalizations of sensitivity and specificity that measure the accuracy with which a marker can distinguish those subjects who are expected to have a more favorable outcome under a specific treatment choice from those subjects who are expected to have a more favorable outcome under alternative treatment options. For quantitative markers we propose principal receiver operating characteristic curves that display the full range of potential sensitivity and specificity. We use simulations to demonstrate the properties of proposed estimators, and we illustrate the methods using candidate neuroimaging and electro-diagnostic markers that could be used to select patients for carpal tunnel surgery.

Keywords

Treatment selection; Longitudinal data analysis; Causal inference; Classification and prediction

1. Introduction

Individuals with a given medical condition are often faced with two or more options for treatment. Clinical trials that compare available treatments typically focus inference on the average effect of treatment for a well-defined set of eligible patients. An important secondary objective is the identification of subgroups, or individuals who benefit more from one treatment as compared to other options. That is, are there demographic, imaging,

genomic, or other clinical measures that are useful in determining which treatment is best for a particular person? If markers are associated with treatment response then how accurate are they at identifying those individuals who are expected to benefit? For example, in subjects with Carpal Tunnel Syndrome (CTS), common therapeutic options are to treat with physical therapy or to pursue a more aggressive surgical approach. The conservative approach of physical therapy may be a less costly and less invasive treatment when compared to surgery. On the other hand, surgery may bring permanent relief and may do so in less time than the typical course of physical therapy treatments. Randomized controlled trials have shown that people with CTS who receive surgical treatment have better outcomes, on average, than those who receive more conservative nonsurgical treatments [1, 2, 3, 4]. However, overall treatment benefit is small, so there is interest in identifying a subset of people for whom surgery is most beneficial, potentially through magnetic resonance neurographic imaging (MRNI) and electrodiagnostic studies (EDS) which can provide an evaluation of the underlying pathology.

In this manuscript, we propose new methods for evaluating the usefulness of candidate markers that may guide the selection of treatment. We call a marker that is used to select treatment a *response-predictive* marker to distinguish it from diagnostic and prognostic markers.

Methods for determining the clinical utility of a potential response-predictive marker are not well established. Common approaches include assessing the prognostic value of the marker among treated subjects, evaluating treatment effects in subjects with a restricted range of marker values, and testing for a statistical interaction between treatment and marker [5]. Interactions have been used in analyses of randomized [4] and observational [6] CTS data. Strong interactions, while necessary, are not sufficient for identifying markers that can be used clinically to predict response to treatment [7, 5]. Gunter et al. [7] discuss two main reasons for this: 1) a significant interaction may be non-qualitative [8] in that it predicts magnitude of response but does not affect the optimal treatment decision, and 2) depending on the distribution of the marker, a significant interaction may identify only a small subset of subjects whose optimal treatment decision is influenced by the marker.

Alternate methods typically focus on population-level outcomes, i.e. they evaluate whether, on average, subjects with a particular marker value will have better outcomes with treatment versus without [7, 5]. Here, we address the marker's performance at the individual level, where we quantify the ability of a marker to classify individuals into those who benefit from treatment and those who do not. Our focus is on the decision accuracy properties of a candidate treatment marker. Subjects with a given marker value may benefit from treatment on average, but whether or not the marker accurately classifies subjects with respect to individual benefit is also important in determining whether the marker is useful for guiding clinical practice and selecting treatment. To estimate classification accuracy at the individual level, we must estimate the distribution of individual-level treatment effects, which we define to be subject-specific differences in potential outcomes that would be achieved with and without treatment. Such effects, which can be viewed as subject-specific principal strata [9], are unobservable, and corresponding parameters are not typically identifiable. However,

with longitudinal data *expected* subject-specific treatment effects can be defined, and provide an alternative to time-specific strata based on individual potential outcomes.

Crossover data allow us to approximate a time-specific observation-level treatment response, by differencing a treated outcome at one time and a non-treated outcome at another time. However, relative to counterfactual outcomes, measurement error exists due to substitution of an observed outcome at one time for a potential outcome at another time. Furthermore, comparing potential outcomes at a single follow-up time is a time-specific contrast when frequently an expected or time-averaged summary may be of more clinical relevance. With longitudinal data, both of these key limitations can be addressed: first, bounds for a time-specific estimate can be constructed by modeling subject heterogeneity using random effects with a longitudinal mixed model, and by specifying the correlation between within-subject potential measurement errors; and second, by defining a time-averaged subject-level treatment response. By adopting a time-averaged subject-level treatment response as the individual summary of interest, the distribution of treatment response is based on random effects variance components which are identified through longitudinal crossover outcomes. We present both options in this manuscript, but in practice when it is appropriate to consider multiple possible treatment times we recommend using the subject-level response contrast that averages over time, as the observation-level (potential) response contrast still requires unverifiable assumptions regarding within-subject potential measurement errors.

In Section 2, we propose methods that combine the distribution of individual-level treatment effects captured through random effects with the marginal marker distribution in order to characterize the “principal” true positive fraction (p-TPF) and the “principal” false positive fraction (p-FPF), which together define a “principal” receiver operating characteristic (p-ROC) curve for the subject-specific treatment response of interest. Such p-ROC curves can be used in the context of time-specific observation-level or time averaged subject-level responses. In Section 3, we define estimators and corresponding asymptotic confidence intervals for the p-ROC curves proposed in Section 2. We evaluate the statistical properties of the p-ROC curve estimates and their confidence intervals via simulations in Section 4. In Section 5 we evaluate potentially response-predictive imaging markers in patients with CTS by applying our methods to data collected in a multi-site randomized trial [10, 4]. In Section 6 we summarize our findings and suggest relevant next steps in clinical and statistical contexts.

2. Models

Our motivating application is surgery where there is a single point treatment. All observations post-surgery are considered as treated, although the magnitude of treatment effects may change over time after surgery. Therefore, we develop detailed methods for time-dependent treatment associated with this specific simple scenario, although generalization to more complex treatment paths is possible. We consider a group of subjects, indexed by $i = 1, \dots, n$, possibly observed at multiple follow-up times $t = 1, 2, \dots, T$. Subjects are assigned to a fixed treatment regimen $\mathbf{X} = [X_1, \dots, X_T]$, where X_t is a binary indicator of treatment received prior to outcome measurement at time t . In the context of

surgical treatments, we assume that treatment can only occur at a single time, denoted s . Subjects are designated as untreated at all times prior to surgery and as treated at all times after surgery, i.e. $X_t = 1_{[t, s]}$. Marker values M are observed at baseline, and the candidate markers of interest are assumed to be pre-specified. We also assume that markers take continuous values, with one possibility that M is a derived composite marker score based on multiple (possibly discrete) candidate markers.

Quantitative observed longitudinal outcomes are denoted as $\mathbf{Y}_i = [Y_{i1}, \dots, Y_{iT}]$. Potential (or counterfactual) outcomes are denoted for each possible surgical time, s , as $\mathbf{Y}_i(s) = [Y_{i1}(s), \dots, Y_{iT}(s)]$. Although we write potential outcomes as functions of the surgery time, we naturally assume that the potential outcome at time t only depends on treatment through time t .

Frangakis and Rubin [9] describe one method for characterizing individual-level causal effects through principal stratification. With longitudinal data there are multiple measures per subject and one simple generalization of principal strata focuses on time-specific or observation-level potential outcomes. In this approach multiple measurement times are addressed by focusing on select times such as patient status one year or two years after enrollment. Strata are defined by the potential outcomes at a given time t , under each treatment path, e.g. by $[Y_{it}(s), Y_{it}(t+)]$, where either surgery occurred at time s or surgery has not occurred by time t (and may not occur at all). Here we denote $Y_{it}(t+)$ as the potential outcome for subject i if through time t they have not received surgery. Note that although we are considering evaluation of outcomes at a fixed follow-up time t we need to consider a second time scale that records when, or whether, surgery has occurred. A “treated” outcome, $Y_{it}(s)$ needs to consider the surgical time, s , and a “non-treated” outcome is where surgery has not yet happened, $Y_{it}(t+)$, and which is assumed to be a common value regardless of when (or if) surgery happens after time t . We define time-specific causal effects as differences in potential outcomes within principal strata; we define the time-specific observation-level treatment benefit as $\Delta_{it}(s) = Y_{it}(s) - Y_{it}(t+)$.

However, in many applications such as surgery the key time scale is the time since surgery, $t - s$, and the specific time of measurement, t , and/or surgery, s , may not be of specific interest since surgery may be an elective procedure that can be considered at multiple candidate times such as with carpal tunnel surgery. In such situations it is plausible to consider a patient-specific expected benefit of treatment where patient-specific averages over candidate treatment times are considered. Such a perspective also naturally parallels the concept of a random treatment effect defined as the difference between an expected outcome under treatment as compared to an expected outcome under no treatment for a specific individual. Therefore, with longitudinal data rather than focus on a single outcome time and surgical time pair (t, s) , we propose classification of subjects based on the time averaged subject-level treatment benefit, defined as:

$$\bar{\Delta}_i(\mathcal{S}) = \int_{\mathcal{S}} \Delta_{it}(s) \cdot w(s, t) ds \, dt \text{ e.g. } = \frac{1}{|\mathcal{S}|} \sum_{(t,s) \in \mathcal{S}} \Delta_{it}(s)$$

for a specific set of outcome and treatment times defined by the set $\mathcal{S}$ with cardinality $|\mathcal{S}|$, and a chosen weight function $w(s, t)$ where $\int w(s, t) ds dt = 1$. For example, $\mathcal{S}_1 = \{ (t, s) : t - s = 1, s \in \Omega \}$ represents all times t that are one year past surgery, $s = (t - 1)$, for s in a set of candidate treatment times, Ω , and therefore considers the expected surgical benefit one year post-surgery, but where the exact time of surgery is not specified. Taking sets $\mathcal{S}$ either as discrete sets of time, or more generally assuming a distribution over time, leads to consideration of expectations that retain a subject-specific interpretation and average over within-subject variation.

A second motivation for considering $\bar{\Delta}_i(\mathcal{S})$ comes from the fact that longitudinal data with individuals measured pre- and post-treatment permit consideration of random treatment effects. A random treatment effect represents an expected benefit, and medically it may be of interest to identify those subjects who have an expected benefit from treatment rather than isolate classification of benefit to a single follow-up time. In section 2.2 we detail structural longitudinal models that provide random effects that are interpreted as expected subject-specific treatment contrasts.

In scenarios when subjects have a subject-specific immediate and sustained effect of surgery that is achieved independent of the exact surgery time, the expected benefit of treatment for subject i reduces to a constant with respect to time, e.g. $\bar{\Delta}_i(\mathcal{S}) = \Delta_i^0$, regardless of the time since surgery. However, in other situations the impact of surgery may either change as a function of the timing of surgery, or as a function of the time since surgery. In these cases the subject-specific expected benefit will depend on the time frame that is considered. In section 2.2 we provide a model framework and give illustration of alternative patient specific treatment benefit summaries.

2.1. Classification Accuracy of Marker for Individual Treatment Effects

Our goal is to evaluate how well a given marker M can classify subjects with respect to the observation-level treatment effect $A_{it}(s)$ or the expected subject-level treatment effect $\bar{\Delta}_i(\mathcal{S})$. We accomplish this goal by using a p-ROC curve to assess the discriminatory ability of the marker to separate cases from controls, where in our context a case is a person who would do “better” on the experimental treatment and a control is a person who would do “better” on the standard treatment. Assuming that higher outcomes are better, and using the observation-level treatment effect, we define the quantity $A_{it}(s) = 1_{[A_{it}(s) > 0]}$, so that $A_{it}(s) = 1$ denotes a case and $A_{it}(s) = 0$ denotes a control. The p-ROC curve is a plot of p-TPF(c) = $P(M_i > c | A_{it}(s) = 1)$ versus p-FPF(c) = $P(M_i > c | A_{it}(s) = 0)$, where the true and false positive fractions are functions of a cutoff point $c \in (-\infty, \infty)$ used to dichotomize the marker M [11]. Comparable quantities can be defined for the expected subject-level treatment effect using $A_i(\mathcal{S}) = 1_{[\bar{\Delta}_i(\mathcal{S}) > 0]}$ instead of $A_{it}(s)$.

By definition, p-TPF and p-FPF are percentiles of the marker distribution given case or control status at time t . Case or control status $A_{it}(s)$ is unobserved since it depends on potential outcomes, so we model p-TPF and p-FPF by combining a model for the distribution of $A_{it}(s)$ given marker status, with a model for the marginal marker distribution, via Bayes formula:

$$p\text{-TPF} = P[M_i > c | A_{it}(s) = 1] = \frac{P[A_{it}(s) = 1 | M_i > c] P[M_i > c]}{P[A_{it}(s) = 1]} \quad (1)$$

$$p\text{-FPF} = P[M_i > c | A_{it}(s) = 0] = \frac{P[A_{it}(s) = 0 | M_i > c] P[M_i > c]}{P[A_{it}(s) = 0]} \quad (2)$$

Again comparable quantities evaluating expected subject-level benefit can be constructed using $A_i(\mathcal{S})$ in place of $A_{it}(s)$. In Section 3 we propose estimation of, and inference for, p-ROC curves by combining parametric estimation of the conditional distribution $[A_t(s)|M]$ (or $[A(\mathcal{S})|M]$) and either parametric or non-parametric estimation of the marker distribution $[M]$. However, in Section 2.2 we first define a model for the potential outcomes $Y_{it}(s)$ that underlies the distribution $[A_t(s)|M]$ (or $[A(\mathcal{S})|M]$). Then we discuss the assumptions needed for estimation of p-ROC curves in Section 2.3.

2.2. Model for Treatment Benefit

We construct a model for the potential outcomes $Y_{it}(s)$, conditional on marker value M_i ; then we use the model for $Y_{it}(s)|M_i$ to construct a model for the time-specific observation-level difference in potential outcomes $_{it}(s)|M_i$ (or the expected subject-level difference $_{i}(\mathcal{S})|M_i$) that underlies case or control status $A_{it}(s)|M_i$ (or $A_i(\mathcal{S})|M_i$).

Specifically, for the potential outcomes, we use a longitudinal structural mixed model (LSMM), as defined by Sitlani et al. [12], which incorporates the following three components: 1) a model for average potential outcomes across individuals, 2) a model for within-subject variation relative to the group average, and 3) a model for measurement error across time within an individual. A general LSMM that applies to treatment at a single time is as follows:

$$Y_{it}(s)|M_i = \mathbf{L}_i(t, M_i)\boldsymbol{\lambda} + 1_{[t \geq s]}\mathbf{G}_i(t, s, M_i)\boldsymbol{\gamma} + \mathbf{Z}_i(t, s)\mathbf{b}_i + e_{it}(s) \quad (3)$$

As noted previously, t is the time of measurement, s is the time of treatment, and M_i is the marker value. The first two terms define the model for the average potential outcomes across individuals, with $\mathbf{L}_i(t, M_i)\boldsymbol{\lambda}$ representing the population average potential outcome conditional on the marker value in the absence of treatment and $\mathbf{G}_i(t, s, M_i)\boldsymbol{\gamma}$ representing the change in the average potential outcome attributable to treatment. Both $\mathbf{L}$ and $\mathbf{G}$ are matrices that can include any relevant covariates, here limited to t , s , and M_i . The model for within-subject variation from the group average incorporates subject-specific latent effects, represented by $\mathbf{b}_i$. The random-effects design matrix $\mathbf{Z}_i(t, s)$ allows the random-effects structure to vary with time and surgery time. Here we also note that while $\mathbf{L}_i(t, M_i)\boldsymbol{\lambda} + 1_{[t \geq s]}\mathbf{G}_i(t, s, M_i)\boldsymbol{\gamma}$ denotes a group or population average as a function of time, the inclusion of subject-specific random effects yields an individual average as a function of time: $\mathbf{L}_i(t, M_i)\boldsymbol{\lambda} + 1_{[t \geq s]}\mathbf{G}_i(t, s, M_i)\boldsymbol{\gamma} + \mathbf{Z}_i(t, s)\mathbf{b}_i$. Contrasts in subject-specific means associated with alternative treatment strategies can characterize the subject-to-subject variation in response to treatment and are key to a model-based summarization of subject-level expected benefit.

Finally, the model for measurement error is incorporated in $e_{it}(s)$, and can also vary with time and surgery time.

One specific example that will be used in subsequent methods development follows:

$$Y_{it}(s)|M_i = [\lambda_0 + \lambda_1 t + \lambda_2 M_i] + 1_{[t \geq s]} [\gamma_0 + \gamma_1 M_i] + [b_{i0} + b_{i1} t + b_{i2} 1_{[t \geq s]}] + [e_{it}^0 1_{[t < s]} + e_{it}^1 1_{[t \geq s]}] \quad (4)$$

The observation-level treatment benefit is:

$$\text{observation-level: } \Delta_{it}(s) = \gamma_0 + \gamma_1 M_i + b_{i2} + [e_{it}^1 - e_{it}^0]$$

and the expected subject-level treatment benefit will depend on the set of candidate surgical times and/or follow-up times considered. For example, if clinical interest is in the impact of surgery one year after treatment, but a discrete range of candidate surgical times is plausible then one subject-level expected benefit would be defined by averages over the set $\mathcal{S}_1 = \{ (t, s) : t - s = 1, s \in \Omega \}$ which yields:

$$\text{subject-level for } \mathcal{S}_1: \bar{\Delta}_i(\mathcal{S}_1) = E_{\mathcal{S}_1} [\gamma_0 + \gamma_1 M_i + b_{i2} + e_{it}^1 - e_{it}^0 | \mathbf{b}_i] = \gamma_0 + \gamma_1 M_i + b_{i2} + \bar{e}_i^1 - \bar{e}_i^0$$

where $\bar{e}_i^k, k=0/1$ denotes the mean measurement error over the set $\mathcal{S}_1$. Note that if the set $\mathcal{S}_1$ is of size m then as m increases the impact of measurement error variability decreases to zero. In the limit, or for sets of infinite size, $\mathcal{S}_2$, such as with an interval of surgical start times we obtain:

$$\text{subject-level for } \mathcal{S}_2: \bar{\Delta}_i(\mathcal{S}_2) = \gamma_0 + \gamma_1 M_i + b_{i2}$$

which now represents a subject-level expected treatment benefit that is independent of the measurement errors. In this case the conditional distribution of $\Delta_i(\mathcal{S}_2)$ is only a function of mean parameters and random effects parameters and not the measurement error variance, or most importantly does not depend on the unidentifiable parameter $\text{corr}(e_{it}^1, e_{it}^0)$.

The LSMM in equation (4) assumes that there is an immediate and sustained treatment effect, denoted $\gamma_0 + \gamma_1 M_i$. Another option would be a model that assumes an immediate, then linearly growing treatment effect, such as the following:

$$Y_{it}(s)|M_i = [\lambda_0 + \lambda_1 t + \lambda_2 M_i] + 1_{[t \geq s]} [\gamma_0 + \gamma_1 M_i + \gamma_2(t-s)] + [b_{i0} + b_{i1} t + b_{i2} 1_{[t \geq s]}] + [e_{it}^0 1_{[t < s]} + e_{it}^1 1_{[t \geq s]}] \quad (5)$$

Under this assumption, the observation-level treatment benefit now explicitly depends on time since treatment: $\Delta_{it}(s) = \gamma_0 + \gamma_1 M_i + \gamma_2(t-s) + b_{i2} + [e_{it}^1 - e_{it}^0]$, and the expected subject-level treatment benefit one year after surgery is:

$$\bar{\Delta}_i(\mathcal{S}_2) = E_{\mathcal{S}_2} [\gamma_0 + \gamma_1 M_i + \gamma_2(t-s) + b_{i2} + e_{it}^1 - e_{it}^0 | \mathbf{b}_i] = \gamma_0 + \gamma_1 M_i + \gamma_2 + b_{i2}, \text{ which again only depends on the mean parameters and the distribution of random effects.}$$

2.3. Assumptions for Estimation of p-ROC Curves

The study design plays a critical role in the ability to assess the value of a marker for treatment selection. Ideally, to identify individual variability related to treatment, i.e. variability in $y_{it}(s)$, which is necessary for identification of time-specific observation-level p-ROC curves, multiple observed subject-specific responses are needed [13, 14]. Thus, identification of the distribution of $y_{it}(s)$ requires multiple crossover periods per subject. On the other hand, variability in $\bar{y}_i(\mathcal{S})$, which is necessary for identification of expected subject-level p-ROC curves, requires less complex data. Time-specific measurement error is averaged so that multiple observed responses, but not multiple crossover periods, are needed to identify the distribution of $\bar{y}_i(\mathcal{S})$.

In a randomized controlled trial (RCT) with only a single post-treatment outcome measurement, estimation of the population distribution of either $y_{it}(s)$ or $\bar{y}_i(\mathcal{S})$ requires strong assumptions. The key issue is well known: either a single treated or an untreated potential outcome is observed, but there is no model free information about the other potential outcome, so characterizing the distribution of the difference in outcomes, $y_{it}(s)$ or $\bar{y}_i(\mathcal{S})$ relies on model assumptions. One approach would be to impute the unobserved potential outcome from an assumed joint distribution of the two potential outcomes, but such imputation would require an assumption about the within-person correlation between potential outcomes under different treatments, a quantity about which we have no information. Thus strong assumptions must be made to estimate p-ROC curves when only one post-treatment outcome is observed. In certain applications covariate information can be incorporated in order to explain components of variation, but even with additional predictors the conditional correlation between potential outcomes is not identified.

On the other hand, if the effect of treatment occurs immediately and does not change over time, then a simple two period crossover design can be employed to estimate individual change associated with treatment. Such longitudinal pre-post contrasts capture the basic idea behind a more formal principal strata contrast, but are subject to either strong assumptions or intrinsic measurement error.

With longitudinal data we could consider naive “expected benefit” summaries defined as the average of multiple pre-treated outcomes minus multiple post-treated outcomes. However, similar to the simple crossover scenario these statistics do not fully acknowledge within-subject variation and also may not account for large scale temporal trends. Therefore, we focus on methods that adopt a longitudinal regression model in order to estimate treatment accuracy summaries. By assuming that potential outcomes follow a LSMM, with correctly defined random effects, we can separate out within-subject and between-subject variability via $\mathbf{b}_i$ and $\mathbf{e}_i$. In partitioning the variability, we narrow the bounds on $\text{Var}[y_{it}(s)]$, which is crucial for characterizing a good response predictive marker [14]. Thus, in the development that follows, we propose using available longitudinal data to generate either bounds on time-specific p-ROC curves or estimates of expected p-ROC curves. For the time-specific curves, we acknowledge that we cannot fully characterize the curves even with our assumptions, so we recommend consideration of expected p-ROC curves in clinical scenarios where the exact timing of treatment may occur within a time window.

The following assumptions are required for estimating either bounds on time-specific p-ROC curves or expected p-ROC curves:

- A1.** Consistency and the stable unit treatment value assumption (SUTVA), which implies that the potential outcomes for one individual are unrelated to the treatment status of other individuals [15].
- A2.** Exchangeability of treated and untreated subjects at all times.

A1 is standard in literature on causal inference. A2 can be achieved via data from randomized trials with good compliance, or designed use of observational data.

The remaining assumptions relate to specification of the joint distribution of potential outcomes with and without treatment. Stronger assumptions are needed to estimate bounds for a time-specific p-ROC curve, as compared to estimating an expected p-ROC curve. Only one surgery time is observable in most surgery trials, but the joint distribution of individual potential outcomes at two different surgery times is necessary for estimating the performance of markers in predicting time-specific treatment benefit at the level of the individual. Therefore, we cannot fully identify the distribution of time-specific treatment benefit, and instead make plausible assumptions about correlation between subject-specific potential outcomes under different treatments, leading to bounds on time-specific p-ROC curves. On the other hand, in estimating expected treatment benefit, we average over within-subject variability over time, so the distribution of expected subject-level treatment benefit can be identified without full specification of the joint distribution of potential outcomes.

With reference to the LSMM in equation (4), assumptions A3.1 and A3.2 are required for either the time-specific observation-level or the expected subject-level p-ROC curves, while assumption A3.3 is only required for time-specific p-ROC curves:

- A3.1.** The random effects $\mathbf{b}$ are independent of the measurement errors $\mathbf{e}$.
- A3.2.** Both $\mathbf{b}$ and $\mathbf{e}$ are normally distributed, with mean zero.
- A3.3.** Measurement errors $\mathbf{e}$ at the same time, but under different treatments, e.g.

e_{it}^1 and e_{it}^0 , are jointly normally distributed, so that pairs of potential outcomes, e.g. $Y_{it(s)}|M$ and $Y_{it(t+)}|M$, are jointly normally distributed:

$$\begin{pmatrix} e_{it}^0 \\ e_{it}^1 \end{pmatrix} \sim N \left(\mathbf{0}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \right), \text{ where } \sigma_{11} = \sigma_{22} = \sigma_e^2 \text{ and } \sigma_{12} = \sigma_{21} = \rho \sigma_e^2. \text{ A3.3}$$

implies that $_{it(s)}|M_i$ is normally distributed.

To estimate a time-specific p-ROC curve, we need to estimate the distribution of $_{it(s)}$, conditional on marker value M_i . Using the LSMM in equation (4), with either cross-sectional or longitudinal data, and assumption A2, γ_0 and γ_1 are identifiable, so we can estimate the mean $E[_{it(s)}|M_i] = \gamma_0 + \gamma_1 M_i$. Estimation of the variance requires further assumptions. With cross-sectional data, $\text{Var}[_{it(s)}|M_i]$ includes both subject-to-subject variation due to treatment and within-subject variation, so $\text{corr}[Y_{it(s)}, Y_{it(t+)}|M_i]$ is unrestricted. However, with longitudinal data, b_{i2} is identifiable, so $\text{Var}[_{it(s)}|M_i]$ can be partitioned into $\text{Var}[b_{i2}]$ and $\text{Var}[e_{it}^1 - e_{it}^0]$, narrowing the bounds on $\text{corr}[Y_{it(s)}, Y_{it(t+)}|M_i]$.

Details are in Supplementary Appendix 1. Using assumption A3.3, with $\rho = \text{corr}[e_{it}^1, e_{it}^0]$ the variance of the time-specific treatment effect is

$\text{Var}[b_{i2}] + \text{Var}[e_{it}^1 - e_{it}^0] = \text{Var}[b_{i2}] + 2(1 - \rho)\sigma_e^2$. Although the distribution of $e_{it}(s)|M_i$ could be estimated under any assumption of interest for ρ , we will consider upper and lower bounds for ρ to characterize the range of plausible estimates.

Two natural assumptions are 1) that the e_{it} don't vary with treatment time, i.e. $\rho = 1$, or 2) that they vary, but are independent of each other, i.e. $\rho = 0$. The latter assumption will often be preferable as the former implies rank preservation which is a strong assumption. However, conducting two separate analyses, one making each assumption, provides answers at two natural extremes and therefore constructs bounds for the quantity of interest. Note that the mathematical extremes would be no variance with treatment time, i.e. $\rho = 1$, and opposite values at the two treatment times of interest, i.e. $\rho = -1$, but we have chosen to consider more plausible extremes since negative correlation is difficult to justify.

Alternatively, we can estimate p-ROC curves that evaluate expected treatment benefit. The subject-specific time-averaged benefit does not depend on the residual errors, and therefore the distribution for $\bar{e}_i(\mathcal{S}_1)|M_i$ is completely identifiable from observed longitudinal data, allowing us to estimate its distribution. The estimate of its mean is the same as that of the time-specific treatment benefit: $E[\bar{e}_i(\mathcal{S}_1)|M_i] = \gamma_0 + \gamma_1 M_i$. However, the measurement error terms are averaged out, so the estimate of its variance is simpler than that of the time-specific treatment benefit: $\text{Var}[\bar{e}_i(\mathcal{S}_1)|M_i] = \text{Var}[b_{i2}]$. Consequently, we recommend use of expected p-ROC curves in most practical situations.

3. Estimation and Inference: p-ROC Curves

Having discussed the assumptions necessary for estimation of the distribution of time-specific and expected treatment effects, we now use the framework presented in Section 2.1 to discuss generation of p-ROC curves to evaluate the response-predictive ability of markers. In practice, estimation of p-TPF and p-FPF is not sufficient for determining whether a given marker is useful, either on its own or as compared to other markers of interest. Therefore we also discuss generation of pointwise confidence intervals, based on the asymptotic variance of the p-ROC curve.

3.1. Estimation of p-ROC Curves

We combine estimates of the conditional distribution $[A_t(s)|M]$ (or $[A(\mathcal{S})|M]$) and the marker distribution $[M]$ to estimate p-TPF and p-FPF. We denote the numerators in equations (1) and (2) to be p-TPF₁ and p-FPF₁, with corresponding denominators being p-TPF₂ and p-FPF₂. Ideas are developed in the context of time-specific treatment benefit, but extend directly to expected benefit.

Estimation of $[A_t(s)|M]$ Estimation of the numerators in equations (1) and (2) requires estimation of the conditional distribution $[A_t(s)|M > c]$. We estimate $[Y_t(s)|M]$, then assume bivariate normality so that $[A_t(s)|M] = [Y_t(s) - Y_t(t+)|M]$ is normally distributed. The product of a quantile of this conditional distribution and the marker distribution can be integrated

over marker values greater than c to give expressions for the numerators in equations (1) and (2):

$$p\text{-TPF}_1 = P[A_{it}(s)=1|M_i>c]P[M_i>c] = \int_c^\infty P[A_{it}(s)=1|M_i=m]P[M_i=m]dm \quad (6)$$

$$p\text{-FPF}_1 = P[A_{it}(s)=0|M_i>c]P[M_i>c] = \int_c^\infty P[A_{it}(s)=0|M_i=m]P[M_i=m]dm. \quad (7)$$

Using the LSMM in Equation (4), the joint distribution of conditional potential outcomes is:

$$\begin{pmatrix} Y_{it}(s) \\ Y_{it}(t+) \end{pmatrix} | M_i \sim N \left(\begin{pmatrix} \lambda_0 + \lambda_1 t + \lambda_2 M_i + \gamma_0 + \gamma_1 M_i \\ \lambda_0 + \lambda_1 t + \lambda_2 M_i \end{pmatrix}, \begin{pmatrix} \text{Var}[Y_{it}(s)|M_i] & \text{Cov}[Y_{it}(s), Y_{it}(t+)|M_i] \\ \text{Cov}[Y_{it}(s), Y_{it}(t+)|M_i] & \text{Var}[Y_{it}(t+)|M_i] \end{pmatrix} \right)$$

Thus, the mean $\mu_{|M} = E[Y_{it}(s)|M_i] = \gamma_0 + \gamma_1 M_i$, and the conditional variance

$$\sigma_{\Delta|M}^2 = \text{Var}[\Delta_{it}(s)|M_i] = \text{Var}[b_{i2}] + \text{Var}[e_{it}^1 - e_{it}^0].$$

Under the assumptions in Section 2.3, a linear mixed effects model can be fitted to observed longitudinal data to obtain estimates of γ , $\text{Var}[b_{i2}]$, and σ_e^2 . We obtain an estimate of $\mu_{|M}$ by plugging in estimates of γ . To estimate bounds for time-specific p-ROC curves, we obtain

two different estimates of $\sigma_{\Delta|M}^2$ by plugging in estimates of $\text{Var}[b_{i2}]$ and σ_e^2 into 1)

$\text{Var}[b_{i2}] + 2\sigma_e^2 (\rho=0)$ and 2) $\text{Var}[b_{i2}] (\rho=1)$. The p-ROC curve analysis is conducted using each of these conditional variance estimates to illustrate the range of discriminatory ability of the marker.

The expected p-ROC curve analysis has the same $\mu_{|M}$, with $\hat{\sigma}_{\Delta|M}^2 = \text{Var}[\hat{b}_{i2}]$.

Estimation of $[M]$ An estimate of the marker distribution is also necessary. The methods for estimation and inference presented in this manuscript permit the choice of any parametric model for the marker distribution. For simplicity, we assume a normal distribution, characterized by the sample mean $\bar{M}$ and the sample variance S_M^2 .

Combining normality of the conditional difference in outcomes with normality of the marker, we estimate the numerators of equations (1) and (2) by plugging in estimates of parameters of normal distributions, and integrating over marker values greater than c , as specified in equations (6) and (7). Integration over all values gives estimates of the denominators. The ratios of numerator and denominator estimates provide estimates of p-TPF and p-FPF.

3.2. Inference for p-ROC Curves

Inference regarding the response-predictive ability of a marker can be accomplished via pointwise confidence intervals for p-ROC curves. Standard methods for generating pointwise confidence intervals exist [11, 16]. However standard methods rely on the independent use of cases to estimate TPF and controls to estimate FPF. They do not account for the use of all subjects in estimating both TPF and FPF, as we propose here.

We generate pointwise confidence intervals using the asymptotic variance of a point on a p-ROC curve. The p-ROC function is a composite function of p-TPF and p-FPF: $p\text{-ROC}(p) = p\text{-TPF}(p\text{-FPF}^{-1}(p))$, where $p\text{-FPF}^{-1}(p)$ is the cutoff corresponding to a false positive rate of p .

Our estimate $p\text{-}\hat{\text{ROC}}(p)$ is a function of both the LSMM parameter and variance estimates, which we denote $\hat{\theta}$, and the estimated marker distribution, which we denote $\hat{F}_M$. We can separate the difference between the estimate $p\text{-}\hat{\text{ROC}}(p)$ and its true value $p\text{-ROC}(p)$ into two independent parts, the first a function of $\hat{\theta}$ conditional on marker value M and the second a function of $\hat{F}_M$:

$$\left[p\text{-}\hat{\text{ROC}}(p) - p\text{-ROC}(p) \right] = \underbrace{R(p, \hat{\theta}, \hat{F}_M) - R(p, \theta, \hat{F}_M)}_{R_A} + \underbrace{R(p, \theta, \hat{F}_M) - R(p, \theta, F_M)}_{R_B}. \quad (8)$$

Note that we use R as an abbreviation for p-ROC when indicating the dependence on θ and

F_M . Using this formulation, $\text{Var} \left[p\text{-}\hat{\text{ROC}}(p) - p\text{-ROC}(p) \right] = \text{Var}(R_A) + \text{Var}(R_B)$. Because M is considered fixed in the first component, $\text{Var}(R_A)$ is a function of the variance in the estimation of $[A_i(s)|M]$; $\text{Var}(R_B)$ is a function of the variance in the estimation of $[M]$. The asymptotic variances $\text{Var}(R_A)$ and $\text{Var}(R_B)$ can be calculated using multivariate delta methods [17]:

$$\text{Var}(R_A) = \left(\frac{\partial R(p, \hat{\theta}, \hat{F}_M)}{\partial \hat{\theta}} \right)' (\text{Var}(\hat{\theta})) \left(\frac{\partial R(p, \hat{\theta}, \hat{F}_M)}{\partial \hat{\theta}} \right) \quad (9)$$

$$\text{Var}(R_B) = \left(\begin{array}{c} \frac{\partial R(p, \theta, \hat{F}_M)}{\partial \bar{M}} \\ \frac{\partial R(p, \theta, \hat{F}_M)}{\partial S_M^2} \end{array} \right)' \left(\begin{array}{cc} \text{Var}\{\bar{M}\} & \text{Cov}\{\bar{M}, S_M^2\} \\ \text{Cov}\{\bar{M}, S_M^2\} & \text{Var}\{S_M^2\} \end{array} \right) \left(\begin{array}{c} \frac{\partial R(p, \theta, \hat{F}_M)}{\partial \bar{M}} \\ \frac{\partial R(p, \theta, \hat{F}_M)}{\partial S_M^2} \end{array} \right) \quad (10)$$

Detailed calculations of equations (9) and (10) that are specific to the parametric assumptions made in Section 3.1 are provided in Supplementary Appendix 2. A logistic transformation can be used to constrain the limits of the pointwise confidence intervals to be between zero and one.

3.3. Semi-parametric Estimation and Inference for p-ROC Curves: Flexible Marker Distribution

The p-ROC curve estimator proposed in Section 3.1 requires parametric assumptions about the marker distribution. Not every marker distribution can be modeled with a simple parametric form. Therefore, in this section, we propose a modified p-ROC curve estimator that allows the marker distribution to be estimated more flexibly.

One can obtain a fully nonparametric estimator of the marker distribution by replacing the parametric distribution with an empirical cumulative distribution function (cdf):

$$\hat{F}_n(m) = \frac{1}{n} \sum_{i=1}^n 1_{[M_i \leq m]} \quad [17].$$

Combined with estimates of the conditional distribution

$[A_T(s)|M]$, semi-parametric estimates of the p-ROC curve follow from estimates of the numerators and denominators in equations (1) and (2). The ratios give p-ROC curve estimates, which are functionals of empirical estimates. Asymptotic inference for functionals of empirical estimates typically proceeds via the functional delta method, which follows from empirical process theory. However, the discontinuity in the cutoff c that is defined by the inverse p-FPF prevents the differentiability conditions that are necessary for use of the functional delta method from being satisfied. Therefore we employ smoothed empirical process theory instead.

Empirical process theory proves the asymptotic convergence of empirical cdf estimates to Brownian bridge processes [17]. Smoothed empirical process theory extends these results to smoothed cdf estimates [18, 19]. One type of smoothed cdf estimate incorporates a kernel estimate of the corresponding probability density function (pdf) [18, 19]:

$$\hat{f}(m) = \frac{1}{n} \sum_{i=1}^n K_h(m - m_i)$$

where m is a one-dimensional marker value for which we have data values m_i with $i = 1, \dots, n$, and the kernel $K_h(m)$ has bandwidth h such that $K_h(m) = \frac{1}{h} K(\frac{m}{h})$. An analog to the functional delta method has been established for functionals of kernel density estimates [20]. We construct semi-parametric p-ROC curve estimates by combining a kernel density estimate of the marker distribution with the estimate of the conditional distribution $[A_T(s)|M]$ that we developed in Section 3.1.

The estimation procedure is the same as in Section 3.1, except for estimation of the marker density. Instead of using normal density estimates for $P[M = m]$, we now use kernel density estimates. For details regarding kernel and bandwidth selection, please see Supplementary Appendix 3. For details regarding confidence bands and inference, please see Supplementary Appendix 4.

4. Simulations

We conducted simulation studies using longitudinal data with several goals in mind: 1) to illustrate the unbiasedness of our proposed estimator of the p-ROC curve, 2) to evaluate coverage of asymptotic confidence intervals in finite samples, and 3) to evaluate the robustness of our proposed methods to model misspecification. For each scenario of interest we generated 1000 simulated datasets.

Each simulated dataset included 500 subjects, each with 5 observations. To reflect a randomized trial with longitudinal follow-up, all subjects have a baseline outcome measurement; then half of the subjects receive treatment, after which all of the subjects have 4 equally-spaced follow-up measurements. For the primary scenario, with a correctly-specified model, the baseline response-predictive marker is generated from a normal distribution with mean=0 and variance=4, while the underlying LSMM from which data are generated is:

$$Y_{it}(s) | M_i = 28 + 0.1 \cdot t + 0.5 \cdot M_i + 1_{[t \geq s]} [1 + 0.2 \cdot M_i] + b_{i0} + b_{i1}t + b_{i2}1_{[t \geq s]} + e_{it}$$

with $\mathbf{b}_i \perp \mathbf{e}_i$, both $\mathbf{b}_i$ and $\mathbf{e}_i$ normally distributed, and correlation between pairs of random effects set to 0.5.

We also explored three different scenarios where the model used to generate the data did not satisfy the assumptions underlying our methods:

- S1.** Non-normal errors (Supplementary Figure 2): measurement errors $\mathbf{e}$ were sampled from a chi-squared distribution with three degrees of freedom that was centered at 0 and standardized to have the same variance (1.33) that was used in the primary scenario.
- S2.** Non-normal random effects (Supplementary Figure 3): $\mathbf{b}_i$ were sampled from a 30–70 mixture of two multivariate normal distributions: one centered at (0,0,0) and one centered at (1.5,0.5,1.5), each with variances equal to one-third of those in the primary scenario so that the variances of the mixture were similar to those in the primary scenario. To ensure that the random effects had expectation zero, we subtracted off the weighted means (1.05,0.35,1.05).
- S3.** Non-normal marker distribution (Supplementary Figure 4): M was simulated from a 30–70 mixture of normal distributions with means at -1 and 1 , each with variance of 2.2 so that the variance of the mixture was similar to the variance in the primary scenario.

When the underlying distributions of the marker and the potential outcomes are normal, we can use the known true marginal distributions of the marker M and the treatment effect of interest, as well as their joint distribution, to compute the true p-TPF and p-FPF for each cutoff c :

$$\text{p-TPF} = \frac{P[M > c, \Delta > 0]}{P[\Delta > 0]} \text{ and } \text{p-FPF} = \frac{P[M > c, \Delta > 0]}{P[\Delta \leq 0]}.$$

When the underlying distributions of the marker and/or the potential outcomes are not normal, the joint distribution is more complicated. Therefore, in the three scenarios where the model used to generate the data included some non-normality, we found the true joint distribution by generating 10 million (M, Δ) pairs, then computing the empirical p-FPF and p-TPF values at a series of cutoffs c . That is, for each cutoff c , the true p-TPF and p-FPF are:

$$\text{p-TPF} = \frac{\# \text{ of pairs with } M > c \text{ and } \Delta < 0}{\# \text{ of pairs with } \Delta > 0} \quad \text{p-FPF} = \frac{\# \text{ of pairs with } M > c \text{ and } \Delta \leq 0}{\# \text{ of pairs with } \Delta \leq 0}.$$

Regardless of the method used to compute true values, we used 1000 equally-spaced cutoffs within three standard deviations above and below the mean of the marker.

To compare the estimated curves to the true one, we interpolated the estimated p-TPF value at each of the 1000 true p-FPF points for each of the 1000 simulations. We also interpolated the upper and lower confidence interval estimates. We then plotted the true p-ROC curve, the average estimated p-TPF, and the average estimated confidence intervals.

Figure 1 displays the average estimated bounds and the true p-ROC curve, when evaluating the time-specific observation-level treatment benefit in the primary scenario. The lower estimated bound corresponds to the true curve because the simulated truth and the lower bound both assume independent errors. In applications, the true correlation is likely to be between 0 and 1, so that the true curve would be between the two bounds, rather than coincident with one of the bounds. Assuming the true correlation to be equal to 0 or 1, the confidence intervals for the estimated bounds cover the true value approximately 95% of the time, except when the p-FPF is very close to 0 or 1 (data not shown). The confidence intervals using the logistic transformation show more consistent coverage across the range of p-FPF values than the untransformed confidence intervals do.

Figure 2 displays the average estimated p-ROC curve and the true p-ROC curve, when evaluating the expected subject-level treatment benefit in the primary scenario. The estimated curve is quite close to the true curve, with the confidence intervals displaying similar behavior to that described for the time-specific p-ROC curves.

In scenario S1, where we simulated non-normal errors, performance of our methods for estimation of both the expected p-ROC curve and the bounds on the time-specific p-ROC curve was quite similar to that seen with normal errors, implying robustness to misspecification of the distribution of measurement errors. Details are in Supplementary Figures 5 and 6.

In scenario S2, where we simulated non-normal random effects, there was some loss of accuracy of our methods, but it was not substantial; details are in Supplementary Figures 7 and 8. To ensure validity of p-ROC curve estimates, we recommend that investigators evaluate the assumption of normality of the random effects [21]. Should this assumption be violated, our methods could be extended to relax the assumption via more complicated estimation of the LSMM parameters, e.g. using a Bayesian approach with a normal mixture or Dirichlet process prior [22, 23].

In scenario S3, where we simulated a non-normal marker, there was more substantial loss of accuracy, particularly for expected p-ROC curves (Figure 3 and Supplementary Figure 9). Therefore Section 3.3 describes an extension of our methods that includes nonparametric estimation of the marker distribution. This modified method performs well even when the marker is not normally distributed, as shown in Figure 4 and Supplementary Figure 10.

5. Illustration

In this section we evaluate potential response-predictive markers in the context of treatment for CTS. We use data from a multi-site clinical trial comparing surgical to non-surgical therapy in subjects with CTS ($n=116$) [10]. The outcome of interest is functional status in

the hand, which was assessed at baseline and every 3 months thereafter over a 12 month period. Crossover to surgery was allowed after 3 months.

Markers of interest measured by EDS include median motor latency (MML) (n=107), median motor amplitude (MMA) (n=104), and median-ulnar sensory latency difference (n=62). Markers of interest measured by MRNI include median nerve signal (n=65). We conduct separate analysis for each marker, with subjects excluded from analyses for which they are missing data. Two treatment-marker interactions have been found in these data - one for MML as measured by EDS and the other for median nerve signal as measured by MRNI [4].

In estimating p-ROC curves, we assume that markers are normally distributed. To ensure that this assumption is at least approximately true, we investigate simple transformations of marker data. Using histograms and kernel density estimates (Supplementary Figure 1), as well as corresponding normal qq-plots, we conclude that untransformed values of MMA should be used in analyses, while log-transformed MML and square-root-transformed sensory latency difference and median nerve signal should be used in analyses. We use the sample mean and variance of each marker as parameter estimates for the normal distribution.

To estimate the distribution of $A_t|M$, we fit the following linear mixed effects model, based on a priori specification of a plausible model for these data:

$$E[Y_{it}] = \lambda_0 + \lambda_1 \text{age}_i + \sum_{j=2}^4 \lambda_j \text{site}_{ij} + \lambda_5 t + \lambda_6 M_i + 1_{[t \geq s_i]} [\gamma_0 + \gamma_1 M_i]$$

with random intercepts, random slopes on time t , and random effects on the treatment $1_{[t \geq s_i]}$. Time was modeled linearly for simplicity; however, exploratory analyses adding a quadratic term were performed. Model fit, as measured by Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), was worse with the higher-order term for time, supporting our decision to use the simpler model. Valid estimates of the LSMM parameters require that the random effects are correctly specified and that either treatment is randomized or selection to treatment does not depend on factors that are not included in the model [12]. The first assumption is plausible given the complex structure that we've assumed for random effects. We included trial site in the model to satisfy the second assumption, due to the finding that in these data noncompliance with randomized treatment assignment is best predicted by previous outcomes of interest and trial site. Parameter estimates, standard errors, and variance component estimates are in Supplementary Table 1. An inverse relationship exists between MML and outcomes, so negative MML values are used to reverse the relationship.

In practice we believe that expected p-ROC curves are generally more useful than estimated bounds on time-specific p-ROC curves, due to the weaker assumptions required for validity, and due to the focus on an expected benefit rather than potentially arbitrary time-specific benefit. However, in some applications there is a single primary outcome time such as health status one year after surgery, and therefore a marker for time-specific treatment benefit

could be evaluated using an observation-level p-ROC curve. Therefore we have estimated both expected p-ROC curves (Figure 5) and observation-level p-ROC curves (Figure 6) for each marker of interest. We have also estimated the area under the curve (AUC) corresponding to these p-ROC curves to facilitate comparisons across markers and with other metrics. In addition, we think it is important to illustrate the observation-level p-ROC curve to highlight the difference in scientific focus and difference in statistical properties as compared to the expected p-ROC curve.

Based on the expected p-ROC curve estimates, sensory latency difference is the most response-predictive of expected treatment benefit (Figure 5). However, none of the markers have particularly strong predictive ability, as the largest AUC in the analysis of expected benefit is 0.63. Similarly, none of the markers have strong predictive ability in the context of observation-level treatment benefit (Figure 6). The predictive ability is particularly poor when we assume that observation-level measurement errors at treated times are independent of those at untreated times. If observation-level measurement errors at treated times are correlated with those at untreated times, then the performance of the markers in predicting individual response to treatment improves.

Asymptotic confidence intervals reveal that based on these data, none of the four markers is more useful than ignoring markers in making treatment decisions, regardless of whether expected p-ROC curves or observation-level p-ROC curves are used. Mixed models including the random effect on treatment do not have substantially better fit than those without the random effect on treatment, indicating that the underlying propensity to benefit from surgery does not vary much across participants. Some variability in the underlying propensity for treatment benefit must exist for markers to be useful in identifying individuals who will benefit, so this lack of variability is consistent with the lack of predictive accuracy of potential markers for treatment. Although these results illustrate the use of the methods proposed in this manuscript, further analysis with larger sample sizes would be necessary to confirm clinical conclusions.

6. Discussion

Identifying the patients who will benefit most from treatment is a common goal of clinical studies. By constructing models for potential outcomes that incorporate information from baseline markers, we delineate the assumptions that are necessary to bound the distribution of individual treatment effects. Combining the model for outcome given marker status with a model for the marker distribution, we propose a method for examining the predictive accuracy of a marker with respect to either time-specific or time averaged subject-level benefit from treatment. Time-specific benefit incorporates specific outcome and treatment times, whereas expected benefit averages over outcome times. Both due to this average interpretation, and to the weaker assumptions that are needed for estimation, we recommend use of expected benefit in practical analyses. Asymptotic confidence intervals can be calculated to facilitate evaluation of the usefulness of markers, both in relation to not using a marker to make treatment decisions and in relation to other markers.

Standard methods for evaluation of the usefulness of markers include evaluation of the existence of interactions between the marker of interest and the treatment effect. However, differential benefit by marker value is not necessarily the correct criterion for establishing the usefulness of a response-predictive marker. By incorporating models for subject-specific potential outcomes, we ensure that we will not identify markers that predict treatment effects that vary by marker value, but do not influence optimal treatment. We also incorporate the marker distribution into our evaluation of accuracy, ensuring that we will not identify only small subgroups of people who will benefit from the use of a marker in treatment decisions. In data from subjects with CTS, previous evaluation of interaction provided different results than did our evaluation of accuracy; however, further analyses using larger sample sizes is recommended to inform clinical practice.

Huang et al. [24] recently published a similar individual-level, potential-outcomes approach for assessing treatment-selection markers. Although their approach uses the same underlying framework as ours, they use a monotonicity assumption to identify parameters of interest in the context of randomized treatments with binary outcomes, under standard assumptions for causal inference [24]. The monotonicity assumption implies that treatment either will not do any harm or will not have any benefit. Although this assumption is reasonable in their examination of vaccine effects on HIV status, it will not be reasonable in many other contexts; therefore, further assumptions about the correlation between potential outcomes will often be needed to identify the parameters of interest, as they discuss in their section 2.4. As we have pointed out here, longitudinal data will help narrow the range of plausible assumptions.

Further, their methods only apply to binary and survival outcomes, while our methods permit quantitative outcomes. Further research would be needed to verify the usefulness of markers identified via our methods. Simon [25, 26] has discussed possible designs of clinical trials that would evaluate the use of markers in making treatment decisions. However, before such trials could be conducted, work would need to be done on methods for choosing the best cutoff to use for a particular marker. A biomarker-adaptive design has been proposed for this purpose [27], but may need to be modified to incorporate the estimation of accuracy that we have proposed.

In this manuscript, we focus on simple exogenous binary treatments. The concepts that we discuss apply to any type of treatment; however, the specific assumptions informing the structure of the LSMM, and the corresponding estimate of treatment benefit, would need to be modified for application to other types of treatment.

The methods that we propose for generating and evaluating p-ROC curves are subject to several limitations. Estimation of bounds on the distribution of individual treatment effects relies on assumptions about the measurement error in outcomes. The use of longitudinal data permits estimation of bounds under a narrower range of assumed correlations between measurement error before and after treatment, but cannot eliminate the need for such assumptions when estimating time-specific p-ROC curves. Thus we recommend using time averaged p-ROC curves, subject to weaker assumptions, when analyzing clinical data. Additionally, both the marker distribution and the LSMMs used to estimate the effect of the

marker on outcomes require parametric assumptions. When a marker can be easily transformed to achieve normality, as in our illustration, concerns about the parametric assumptions on the marker distribution are alleviated. However, if normality cannot be achieved by transformation, then we recommend using the modified methods included in this manuscript to relax the parametric assumptions on the marker distribution by using a non-parametric marker distribution estimator such as a kernel density estimator. Further work is needed to relax the LSMM assumptions, so they should be checked to justify use of the methods in their current form.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This manuscript makes use of data generated in studies led by Dr. Jeffrey Jarvik and managed by Bryan Comstock. The authors thank both of them for their support in conducting this research.

Contract/grant sponsor: NIH UL1 TR000423, NIH R01 HL072966, NIH P01 CA053996, NIH T32 HL07183

References

1. Gerritsen A, de Vet H, Scholten R, Bertelsmann F, de Krom M, Bouter L. Splinting vs surgery in the treatment of carpal tunnel syndrome: a randomized controlled trial. *JAMA*. 2002; 288:1245–1251. [PubMed: 12215131]
2. Hui A, Wong S, Leung C, Tong P, Mok V, Poon D, Li-Tsang C, Wong L, Boet R. A randomized controlled trial of surgery vs steroid injection for carpal tunnel syndrome. *Neurology*. 2005; 64(12): 2074–2078. [PubMed: 15985575]
3. Ly-Pen D, Andreu J, de Blas G, Sanchez-Olaso A, Millan I. Surgical decompression versus local steroid injection in carpal tunnel syndrome: a one-year, prospective, randomized, open, controlled clinical trial. *Arthritis Rheum*. 2005; 52:612–619. [PubMed: 15692981]
4. Jarvik J, Comstock B, Kliot M, Turner J, Chan L, Heagerty P, Hollingworth W, Kerrigan C, Deyo R. Surgery versus non-surgical therapy for carpal tunnel syndrome: a randomised parallel-group trial. *Lancet*. 2009; 374(9695):1074–1081. [PubMed: 19782873]
5. Janes H, Pepe M, Bossuyt P, Barlow W. Measuring the Performance of Markers for Guiding Treatment Decisions. *Ann Intern Med*. 2011; 154:253–259. [PubMed: 21320940]
6. Jarvik J, Comstock B, Heagerty P, Haynor D, Fulton-Kehoe D, Kliot M, Franklin G. Magnetic resonance imaging compared with electrodiagnostic studies in patients with suspected carpal tunnel syndrome: predicting symptoms, function, and surgical benefit at 1 year. *J Neurosurgery*. 2008; 108:541–550.
7. Gunter L, Zhu J, Murphy S. Variable selection for qualitative interactions. *Statistical Methodology*. 2011; 8(1):42–55. [PubMed: 21179592]
8. Gail M, Simon R. Testing for Qualitative Interactions between Treatment Effects and Patient Subsets. *Biometrics*. 1985; 41(2):361–372. [PubMed: 4027319]
9. Frangakis C, Rubin D. Principal Stratification in Causal Inference. *Biometrics*. 2002; 58:21–29. [PubMed: 11890317]
10. Martin B, Levenson L, Hollingworth W, Kliot M, Heagerty P, Turner J, Jarvik J. Randomized clinical trial of surgery versus conservative therapy for carpal tunnel syndrome. *BMC Musculoskelet Disord*. 2005; 6(2)
11. Pepe, M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press; 2003.

12. Sitlani C, Heagerty P, Blood E, Tosteson T. Longitudinal structural mixed models for the analysis of surgical trials with noncompliance. *Statist Med.* 2012; 31:1738–1760.
13. Senn S. Individual Therapy: New Dawn or False Dawn? *Drug Inf J.* 2001; 35:1479–1494.
14. Senn S. Three things that every medical writer should know about statistics. *J Eur Med Writers Assoc.* 2009; 18:159–162.
15. Bellamy S, Lin J, Have TT. An introduction to causal modeling in clinical trials. *Clin Trials.* 2007; 4:58–73. [PubMed: 17327246]
16. Hsieh F, Turnbull B. Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve. *Ann Stat.* 1996; 24(1):25–40.
17. van der Vaart, A. *Asymptotic Statistics.* Cambridge, UK: Cambridge University Press; 1998.
18. Yukich J. Weak Convergence of Smoothed Empirical Processes. *Scand J Statist.* 1992; 19(3):271–279.
19. van der Vaart A. Weak Convergence of Smoothed Empirical Processes. *Scand J Statist.* 1994; 21(4):501–504.
20. Cuevas A, Romo J. Differentiable Functionals and Smoothed Bootstrap. *Ann Inst Statist Math.* 1997; 49(2):355–370.
21. Lange N, Ryan L. Assessing Normality in Random Effects Models. *Ann Statist.* 1989; 17(2):624–642.
22. Ohlssen D, Sharples L, Spiegelhalter D. Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. *Statist Med.* 2007; 26:2088–2112.
23. Ghosh P, Rosner G. A semi-parametric Bayesian approach to average bioequivalence. *Statist Med.* 2007; 26:1224–1236.
24. Huang Y, Gilbert P, Janes H. Assessing Treatment-Selection Markers using a Potential Outcomes Framework. *Biometrics.* 2012 first published online.
25. Simon R. Development and Evaluation of Therapeutically Relevant Predictive Classifiers Using Gene Expression Profiling. *J National Cancer Inst.* 2006; 98(17):1169–1171.
26. Simon R. Development and validation of biomarker classifiers for treatment selection. *J Stat Plan Infer.* 2008; 138:308–320.
27. Jiang W, Freidlin B, Simon R. Biomarker-Adaptive Threshold Design: A Procedure for Evaluating Treatment With Possible Biomarker-Defined Subset Effect. *J National Cancer Inst.* 2007; 99(13): 1036–1043.

Figure 1.
Average estimated bounds for time-specific p-ROC curve and pointwise confidence intervals.

Figure 2.

Average estimated expected subject-level p-ROC curve and pointwise confidence intervals.

Figure 3.

Marker distribution nonnormal: Average estimated expected subject-level p-ROC curve and pointwise confidence intervals.

Figure 4.

Marker distribution nonnormal, using semiparametric methods: Average estimated expected subject-level p-ROC curve and pointwise confidence intervals.

Figure 5.
Estimates of expected subject-level p-ROC curves and pointwise confidence intervals for CTS markers.

Figure 6.

Estimates of observation-level p-ROC curve bounds for EDS and MRNI markers. The upper bounds assume $\rho = 1$, and the lower bounds assume $\rho = 0$. Confidence intervals are not included on these plots; however, as in the previous figure, none of the confidence intervals exclude the diagonal line.