Published in final edited form as:

Stat Med. 2018 April 30; 37(9): 1439–1453. doi:10.1002/sim.7608.

Evaluation of biomarkers for treatment selection using individual participant data from multiple clinical trials

Chaeryon Kang^{1,*}, Holly Janes², Parvin Tajik^{3,4}, Henk Groen⁵, Ben W. J. Mol⁶, Corine M. Koopmans⁷, Kim Broekhuijsen⁷, Eva Zwertbroek⁵, Maria G. van Pampus⁸, and Maureen T M Franssen⁷

¹Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, U.S.A ²Vaccine and Infectious Disease Division and Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, U.S.A ³Department of Clinical Epidemiology & Biostatistics, University of Amsterdam, The Netherlands ⁴Department of Pathology, Academic Medical Center, Amsterdam, The Netherlands ⁵Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands ⁶The Robinson Research Institute, School of Medicine, University of Adelaide, Adelaide, SA, Australia ⁷Department of Obstetrics and Gynecology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands ⁸Department of Obstetrics and Gynecology, Onze Lieve Vrouwe Gasthuis, Amsterdam, The Netherlands

Abstract

Biomarkers that predict treatment effects may be used to guide treatment decisions, thus improving patient outcomes. A meta-analysis of individual participant data (IPD) is potentially more powerful than a single-study data analysis in evaluating markers for treatment selection. Our study was motivated by the IPD that were collected from two randomized controlled trials of hypertension and pre-eclampsia among pregnant women to evaluate the effect of labor induction over expectant management of the pregnancy in preventing progression to severe maternal disease. The existing literature on statistical methods for biomarker evaluation in IPD meta-analysis have evaluated a marker's performance in terms of its ability to predict risk of disease outcome, which do not directly apply to the treatment selection problem. In this study, we propose a statistical framework for evaluating a marker for treatment selection given IPD from a small number of individual clinical trials. We derive marker-based treatment rules by minimizing the average expected outcome across studies. The application of the proposed methods to the IPD from two studies in women with hypertension in pregnancy is presented.

Keywords

HYPITAT trials; Individual participant data; Randomized clinical trial; Treatment selection biomarker

*Correspondence to: Chaeryon Kang, Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, U.S.A. crkang@pitt.edu.

†Please ensure that you use the most up to date class file, available from the SIM Home Page at www.interscience.wiley.com/jpages/0277-6715

1. Introduction

There is great interest in understanding treatment effect heterogeneity across subjects in clinical trials and drug development. In the presence of treatment effect heterogeneity, subjects respond to the same treatment differently and thus it is of interest to identify subgroups of subjects who would be more likely to benefit from given treatment based on subject characteristics. A treatment selection marker is a biomarker that can predict treatment effect heterogeneity in a certain target population and thus useful for clinicians and patients to make more informed treatment decisions. There is a growing literature on development and evaluation of a marker for treatment selection [1, 2, 3, 4, 5, 6]. The fundamental task is to estimate the treatment effect as a function of a marker, and the most common evaluation metric for evaluating the marker is the expected outcome under marker-based treatment [7, 8, 9, 2, 10, 11, 5, 3, 6, 12].

A meta-analysis of individual participant data (IPD) is potentially more powerful than a single-study data analysis in evaluating a marker for treatment selection. Unlike meta-analysis of aggregated summary statistics over multiple studies, the raw participant-level data are available from individual studies in the IPD setting. Motivating our work, individual patient data from two randomized and controlled trials of hypertensive disorders among pregnant women (HYPITAT: Hypertension and Pre-eclampsia Intervention Trial At Term and HYPITAT II) were collected to evaluate the effect of labor induction over expectant management of the pregnancy in preventing progression to severe maternal disease, morbidity and mortality. Overall, labor induction had a benefit in reducing the risk of progression to severe maternal disease, but previous analyses of the the first study, which we call HYPITAT I [13], showed evidence that not all women with hypertension and pre-eclampsia and their child benefit from labor induction. Therefore, it is of interest to identify a subgroup of women who are not likely to benefit from labor induction over expectant management of the pregnancy.

The vast majority of existing literature on statistical methods for biomarker evaluation in IPD meta-analysis have proposed statistical frameworks for evaluating a marker's discriminative performance (for example, C-statistic) [14, 15, 14, 16, 17, 18]. However, these methods do not directly apply to the treatment selection problem, where the task is to evaluate a biomarker's capacity to predict the treatment effect. A marker that predicts the risk of disease (called "prognostic marker") does not always predict treatment effects, and a marker which does not predict the risk of disease on a given treatment can be useful for treatment selection [10]. For example, gender and histology are strong prognostic makers for patients with advanced non-small cell lung cancer, but these markers do not predict a differential survival benefit from erlotinib over placebo [19]. A good example of the latter case is human epidermal growth factor receptor 2 (HER-2) that predicts the effect of some adjuvant chemotherapies and hormone therapies without a strong association with cancer recurrence in breast cancer patients [20, 21]. For more details, see Clark [19], Simon [22] and Ballman [23].

There has been relatively little work on statistical methods for evaluating a marker for predicting treatment effect, also called a “predictive” or “prescriptive” marker in the context of conventional meta-analysis (meta-analysis of aggregate data) or IPD meta-analysis [24, 25, 26, 27, 28, 29, 30, 31]. Those existing studies have evaluated a marker based on hypothesis testing and estimation of marker-by-treatment interactions. However, a marker-by-treatment interaction coefficient is not an adequate measure for evaluating marker performance for treatment selection due to its lack of clinical interpretation and ability to characterize the benefit of the marker [10, 32]. There is also work on evaluating treatment selection markers in a single-study setting which does not directly apply to multiple studies (for example, evaluating markers using the STEPP method [33]), and much of this has also focused on interaction testing (for example, a proportional interaction model [34, 35]).

Therefore, this paper aims to propose a statistical framework for developing a rule for guiding treatment based on marker, and for evaluating its performance, given IPD consisting of a small number of randomized controlled trials.

The remainder of this article is organized as follows. In Section 2, we provide detailed data information of our motivation data. We establish a statistical framework for treatment selection marker evaluation in IPD from multiple clinical trials in Section 3. A simulation study is presented in Section 4. Then we apply the methods to the studies in women with hypertension in pregnancy in Section 5. We conclude with our thoughts for future work in Section 6.

2. Motivating study: the HYPITAT I and HYPITAT II trials

The present study is motivated by an IPD consisting of two randomized controlled trials of non-sever hypertensive disorders among pregnant women: HYPITAT I ($n = 756$) [13] and HYPITAT II ($n = 680$) [36, 37]. The HYPITAT I and HYPITAT II trials enrolled pregnant women with pregnancy induced hypertension or mild pre-eclampsia and singleton pregnancies at 36-41 weeks and 34-37 weeks of gestational age, respectively from multiple hospitals in the Netherlands. These trials randomized women to labor induction or expectant management. Under expectant management, no intervention was undertaken; both mother and baby were monitored until the onset of a spontaneous labor or until deterioration of the woman or fetus justified delivery. Cervical length and the Bishop score [38], a composite of five measurements from vaginal examination, were measured at baseline and are factors known to predict progression to severe maternal disease [39]. Longer cervix or lower Bishop score indicate a less favorable cervical condition, so mothers who have a longer cervix or a lower Bishop score are at higher risk of progression to severe maternal disease. Overall, labor induction had a benefit in reducing the risk of progression to severe maternal disease, defined as a composite measure of several conditions (maternal mortality, maternal morbidity, progression to severe disease and major postpartum hemorrhage). However, analysis of HYPITAT I showed that not all women were likely to benefit from labor induction: for women with a long cervix labor induction reduced the risk of progression to severe maternal disease (31.7% under labor induction; 48.6% under expectant management), but induction had no significant benefit for women with a short cervix (32.9% under labor induction; 38.6% expectant management) [39]. Tajik et al. [39] reported a significant

cervical length-by-treatment interaction on the progression to severe maternal disease (p-value 0.03). Inducing labor poses risks to baby and mother, for example, thought to increase the Caesarean section rate, so avoiding unnecessary induction is compelling.

The HYPITAT I and HYPITAT II trials had some differences in terms of eligibility criteria, treatment protocol, disease prevalence and available covariates as shown in Table 1 of the supplementary materials. HYPITAT II enrolled mothers at an earlier gestational age than HYPITAT I, thus it bears an additional risk of neonatal death due to preterm birth. It is important to take into account such differences between the studies when evaluating a potential treatment selection marker.

Our goal is to use these two data sets as an empirical example to demonstrate our framework to determine whether a subgroup of women can be identified for whom there is no expected benefit of labor induction, as measured by a decreased risk of progression to severe maternal disease. Combining data from HYPITAT I and II allows us to evaluate a treatment selection marker for high-risk pregnant women which may be more broadly applicable (for example, applicable to mothers in a broader range of the gestational age) with a greater statistical power, and yet potential differences between the two studies need to be taken into account.

3. Statistical methods

3.1. Context and notation

Let D denote an adverse outcome following treatment which we refer to as “disease”, and assume that a smaller value of D is preferable. In our study, D is a binary outcome indicating occurrence of progression to severe maternal disease. Suppose that the task is to decide, for each individual subject, between two strategies denoted by T , where we call $T=1$ “treatment” and $T=0$ “no treatment”. We assume that treatment is randomly assigned, and the default treatment strategy is to treat all subjects ($T=1$). In our study, T indicates labor induction, which is the current standard of care among pregnant women with hypertension or pre-eclampsia. Let $Y \in \mathbb{R}^p$ denote a single marker or a vector of subject-level candidate measurements (for example, cervical length, Bishop score, blood pressure, and other biochemical markers). Let S be a scalar random variable indicating study itself. Alternatively, in settings with study-level measurements that completely explain between-study heterogeneity in outcome, S may be a vector of study-level measurements. The main goal is to identify a subgroup of subjects for whom treatment can be avoided, because it is unlikely to benefit them; the opposite scenario where not treating is the default can be handled analogously.

3.2. Evaluating marker performance

We measure the performance of a treatment selection marker by the associated change in the expected outcome under marker-based treatment. Let $\phi(Y, S)$ denote a treatment rule that may depend on Y as well as S . Specifically, $T=0$ is recommended if $\phi(Y, S) = 1$; otherwise the default treatment ($T=1$) is recommended. To characterize the performance of this treatment rule, we assess the difference in the expected disease outcome under default treatment versus under marker-based treatment. This measure, denoted by θ , takes a higher

value for a better marker and varies between 0 for a useless marker and $E(D|T=1)$ for a perfect marker [7, 8, 9, 2, 10, 11, 5, 3, 6, 12].

In the context of IPD meta-analysis, we consider two variations on θ . The first, denoted by θ_s , measures the study-specific performance of the treatment rule:

$$\begin{aligned} \theta_s\{\phi(Y, S)\} &= E(D|T=1, S=s) - \{E(D|T=0, \phi(Y, S)=1, S=s)P(\phi(Y, S)=1|S=s) \\ &+ E(D|T=1, \phi(Y, S)=0, S=s)P(\phi(Y, S)=0|S=s)\} = [E\{D|T=1, \phi(Y, S)=1, S=s\} \\ &- E\{D|T=0, \phi(Y, S)=1, S=s\}] \times P\{\phi(Y, S)=1|S=s\} = \beta_s \times \tau_s, \end{aligned}$$

(1)

where β_s denotes the treatment effect among those recommended $T=0$ and τ_s denotes the proportion recommended $T=0$ [5, 12] in study s , respectively. The second measure is an aggregate of study-specific performance measures:

$$\bar{\theta}\{\phi(Y, S)\} = \sum_{s=1}^S w_s \theta_s\{\phi(Y, S)\}, \text{ where } \sum_{s=1}^S w_s = 1, 0 < w_s < 1. \quad (2)$$

Here w_s is a fixed study-specific weight. For example, it may be chosen to correspond to the expected disease outcome in the population from which study s was sampled. In general θ_s and $\bar{\theta}$ are different; however, if θ_s are homogeneous across studies, then $\theta_s = \bar{\theta} = \theta$, where θ denotes the common benefit of the marker-based treatment across studies. Comparing θ_s across studies provides information on differences and similarities in performance across studies. If performance varies considerably across studies, marker performance might be too heterogeneous to be aggregated. In this case, $\bar{\theta}$ still represents a weighted average of study-specific performances, but it is less informative to measure the performance of markers for treatment selection.

3.3. Marker-based treatment rules

Next we consider the form of rules for recommending for or against treatment. Let $(Y, S) \equiv E(D|T=0, Y, S) - E(D|T=1, Y, S)$ denote the marker and study-based treatment effect. One potential rule is to recommend no treatment if $(Y, S) < 0$. This rule can be shown to minimize the expected outcome in the “population” consisting of the pooled populations from which studies $s = 1, \dots, S$ were sampled (equivalently, to maximize θ_s) [40, 9, 41, 3, 6] (see supplementary materials Section 1.1). We call this a “marker and study-based rule” because both the marker value and study are used to determine the recommended treatment. This rule is optimal in the sense that it minimizes the expected disease outcome, or, equivalently, maximizes θ_s . However, the rule is unappealing because it is unclear how to apply it to future subjects who are not associated with any of the studies used to develop the rule.

We consider an alternative rule which is appealing in that it depends on a subject's marker Y , but not study, and thus generalizes to future subjects. It is motivated by the objective of maximizing $\bar{\theta}$, or equivalently, of minimizing the average expected outcome across studies. Specifically, let $D^*(\phi)$ denote the potential outcome for a subject that would be observed had the subject been treated according to rule $\phi \in \{0,1\}$. In Section 1.1 in the supplementary materials, we show that

$$\bar{\phi}(Y) = \arg \max_{\phi} \bar{\theta}\{\phi(Y)\} = \arg \min_{\phi} \sum_{s=1}^S w_s E[D * \{\phi(Y)\} | S = s] = \mathbf{1}\{\bar{\Delta}(Y) \leq 0\}, \quad (3)$$

where $\bar{\Delta}(Y) = \sum_{s=1}^S w_s \Delta(Y, s) f(Y|s)$, $f(Y|s)$ is a conditional density of marker Y given study $S = s$ (study-specific marker distribution), and $\mathbf{1}\{\cdot\}$ is the indicator function. We call $\bar{\phi}(Y)$ a “marker-based rule”. The form of the rule suggests that for a subject with marker value $Y = y$, the treatment recommendation should be based on the treatment effect averaged across studies, where studies are upweighted when y is more common.

3.4. Estimating marker performance

Given IPD from multiple randomized controlled trials comparing $T = 0$ to $T = 1$, consisting of i.i.d. copies $\{(D_i, T_i, Y_i, S_i)\}_{i=1}^n$, we estimate marker performance as follows. Given an

estimated treatment rule $\hat{\phi}$,

$$\hat{\theta}_s(\hat{\phi}) = [\hat{E}\{D|T = 1, \hat{\phi} = 1, S = s\} - \hat{E}\{D|T = 0, \hat{\phi} = 1, S = s\}] \times \frac{\mathbb{P}_n\{\hat{\phi} = 1, S = s\}}{\mathbb{P}_n\{S = s\}} \text{ and}$$

$$\hat{\theta}(\hat{\phi}) = \sum_{s=1}^S w_s \hat{\theta}_s(\hat{\phi}), \text{ where the standard empirical measure } \mathbb{P}_n(A) \equiv \sum_{i=1}^n n^{-1} \mathbf{1}_A\{X_i\} \text{ for a}$$

measurable set A . In our setting where the outcome is binary, we set w_s to the sample proportion of the disease outcome and scale so that $\sum_s w_s = 1$. Confidence intervals for $\hat{\theta}_s$ and $\hat{\theta}$ can be computed numerically using the bootstrap resampling method.

3.5. Estimating marker-based treatment rules

In this study, we estimate the marker-based treatment rule, $\bar{\phi}(Y)$ in (3) using three methods developed for the single-study setting: outcome modeling [5], inverse-probability weighting [3, 42], and augmented inverse-probability weighting [3, 42] methods. We adapt the methods to the IPD setting and these methods are elaborated on in more details below.

3.5.1. Outcome modeling—Outcome modeling is the most obvious approach to estimating a marker-based treatment rule. To find the rule $\bar{\phi}(Y) = \mathbf{1}\{\bar{\Delta}(Y) \leq 0\}$, we estimate the constituents of $\bar{\Delta}(Y)$ by modeling $E(D|T, Y, S)$. Specifically, we pool the IPD from multiple studies and fit a model for $E(D|T, Y, S)$ including marker-by-treatment interactions and study-by-treatment interactions as fixed effects, for example using a generalized linear regression model with a link function g :

$$E(D|T, Y, S) = \mu(T, Y, S; \alpha) = g^{-1}\{\alpha_1 + \alpha_2 S + \alpha_3 Y + \alpha_4 SY + T(\alpha_5 + \alpha_6 Y + \alpha_7 S + \alpha_8 SY)\}.$$

(4)

In its most general form, the model can include a main effect for study; an interaction between study and marker; and a three-way interaction between study, treatment, and marker. In practice, the decision as to whether to include these terms will be based on prior knowledge as well as statistical significance. This fitted model is then used to calculate $\hat{\Delta}(Y, S) = \hat{E}(D|T = 0, Y, S) - \hat{E}(D|T = 1, Y, S)$.

The remaining ingredient for calculating $\bar{\Delta}(Y)$ is $f(Y|S)$. This can be difficult to estimate in practice, particularly for a high-dimensional marker. We consider two approaches. The first uses kernel density estimation to provide estimates of $f(Y|S)$. The second is simpler, but relies on the assumption that the marker distribution is independent of study, in which case $\bar{\Delta}(Y) = \tilde{\Delta}(Y) \equiv \sum_{s=1}^S w_s \Delta(Y, s)$ and $\tilde{\phi}(Y) = \tilde{\phi}(Y) \equiv \mathbf{1}\{\tilde{\Delta}(Y) \leq 0\}$. As shown in Section 1.1 in the supplementary materials, $\tilde{\phi}(Y)$ is suboptimal unless the assumption of a common marker distribution is correct, but is practically appealing in that it does not require estimating the conditional marker distribution.

3.5.2. Inverse probability weighted estimator—Next we consider treatment assignment rules which are derived without specifying a full outcome model. These are based on work by Zhang et al. [42, 3]. Zhang and colleagues proposed to directly search for a rule that minimizes the expected outcome under marker-based treatment, or equivalently, that maximizes the change in expected outcome under marker-based treatment in a single study data setting. They minimize one of two Horwitz-Thompson-type estimators of $E\{D^*(\phi)\}$: an inverse-probability weighted (IPW) estimator or an augmented inverse-probability weighted (AIPW) estimator. We call this “direct searching” (see Laber et al. [43]).

Adapting this to the IPD setting, we define the objective as minimizing the average expected outcome across studies, $\bar{E}\{D^*(\phi)\} \equiv \sum_s w_s E[D^*(\phi(Y)) | s]$. Both IPW and AIPW estimators

of $\{D^*(\phi)\}$ are considered, for a class of simple linear rules of the form $\phi(Y; \eta) = \mathbf{1}\{1 + Y\eta \geq 0\}$. The class of linear rules is written Φ_η and indexed by η . Specifically, the IPW estimator of $E\{D^*(\phi)\}$ is written as

$$IPWE_{IPD}(\eta) = \mathbb{P}_n \phi(Y; \eta) \hat{B}_{IPW}(Y) + \mathbb{P}_n \sum_{s=1}^S w_s \frac{n}{n_s} \mathbf{1}\{S = s\} \frac{T}{\hat{\pi}(Y, S)} D, \quad (5)$$

where $\hat{B}_{IPW}(Y) = \sum_{s=1}^S w_s \frac{n}{n_s} \mathbf{1}\{S=s\} \left\{ \frac{1-T}{1-\hat{\pi}(Y,S)} D - \frac{T}{\hat{\pi}(Y,S)} D \right\}$, $\hat{\pi}(s) = \hat{P}(T=1|S=s)$ is the observed fraction randomized to treatment in study s , n_s denotes the sample size of study s , and $n = \sum_{s=1}^S n_s$. The AIPW estimator of $E\{D^*(\phi)\}$ is given by

$$AIPWE_{IPD}(\eta) = \mathbb{P}_n \phi(Y; \eta) \hat{B}_{AIPW}(Y) + \mathbb{P}_n \sum_{s=1}^S w_s \frac{n}{n_s} \mathbf{1}\{S=s\} \left\{ \frac{T}{\hat{\pi}(Y,S)} D - \frac{T - \hat{\pi}(Y,S)}{\hat{\pi}(Y,S)} \hat{\mu}(1, Y, S) \right\}, \quad (6)$$

where

$$\hat{B}_{AIPW}(Y) = \sum_{s=1}^S w_s \frac{n}{n_s} \mathbf{1}\{S=s\} \left\{ \frac{1-T}{1-\hat{\pi}(Y,S)} D - \frac{T}{\hat{\pi}(Y,S)} D + \frac{T - \hat{\pi}(Y,S)}{1-\hat{\pi}(Y,S)} \hat{\mu}(0, Y, S) + \frac{T - \hat{\pi}(Y,S)}{\hat{\pi}(Y,S)} \hat{\mu}(1, Y, S) \right\}$$

and $\hat{\mu}(1, Y, S) = \mu(1, Y, S; \hat{\alpha})$ by an outcome model with parameters α as defined in (4).

The process of finding the best treatment rules by direct searching therefore consists of two steps as described in detail in supplementary materials, Section 1.2: first, $\hat{B}_{IPW}$ (or $\hat{B}_{AIPW}$) is computed and plugged in to (5) (or (6)). Second, the rule in class Φ_η that minimizes (5) (or (6)), written $\bar{\phi}_{IPW}(Y; \eta) = \arg \min_{\phi \in \Phi_\eta} IPWE_{IPD}(\eta)$ (or $\bar{\phi}_{AIPW}(Y; \eta) = \arg \min_{\phi \in \Phi_\eta} AIPWE_{IPD}(\eta)$), is identified.

In some context treatment rules may be pre-specified by clinicians or by previous studies and the task is only evaluation of the rule. Our framework described in Section 3.2 accommodate evaluation of previously developed treatment rules.

4. Simulation study

4.1. Simulation set-up

A simulation study is designed to compare the performance of different treatment rules, using the two performance measures θ_s and $\bar{\theta}$, under various scenarios. Multiple sets of IPD consisting of two individual studies were simulated under five different scenarios as shown in Table 1. In Scenario 1, observations for the two studies are sampled from the same population and thus differences are due to sampling variability. Scenarios 2-5 reflect situations where different sets of parameters differ across studies while marker performance is reasonably similar (that is, the markers generate reasonably similar treatment rules). The marginal expected disease outcome, marginal treatment effect ($E(D|T=0) - E(D|T=1)$), main effects of the markers (both continuous), strength of marker-by-treatment interactions, and/or marker distributions, differ between studies. Sample sizes and expected disease outcomes are similar to those in the HYPITAT trials. For each scenario, we generated 1,000

training data sets consisting of two studies, each with a sample size of 700. We evaluate the estimated treatment rules using 1,000 independent test data sets with the sample size of 10^5 . Treatment rules are estimated using outcome modeling with logistic regression and maximum likelihood estimation, and by direct search with IPW or augmented IPWE (AIPWE) and linear treatment rules. The study-specific marker distribution is estimated using a product of smooth kernels, a method suitable given a mixture of continuous, nominal, and ordinal markers [44, 45]. This method is implemented in the R package np [46]. For continuous markers, we use the Gaussian kernel. For nominal markers, we use the kernel function introduced by Aitchison et al. [47]. For ordinal markers, the kernel function introduced by Wang et al. [48] is used. See Section 1.3 of the supplementary materials for more details. To aggregate across studies, the study-specific weight w_s is the sample proportion of the study-specific disease outcome.

We compare the following five treatment rules:

Rule 1. Rules that are developed in (and optimized for) Study 1 only: $\phi(Y, S=1) = \mathbf{1}\{ \Delta(Y, S=1) > 0 \}$,

Rule 2. Rules that are developed in (and optimized for) Study 2 only: $\phi(Y, S=2) = \mathbf{1}\{ \Delta(Y, S=2) > 0 \}$,

Rule 3. Rules that ignore study: $\phi(Y) = \mathbf{1}\{ \Delta(Y) > 0 \}$,

Rule 4. Rules that are based on the average treatment effect across studies: $\tilde{\phi}(Y) = \mathbf{1}\{ \tilde{\Delta}(Y) \leq 0 \} = \mathbf{1}\{ \sum_s w_s \Delta(Y, S=s) \leq 0 \}$,

Rule 5. Rules that are optimal for minimizing the average expected outcome across studies: $\bar{\phi}(Y) = \mathbf{1}\{ \bar{\Delta}(Y) \leq 0 \} = \mathbf{1}\{ \sum_s w_s \Delta(Y, s) f(Y|s) \leq 0 \}$.

The true performance of the rule of each form is estimated empirically by averaging estimates of θ_s and $\bar{\theta}$ across 100 large data sets ($n = 10^7$), each simulated under the true outcome model.

4.2. Results of the simulation study

4.2.1. When individual studies are sampled from the same population

(Scenario 1)—Since the IPD consist of two individual studies that are sampled from the same population, all five treatment rules are equivalent and θ_s and $\bar{\theta}$ are equivalent. Therefore, all five methods produce similar results. Outcome modeling produces more accurate estimates of θ_s and $\bar{\theta}$ than both inverse probability weighted estimators, and smaller standard deviations; this is to be expected given that the “working” outcome model is correctly specified. Figure 1 of the supplementary materials provides the results of Scenario 1.

4.2.2. When individual studies are sampled from different populations

(Scenarios 2-5)—In Scenario 2, two populations have different expected disease outcomes and marginal treatment effects, while the other features of the outcome model are constant. The five treatment rules perform similarly as shown in Figure 2 of the supplementary materials. This result suggests that in our scenarios a moderate difference in the marginal

expected outcome and treatment effect between studies has minimal consequence when evaluating a marker for treatment selection.

In Scenarios 3-4, the two populations differ with regard to marker main effect and marker-by-treatment interactions, and this results in heterogeneity in marker performance between studies. The results of Scenario 3 are summarized in Figure 3 of the supplementary materials, and are similar to Scenario 4 which we summarize here (see Figure 1). Although the optimal treatment rule with respect to study-averaged performance as measured by $\bar{\theta}$ is Rule 5, Rules 3 or 4 are found to perform similarly. The highest estimates of study-specific performance as measured by θ_1 and θ_2 are achieved by Rule 1 and Rule 2, respectively, estimated with outcome modeling. Estimates of $\bar{\theta}$ for Rule 1 or Rule 2 are smaller than those obtained by for Rules 3, 4, or 5. These results suggest that the data from the two studies should be pooled to develop rules with good study-averaged performance, but that ignoring study effects does not worsen performance in this scenario.

Figure 2 summarizes the results of Scenario 5, in which two studies are sampled from two populations that differ with respect to all features of interest. Due to the differences between the two studies, we see that the treatment rules estimated from one study alone produce poor estimates of the treatment effect for subjects belonging to the other study (Rules 1 and 2), particularly when a study that has weak predictive markers (Study 2) is used for developing treatment rules. With regard to study-averaged performance, we observe that Rule 3, which ignores differences between studies, performs slightly worse than Rules 4 and 5, particularly when using outcome modeling. Rule 4, which ignores differences in marker distributions between studies, performs slightly worse than Rule 5, particularly when using outcome modeling. Thus, in this scenario, ignoring study effects does yield rules with inferior study-averaged performance.

In summary, we find that marker performance for treatment selection is not significantly influenced by marginal expected disease outcomes or marginal treatment effects differing between studies, while it depends heavily on differences in marker-by-treatment interaction coefficients and in marker distributions. In the presence of moderate-to-large heterogeneity in marker main effects, marker-by-treatment interactions, and marker distributions across studies, treatment rules that pool data across studies and that capture study effects (Rules 4 and 5) have better study-averaged performance than those developed using an individual study or that ignore heterogeneity between studies. In contrast, performance in each study is optimized with rules developed using data from that study only. In our simulations, because the working outcome models are correctly specified, we find that outcome modeling produces better-performing rules than direct searching (average performance and variability in performance). As expected, direct searching using an AIPW estimator of $\{D^*(\phi)\}$, rather than an IPW estimator, results in increased efficiency [49, 3]. More simulation results, including a simulation with ten individual studies, are presented in Section 2 of the supplementary materials.

We do not consider scenarios for extremely heterogeneous individual studies such as the marker-by-treatment interactions effects or the marginal treatment effects are opposite signs between two studies. In the presence of such extreme heterogeneity among individual

studies, the optimal treatment rules among studies are significantly different and there would be no benefit to evaluate treatment selection markers by combining over different studies.

5. Data analysis

5.1. Marker selection and comparison methods

Here we apply the proposed IPD analysis methods to the high risk pregnancy study. HYPITAT I and HYPITAT II are different in several respects. Women in HYPITAT I are at higher gestational age (median = 39 weeks, range = 36 – 41 weeks) than women in HYPITAT II (median = 36 weeks, range = 34 – 37 weeks). A higher proportion of women with pregnancy-induced hypertension vs. pre-eclampsia enrolled in HYPITAT I (68% and 32%) vs. in HYPITAT II (27% and 44%). There is evidence that the treatment effect differs by cervical length in HYPITAT I ($p = 0.01$ for interaction) but not in HYPITAT II ($p = 0.5$ for interaction). Similarly, the treatment effect differs by gestational age in HYPITAT II ($p = 0.01$ for interaction), but not in HYPITAT I ($p = 0.16$ for interaction).

Our primary outcome is a composite of maternal mortality, severe maternal morbidity (eclampsia, HELLP syndrome, pulmonary edema, thromboembolic disease, placental abruption, or major post-partum hemorrhage), and signs of progression to severe disease (systolic blood pressure ≥ 170 mmHg or diastolic blood pressure ≥ 110 mmHg, or proteinuria ≥ 5 g per 24 hours). This was the original primary outcome of HYPITAT I, and is used here to permit a combined analysis of the two studies; the original HYPITAT II outcome was somewhat different (see Supplementary Table 1 and Section 3.3 of the supplementary materials for more details). The numbers of primary events were 283 (37.4%; 43.8% for expectant monitoring and 31% for labor induction) and 302 (46.6%; 52.7% for expectant management and 40.7% for labor induction) in HYPITAT I and II, respectively.

We consider the following markers measured at baseline: cervical length, Bishop score and its constituents (cervical dilation, cervical effacement, cervical consistency, cervical position, and cervical engagement), maternal age (year), diagnosis (pregnancy-induced hypertension/pre-eclampsia/chronic hypertension), gestational age (in weeks), BMI ($\text{weight}/(\text{height}/100)^2$), and systolic and diastolic blood pressures. We select markers whose main effect or interaction with treatment, study, or study-by-treatment interaction are selected in a logistic regression models of the form (4) with Least Absolute Shrinkage and Selection Operator (LASSO [50]) penalty. We apply the LASSO approach 100 times, varying the random seed for the internal 5-fold cross-validation. The internal 5-fold cross-validation returns an estimate of the penalty parameter by minimizing the cross-validated mean squared error. In the first step, we screen out markers whose main effects or interactions are never selected over the first 100 replications. In the second step, among remaining main and interaction terms of markers, we select the markers whose main effects or treatment interactions are among the top 10 variables selected (in at least 88 out of 100 replications) by Lasso. These variables are: the main effects of chronic hypertension, cervical effacement, and cervical dilation; the two-way interaction terms between study and each of $(\text{Gestational Age} - 15)^2$ and pre-eclampsia; the two-way interaction terms between treatment and each of $(\text{Gestational Age} - 15)^2$, hodge 2 level of cervical engagement, cervical length, and pre-eclampsia; and the three-way interaction term between treatment and study and the median

level of cervical position. When a higher order interaction term is selected, we also include main and all possible lower order terms underlying it. We estimate treatment rules using the selected markers (the number of regression parameters of 21 including an intercept term), using outcome modeling and direct search.

The five treatment rules that we introduced in Section 4 are compared. For outcome modeling, we use logistic regression with maximum likelihood estimation. For direct search, we minimize the AIPW estimator and search for the optimal linear rule. Rules are evaluated using θ_1 , θ_2 , and $\bar{\theta}$. Apparent performance estimates are obtained using the original (training) data. In order to reduce bias due to overoptimism associated with screening markers and fitting and evaluating the estimation models on the same data, we also calculate bootstrap bias-corrected estimates of performance, conditional on the markers selected in the original data, using the refined bootstrap method [51] (see Section 3.1 of the supplementary materials). Confidence intervals are calculated using the percentile bootstrap method.

5.2. Analysis results

Results are shown in Table 2. Pooling data across the two HYPITAT studies, in a fashion that acknowledges the differences between studies, appears beneficial. Using the bias-corrected estimates of marker performance, Rule 4, which minimizes the AIPW estimator of study-averaged expected outcome but which assumes a common marker distribution across studies, performs better than other rules. Employing Rule 4 is estimated to decrease the risk of progression to severe maternal disease by 0.6% compared with the default treatment rule that treats all women with labor induction; 22.2% and 4.7% of women in HYPITAT I and HYPITAT II are recommended for expectant management, respectively (see Figure 3). Rule 5, yields the highest apparent measure of study-averaged marker performance, but negative bias-corrected estimates of marker performance, which indicates a high level of overoptimism.

Treatment rules derived in the two individual studies produce good performance in that study, but poorer study-averaged performance. It appears that Rule 2, based on HYPITAT II, fails to estimate the best treatment for women in HYPITAT I. This rule also yields considerably different classification of woman according to treatment benefit based as compared to other rules (see the supplementary materials Section 3.4 and Tables 3 and 4).

We explore the impact of treatment rules selected to minimize the risk of progression to severe maternal disease on secondary outcomes. Secondary outcomes are adverse neonatal outcomes (5-minute Apgar score < 7, umbilical artery pH < 7.05, or admission to a neonatal intensive-care unit) and caesarean delivery. Risks of secondary outcomes under marker-based rules are similar to those under labor induction. More details are given in Section 3.5 and Table 5 of the supplementary materials.

While the benefit of the markers is small, as measured by θ , the identification of a subgroup who can avoid labor induction may still have value. In other words, a marker-based treatment rule that avoids labor induction in a subgroup, without meaningfully impacting the risk of progression to severe maternal disease, will still avoid the cost and risk associated with labor induction for the selected subgroup of women. However it appears that the size of

this subgroup differs between the two HYPITAT trials. For example, using Rule 4 estimated using direct search, 138 out of 756 women (22.2%) in HYPITAT I are recommended to avoid labor induction while only 32 out of 648 women (4.7%) in HYPITAT II are recommended to avoid labor induction. Our results suggest that different distributions of markers, particularly hodge 2 engagement status, may explain the different subgroup sizes (see Section 3.6 and Figure 5 of the supplementary materials).

The application of the proposed IPD analysis method to the high risk pregnancy study identified a significant interaction effect between a marker indicating history of pre-eclampsia and treatment, which was not identified by a single-study analysis. The IPD analysis also allows us to evaluate treatment selection markers and to derive treatment rules that are applicable to high-risk pregnant women over a wider range of gestational ages than would be possible with single-study analysis.

The results of this data analysis suggest that cervical length, Bishop score or constituents of the Bishop score, and currently available maternal characteristics, despite being strong prognostic markers (that is, predict the risk of maternal complications), are relatively weak predictors of the treatment effect on the risk of progression to severe maternal disease. The bias-corrected estimates of marker performance are very small with wide confidence intervals containing zero. More research is warranted to identify biomarkers that are not available in the current HYPITAT trails yet have potential to predict the treatment effect better than cervical length, Bishop score, or constituents of the Bishop score (or by combining with these markers) for pregnant women with pregnancy induced hypertension or mild pre-eclampsia.

6. Discussion

We proposed a statistical framework for evaluating treatment selection markers, given individual participant data from multiple randomized controlled trials. We described methods for discovering and evaluating marker-based treatment rules designed to be optimal in a study-averaged sense, as compared to rules optimized within each study or that ignore study differences. Importantly, the rules do not depend on study and thus can be used to determine the best treatment for future subjects based on their marker value. Based on the simulation results, we recommend that marker-based treatment rules based on the treatment effect averaged across studies (such as $\bar{\phi}$ or $\tilde{\phi}$) be considered to optimize the study-averaged performance measurement ($\bar{\theta}$). Due to the difficulty in reliable estimation of the conditional density, it may be often more practical to use $\tilde{\phi}$ for high-dimensional markers. Also, in addition to the study-averaged performance measurement, we recommend one computes the proportion of participants recommended the non-default treatment (τ_s) and the magnitude of the treatment effect (β_s) in this subgroup, for each study, to investigate sources of heterogeneity across studies.

This work goes beyond previous works on biomarkers for treatment selection in the IPD setting, which are based on hypothesis tests and estimation of marker-by-treatment interaction or subgroup analysis [25, 26, 27, 28, 29, 30, 31]. We derive treatment rules that

optimize study-averaged performance, and compare them to rules optimized within individual studies.

Our method proposed in current article has its limitations which suggest several avenues for future research. In this study, we fit fixed effect models, which are better-justified than random effects models in settings with a low number of individual studies [52]. In settings with a large number of individual studies, random effects models deserve investigation. Also, in the presence of heterogeneity across studies, investigation of the source of the heterogeneity would be important and might help in interpreting, and in increasing the generalizability of, the results derived from multiple studies. A limitation of our approach is that differences in study performance are not sufficient to explain or diagnose reasons for the differences. Exploratory analysis, along the lines of what was done in our data example, can be used to help understand why the differences occur. Other methods for better-characterizing the sources of between-study heterogeneity that may affect marker performance deserve investigation. We have focused on the “gold standard” setting of combining data from multiple randomized and controlled trials. Extension of the methods to accommodate combining observational studies deserve consideration. We note that the IPW and AIPW methods for deriving treatment rules can be applied to single observational studies, but need extension to accommodate combining multiple studies. Also, our method has been developed only for evaluating treatment rules in the population from which trial participants were sampled. Methods for generalizing to larger populations deserve investigation. A final issue, very common in statistics and not specific to our problem, is doing proper inference after learning from the data. Ideally, a separate study or studies would be available to validate the treatment rules derived using the multiple trials.

The clinical interpretation of our results for the high risk pregnancy study is constrained by our choice of outcome, which includes not only severe maternal outcomes but also some signs of progression to severe disease that does not necessarily require an intervention. Future efforts to identify a clinically applicable subgroups of women who could benefit more from labor induction or expectant management using from the IPD of HYPITAT trials would ideally use a composite of maternal mortality and severe maternal morbidity as well as simultaneous modeling of maternal outcomes and neonatal outcomes.

This research raises an interesting question as to when (and how much) information is gained by combining IPD from multiple studies, relative to individual studies. There is no definite criteria to judge how similar is similar enough, and existing approaches to assess heterogeneity across studies (for example, I^2 test [53]) are not appropriate when a small number of individual studies is combined [54]. The two HYPITAT trials are considered similar in clinical (for example, treatment and primary outcome) and methodological (for example, study design) aspects, and previous individual analyses of two HYPITAT trials show consistent result on greater marginal effect of labor induction over expectant monitoring. We do not assess heterogeneity, and this deserves future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by the National Institutes of Health (NIH) R01CA152089. Parvin Tajik is supported by AXA Research Fund.

References

1. McKeague I, Qian M. Evaluation of treatment policies via sparse functional linear regression 2011. Preprint.
2. Qian M, Murphy S. Performance guarantees for individualized treatment rules. *Annals of Statistics*. 2011; 39(2):1180. [PubMed: 21666835]
3. Zhang B, Tsiatis A, Laber E, Davidian M. A robust method for estimating optimal treatment regimes. *Biometrics*. 2012; 68(4):1010–1018. [PubMed: 22550953]
4. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*. 2012; 107(499):1106–1118. [PubMed: 23630406]
5. Janes H, Brown MD, Huang Y, Pepe MS. An approach to evaluating and comparing biomarkers for patient treatment selection. *The International Journal of Biostatistics*. 2014; 10(1):99–121. [PubMed: 24695044]
6. Kang C, Janes H, Huang Y. Combining biomarkers to optimize patient treatment recommendations. *Biometrics*. 2014; 70(3):713–716. [PubMed: 24889265]
7. Song X, Pepe M. Evaluating markers for selecting a patient's treatment. *Biometrics*. 2004; 60(4): 874–883. [PubMed: 15606407]
8. Gunter L, Zhu J, Murphy S. Variable selection for qualitative interactions in personalized medicine while controlling the family-wise error rate. *Journal of Biopharmaceutical Statistics*. 2011; 21(6): 1063–1078. [PubMed: 22023676]
9. Brinkley J, Tsiatis A, Anstrom K. A generalized estimator of the attributable benefit of an optimal treatment regime. *Biometrics*. 2010; 66(2):512–522. [PubMed: 19508237]
10. Janes H, Pepe M, Bossuyt P, Barlow W. Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine*. 2011; 154(4):253–259. [PubMed: 21320940]
11. Janes H, Pepe MS, Huang Y. A framework for evaluating markers used to select patient treatment. *Medical Decision Making*. 2014; 34(2):159–167. [PubMed: 23811760]
12. Janes H, Brown MD, Pepe MS. Designing a study to evaluate the benefit of a biomarker for selecting patient treatment. *Statistics in Medicine*. 2015; 34(27):3503–3515. [PubMed: 26112650]
13. Koopmans CM, Bijlenga D, Groen H, Vijgen S, Aarnoudse JG, Bekedam DJ, van den Berg PP, de Boer K, Burggraaff JM, Bloemenkamp KW, et al. Induction of labour versus expectant monitoring for gestational hypertension or mild pre-eclampsia after 36 weeks' gestation (HYPITAT): a multicentre, open-label randomised controlled trial. *The Lancet*. 2009; 374(9694):979–988.
14. Noordzij JP, Lee SL, Bernet VJ, Payne RJ, Cohen SM, McLeod IK, Hier MP, Black MJ, Kerr PD, Richards ML, et al. Early prediction of hypocalcemia after thyroidectomy using parathyroid hormone: an analysis of pooled individual patient data from nine observational studies. *Journal of the American College of Surgeons*. 2007; 205(6):748–754. [PubMed: 18035257]
15. Xu J, Ye Y, Zhang H, Szmitkowski M, Mäkinen M, Li P, Xia D, Yang J, Wu Y, Wu H. Diagnostic and prognostic value of serum interleukin-6 in colorectal cancer. *Medicine*. 2016; 95(2):e2502. [PubMed: 26765465]
16. Sylvester RJ, van der Meijden AP, Oosterlinck W, Witjes JA, Bouffieux C, Denis L, Newling DW, Kurth K. Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. *European Urology*. 2006; 49(3):466–477. [PubMed: 16442208]
17. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JDF, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLOS Medicine*. 2008; 5(8):e165. [PubMed: 18684008]

18. Pennells L, Kaptoge S, White IR, Thompson SG, Wood AM, Tipping RW, Folsom AR, Couper DJ, Ballantyne CM, Coresh J, et al. Assessing risk prediction models using individual participant data from multiple studies. *American Journal of Epidemiology*. 2014; 179(5):621–632. [PubMed: 24366051]
19. Clark GM. Prognostic factors versus predictive factors: examples from a clinical trial of erlotinib. *Molecular oncology*. 2008; 1(4):406–412. [PubMed: 19383314]
20. Clark G. Prognostic and predictive factors for breast cancer. *Breast Cancer*. 1995; 2(2):79–89. [PubMed: 11091537]
21. Henry N, Hayes D. Uses and abuses of tumor markers in the diagnosis, monitoring, and treatment of primary and metastatic breast cancer. *The Oncologist*. 2006; 11(6):541–552. [PubMed: 16794234]
22. Simon R. Lost in translation: problems and pitfalls in translating laboratory observations to clinical utility. *European Journal of Cancer*. 2008; 44(18):2707–2713. [PubMed: 18977655]
23. Ballman KV. Biomarker: predictive or prognostic? *Journal of Clinical Oncology*. 2015; 33(33):3968–3971. [PubMed: 26392104]
24. Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology*. 2002; 55(1):86–94. [PubMed: 11781126]
25. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in Medicine*. 2002; 21(3):371–387. [PubMed: 11813224]
26. Teramukai S, Matsuyama Y, Mizuno S, Sakamoto J. Individual patient-level and study-level meta-analysis for investigating modifiers of treatment effect. *Japanese Journal of Clinical Oncology*. 2004; 34(12):717–721. [PubMed: 15640501]
27. Thompson SG, Higgins J. Can meta-analysis help target interventions at individuals most likely to benefit? *The Lancet*. 2005; 365(9456):341–346.
28. Simmonds M, Higgins J. Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. *Statistics in Medicine*. 2007; 26(15):2982–2999. [PubMed: 17195960]
29. Hayden JA, Cartwright JL, Riley RD, Vantulder MW, et al. Exercise therapy for chronic low back pain: protocol for an individual participant data meta-analysis. *Systematic Reviews*. 2012; 1(1):64. [PubMed: 23259855]
30. Stewart GB, Altman DG, Askie LM, Duley L, Simmonds MC, Stewart LA. Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. *PLOS One*. 2012; 7(10):e46 042.
31. Jansen JP, Cope S. Meta-regression models to address heterogeneity and inconsistency in network meta-analysis of survival outcomes. *BMC Medical Research Methodology*. 2012; 12(1):152. [PubMed: 23043545]
32. Huang Y, Gilbert P, Janes H. Assessing treatment-selection markers using a potential outcomes framework. *Biometrics*. 2012; 68(3):687–696. [PubMed: 22299708]
33. Lazar AA, Cole BF, Bonetti M, Gelber RD. Evaluation of treatment-effect heterogeneity using biomarkers measured on a continuous scale: subpopulation treatment effect pattern plot. *Journal of Clinical Oncology*. 2010; 28(29):4539–4544. [PubMed: 20837942]
34. Follmann DA, Proschan MA. A multivariate test of interaction for use in clinical trials. *Biometrics*. 1999; 55(4):1151–1155. [PubMed: 11315061]
35. Kovalchik SA, Varadhan R, Weiss CO. Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. *Statistics in Medicine*. 2013; 32(28):4906–4923. [PubMed: 23788362]
36. Langenveld J, Broekhuijsen K, van Baaren GJ, van Pampus M, van Kaam A, Groen H, Porath M, Oudijk M, Bloemenkamp K, Groot C, et al. Induction of labour versus expectant monitoring for gestational hypertension or mild pre-eclampsia between 34 and 37 weeks' gestation (HYPITAT-II): a multicentre, open-label randomised controlled trial. *BMC Pregnancy and Childbirth*. 2011; 11(1):50. [PubMed: 21736705]

37. Broekhuijsen K, van Baaren GJ, van Pampus MG, Ganzevoort W, Sikkema JM, Woiski MD, Oudijk MA, Bloemenkamp KW, Scheepers HC, Bremer HA, et al. Immediate delivery versus expectant monitoring for hypertensive disorders of pregnancy between 34 and 37 weeks of gestation (HYPITAT-II): an open-label, randomised controlled trial. *The Lancet*. 2015; 385(9986): 2492–2501.
38. BISHOP EH. Pelvic scoring for elective induction. *Obstetrics & Gynecology*. 1964; 24(2):266–268. [PubMed: 14199536]
39. Tajik P, van der Tuuk K, Koopmans C, Groen H, van Pampus M, van der Berg P, van der Post J, van Loon A, de Groot C, Kwee A, et al. Should cervical favourability play a role in the decision for labour induction in gestational hypertension or mild pre-eclampsia at term? An exploratory analysis of the HYPITAT trial. *BJOG: An International Journal of Obstetrics & Gynaecology*. 2012; 119(9):1123–1130. [PubMed: 22703475]
40. Vickers A, Kattan M, Sargent D. Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials*. 2007; 8(1):14. [PubMed: 17550609]
41. Lu W, Zhang H, Zeng D. Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*. 2011; (Online print). doi: 10.1177/0962280211428383
42. Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. Estimating optimal treatment regimes from a classification perspective. *Stat*. 2012; 1(1):103–114. [PubMed: 23645940]
43. Laber EB, Tsiatis AA, Davidian M, Holloway ST. Discussion of combining biomarkers to optimize patient treatment recommendation. *Biometrics*. 2014; 70(3):713–716. [PubMed: 24889265]
44. Ahmad IA, Cerrito PB. Nonparametric estimation of joint discrete-continuous probability densities with applications. *Journal of Statistical Planning and Inference*. 1994; 41(3):349–364.
45. Li Q, Racine J. Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*. 2003; 86(2):266–292.
46. Hayfield T, Racine JS. Nonparametric econometrics: The np package. *Journal of Statistical Software*. 2008; 27(5) URL <http://www.jstatsoft.org/v27/i05/>.
47. Aitchison J, Aitken CG. Multivariate binary discrimination by the kernel method. *Biometrika*. 1976; 63(3):413–420.
48. Wang MC, Van Ryzin J. A class of smooth estimators for discrete distributions. *Biometrika*. 1981; 68(1):301–309.
49. Robins J, Rotnitzky A, Zhao L. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89(427):846–866.
50. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 1996:267–288.
51. Efron, B., Tibshirani, RJ. An introduction to the bootstrap. CRC press; 1994.
52. Riley RD. Commentary: like it and lump it? meta-analysis using individual participant data. *International Journal of Epidemiology*. 2010; 39(5):1359–1361. [PubMed: 20660642]
53. Higgin J, Thompson S, Deeks J, Altman D. Measuring inconsistency in meta-analysis. *British Medical Journal*. 2003; 327(7414):557–560. [PubMed: 12958120]
54. Zlowodzki M, Poolman RW, Kerkhoffs GM, Tornetta P III, Bhandari M, Group IE BOSW. et al. How to interpret a meta-analysis and judge its value as a guide for clinical practice. *Acta Orthopaedica*. 2007; 78(5):598–609. [PubMed: 17966018]

Figure 1.

Results of the simulation study under simulation Scenario 4: two studies sampled from different populations in terms of marker-by-treatment effect, disease prevalence, and marker distribution. Training sample sizes for the two studies were $n_1 = n_2 = 700$. Five treatment rules were considered: Rule 1 is optimized for study 1; Rule 2 is optimized for study 2; Rule 3 ignores studies; Rule 4 is based on the average treatment effect across studies; and Rule 5 is optimal for minimizing expected outcomes averaged across studies. For each type of rule, results of three estimating methods are presented: 1) outcome modeling using logistic regression with maximum likelihood estimation (OM) 2) minimizing the inverse probability weighted estimator of the study-averaged expected outcome (IPW) and 3) minimizing the augmented inverse probability weighted estimator of the study-averaged expected outcome (AIPW). Each treatment rule was evaluated with respect to its performance in each study (θ_s) and the study-averaged performance ($\bar{\theta}$). Mean and standard deviation of performance measures were computed using 1,000 large ($N=10^5$) test data sets. Points represent means and error bars show two standard deviations from the mean. The performance measurement of the true treatment rule of each form is denoted by gray lines.

Figure 2.

Result of the simulation study under simulation Scenario 5: two studies sampled from different populations in terms of the strength of main effect, marker-by-treatment interactions, marker-by-treatment effect, marker-by-treatment effect, disease prevalence, and marker distribution. Training sample sizes for the two studies were $n_1 = n_2 = 700$. Five treatment rules were considered: Rule 1 is optimized for study; Rule 2 is optimized for study 2; Rule 3 ignores studies; Rule 4 is based on the average treatment effect across studies; and Rule 5 is optimal for minimizing expected outcomes averaged across studies. For each type of rule, results of three estimating methods are presented: 1) outcome modeling using logistic regression with maximum likelihood estimation (OM) 2) minimizing the inverse probability weighted estimator of the study-averaged expected outcome (IPW) and 3) minimizing the augmented inverse probability weighted estimator of the study-averaged expected outcome (AIPW). Each treatment rule was evaluated with respect to its performance in each study (θ_s) and the study-averaged performance ($\bar{\theta}$). Mean and standard deviation of performance measures were computed using 1,000 large ($N=10^5$) test data sets. Points represent means and error bars show two standard deviations from the mean. The performance measurement of the true treatment rule of each form is denoted by gray lines.

Figure 3.

Estimated distribution of marker-specific treatment effect for the high risk pregnancy study.

The treatment effect is defined as $\tilde{\Delta}(Y) = \sum_{s=1}^S w_s \Delta(Y, s)$, where (Y, S) is the difference in risk of progression to severe maternal disease given markers in study s , and w_s is a weight for study s . The treatment effect is estimated using logistic regression. The plot is the inverse of a cumulative distribution function; the y-axis shows each treatment effect value and the x-axis shows the percent below each value.

Scenario	Study 1	Study 2	
Two studies were sampled from the same population.	1	True risk model	$\text{logit } RD = 1 T; Y = -0.1 + 0.3 Y_1 - 0.3 Y_2 + \pi(-0.6 - 1.3 Y_1 + 0.5 Y_2)$
		$RD = 1 T = 0$	0.476
		$RD = 1 T = 1$	0.360
		True θ_s	0.071
	2	True risk model	$\text{logit } RD = 1 T; Y = -0.5 + 0.1 Y_1 + 0.3 Y_2 + \pi(-0.1 - 0.8 Y_1 - 0.4 Y_2)$
		$RD = 1 T = 0$	0.380
		$RD = 1 T = 1$	0.368
		True θ_s	0.073
	3	True risk model	$\text{logit } RD = 1 T; Y = -0.125 + 0.1 Y_1 + 0.3 Y_2 + \pi(-0.55 - 1.2 Y_1 - 0.3 Y_2)$
		$RD = 1 T = 0$	0.470
$RD = 1 T = 1$		0.368	
True θ_s		0.061	
Two studies were sampled from different populations.	4	True risk model	$\text{logit } RD = 1 T; Y = -0.4 + 0.1 Y_1 + 0.4 Y_2 + \pi(-0.6 - 1.6 Y_1 - 0.3 Y_2)$
		$RD = 1 T = 0$	0.415
		$RD = 1 T = 1$	0.337
		True θ_s	0.087
	5	True risk model	$\text{logit } RD = 1 T; Y = -0.35 + 0.1 Y_1 + 0.3 Y_2 + \pi(-0.6 - 1.6 Y_1 - 0.3 Y_2)$
		$RD = 1 T = 0$	0.404
		$RD = 1 T = 1$	0.345
		True θ_s	0.095

1. D is a binary disease outcome; T indicates a treatment (0 or 1); and $Y = (Y_1, Y_2)$ is a marker of interest.
2. For Scenarios 1-4, Y_1 and Y_2 are independent $N(0,1)$ in both studies.
3. For Scenario 5, Y_1 and Y_2 are independent $N(0,1)$ in study 1; $N(1.2, 2^2)$ in study 2.
4. Here the true marker performance measurement in each study, denoted by θ_S , is computed using Rule 1 for study 1 and Rule 2 for study 2, separately.

Table 2

Results of the IPD analysis for the high-risk pregnancy study. Five sets of treatment rules were considered: Rule 1 uses HYPITAT I for training; Rule 2 uses HYPITAT II for training; Rule 3 pools data across studies but ignores study effects; Rule 4 is based on the average treatment effect across studies; and Rule 5 is optimal for minimizing expected outcomes averaged across studies. For each rule, two methods were used for estimation: 1) outcome modeling using logistic regression with maximum likelihood estimation (outcome modeling) and 2) direct search by minimizing augmented IPWE (AIPWE) of study-averaged expected outcome. Each treatment rule was evaluated with respect to its performance in each study (θ_s) and with respect to study-averaged performance ($\bar{\theta}$). θ_s decomposes into β_s and τ_s . For Rule 5, the study-specific marker distribution was estimated using the Gaussian kernel. For each method the apparent model performance (Apparent), bootstrap-bias-corrected estimated model performance (Bias-corrected), and bootstrap percentile based 95% confidence intervals (95% CI) are reported.

Rule	Estimation method	Study-specific performance					Study-averaged performance	
		HIPITAT I		HIPITAT II				
		θ_1	β_1	τ_1	θ_2	β_2	τ_2	$\bar{\theta}$
Rule 1	Outcome modeling	Apparent	0.014	0.066	0.206	0.003	0.035	0.073
		Bias-corrected	0.003	0.034	0.207	0.002	0.020	0.073
		95% CI	(-0.021, 0.043)	(-0.105, 0.218)	(0.034, 0.374)	(-0.089, 0.027)	(-0.448, 0.415)	(0.009, 0.443)
	AIPWE	Apparent	0.031	0.244	0.128	-0.018	-0.293	0.062
		Bias-corrected	0.010	0.118	0.128	-0.018	-0.287	0.062
		95% CI	(-0.010, 0.056)	(-0.011, 0.432)	(0.041, 0.377)	(-0.064, 0.008)	(-0.731, 0.192)	(0.014, 0.290)
	Outcome modeling	Apparent	-0.130	-0.133	0.981	0.026	0.125	0.210
		Bias-corrected	-0.130	-0.132	0.982	0.011	0.073	0.210
		95% CI	(-0.196, -0.052)	(-0.207, -0.053)	(0.802, 0.998)	(-0.018, 0.056)	(-0.062, 0.245)	(0.059, 0.419)
Rule 2	AIPWE	Apparent	0	0	0	0	0	0
		Bias-corrected	0.000	-0.007	0	-0.027	-0.143	0.000
		95% CI	(-0.171, 0.000)	(-0.268, -0.007)	(0.000, 0.996)	(-0.027, 0.060)	(-0.143, 0.513)	(0.000, 0.576)
	Outcome modeling	Apparent	0.005	0.005	0.005	0.010	0.010	0.010
		Bias-corrected	-0.003	0	0.003	0.005	0.005	0.005
		95% CI	(-0.022, 0.039)	(-0.022, 0.039)	(0.001, 0.044)	(-0.026, 0.030)	(-0.026, 0.030)	(0, 0.031)
	AIPWE	Apparent	0.030	0.156	0.194	0.001	0.036	0.023
Rule 3	AIPWE	Apparent	0.008					0.008
		Bias-corrected						0.001
		95% CI						(-0.014, 0.024)

Rule	Estimation method	Study-specific performance					Study-averaged performance	
		HIPTAT I		HIPTAT II			τ_2	$\bar{\theta}$
		θ_1	β_1	τ_1	θ_2	β_2		
Rule 4	Bias-corrected	0.016	0.057	0.195	-0.009	-0.180	0.024	0.003
	95% CI	(-0.014, 0.052)	(-0.099, 0.492)	(0.000, 0.385)	(-0.027, 0.033)	(-0.691, 0.477)	(0.001, 0.203)	(-0.012, 0.031)
	Apparent	0.010	0.036	0.274	0.006	0.150	0.040	0.008
	Bias-corrected	-0.002	0.009	0.272	0.003	0.102	0.040	0.001
	95% CI	(-0.053, 0.035)	(-0.194, 0.265)	(0.025, 0.735)	(-0.012, 0.037)	(-0.548, 0.752)	(0, 0.108)	(-0.023, 0.027)
	Apparent	0.041	0.183	0.222	0.016	0.328	0.048	0.028
	Bias-corrected	0.011	0.043	0.222	0.001	0.158	0.047	0.006
	95% CI	(-0.044, 0.062)	(-0.195, 0.859)	(0.005, 0.491)	(-0.016, 0.058)	(-0.212, 0.705)	(0, 0.271)	(-0.016, 0.043)
	Apparent	0.012	0.059	0.208	-0.004	-0.064	0.066	0.004
	Bias-corrected	-0.003	-0.012	0.276	-0.009	-0.086	0.038	-0.006
Rule 5	95% CI	(-0.028, 0.038)	(-0.207, 0.314)	(0.069, 0.459)	(-0.030, 0.036)	(-0.295, 0.400)	(0.006, 0.318)	(-0.020, 0.025)
	Apparent	0.031	0.106	0.291	0.037	0.281	0.131	0.034
	Bias-corrected	-0.004	-0.085	0.341	0.002	0.066	0.093	-0.001
	95% CI	(-0.040, 0.057)	(-0.210, 0.684)	(0.059, 0.488)	(-0.035, 0.057)	(-0.215, 0.708)	(0.009, 0.347)	(-0.028, 0.040)

Note: β_S denotes the treatment effect among those recommended $T = 0$ and τ_S denotes the proportion recommended $T = 0$ in study s .