 DOI: 10.1111/biom.12448

Bayesian Predictive Modeling for Genomic Based Personalized Treatment Selection

Junsheng Ma, Francesco C. Stingo, and Brian P. Hobbs*

email: bphobbs@mdanderson.org

Summary. Efforts to personalize medicine in oncology have been limited by reductive characterizations of the intrinsically complex underlying biological phenomena. Future advances in personalized medicine will rely on molecular signatures that derive from synthesis of multifarious interdependent molecular quantities requiring robust quantitative methods. However, highly parameterized statistical models when applied in these settings often require a prohibitively large database and are sensitive to proper characterizations of the treatment-by-covariate interactions, which in practice are difficult to specify and may be limited by generalized linear models. In this article, we present a Bayesian predictive framework that enables the integration of a high-dimensional set of genomic features with clinical responses and treatment histories of historical patients, providing a probabilistic basis for using the clinical and molecular information to personalize therapy for future patients. Our work represents one of the first attempts to define personalized treatment assignment rules based on large-scale genomic data. We use actual gene expression data acquired from The Cancer Genome Atlas in the settings of leukemia and glioma to explore the statistical properties of our proposed Bayesian approach for personalizing treatment selection. The method is shown to yield considerable improvements in predictive accuracy when compared to penalized regression approaches.

Key words: Bayesian analysis; Genomics; Partial exchangeability; Personalized medicine; Predictive probability; Unsupervised clustering

1. Introduction

Cancer is a complex, dynamic microevolutionary process. It was the cause of more than 580 thousand deaths in the US alone in 2013, and has remained the second leading cause of death for more than half a century (http://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm). Effective treatment requires an understanding of the alterations within cell signaling pathways that enable cancer cells to evade routine cell death and to proliferate and migrate (Alberts et al., 2002). Variations in the genomes of cancer patients and among cancer cells within the same tumor make the disease inherently heterogeneous (Bedard et al., 2013). The study of individual candidate genes, signaling pathways, behaviors, or environmental exposures in lieu of better approximations of the intrinsically complex biological phenomena has limited our understanding of many areas of oncology (Knox, 2010). For example, after nearly two decades of biomarker investigations in colorectal cancer, only one biomarker has been sufficiently validated to warrant its inclusion in clinical practice: *KRAS* gene mutations that are predictive of resistance to anti-EGFR therapy in metastatic disease (Deschoolmeester et al., 2010). In fact, final market approval was achieved for 6.7% of new cancer agents that initiated clinical testing from 2003 to 2011 (Hay et al., 2014).

Many conventional statistical methods assume that patients are exchangeable within a few disparate biomarker/mutation subtypes, which fails to account for potential subsets within the putative marker cohorts (Maitournam and Simon, 2005; Lee et al., 2010; Kim et al., 2011). Other methods accommodate patient heterogeneity through covariate adjustment Zang et al., 2012; Kang, Janes, and Huang, 2014. These existing methods require assumptions for characterizing treatment-by-covariate interactions, which in practice are difficult to specify and may be limited by generalized linear models. Future breakthroughs in personalized medicine will rely on molecular signatures derived from the synthesis of multifarious interdependent molecular quantities that will require more advanced quantitative methods. Moreover, because each tumor is unique, patients should not be considered statistically "exchangeable." Rather, the extent to which results from previously treated patients inform our expectation of treatment success for a future patient should depend upon our current understanding of the extent to which each new patient's tumor exhibits similarity with the tumors of the previously treated patients.

In this article, we introduce a utility-based Bayesian predictive framework that enables personalized treatment selection for new patients based on the treatment histories and molecular measurements of previously treated patients. The proposed approach involves the three sequential components illustrated in Figure 1. First, we use unsupervised clustering to quantify the relative molecular similarity between any two patients in the presence of the available molecular information acquired for all patients. Second, we use the molecular similarity measure to characterize the extent to which each historical patient is partially informative in relation to each new, heretofore untreated patient. Then we use a utility-based Bayesian model to combine the clinical responses and treatment histories of historical patients with the similarity measures for each new patient, providing a probabilistic ba-

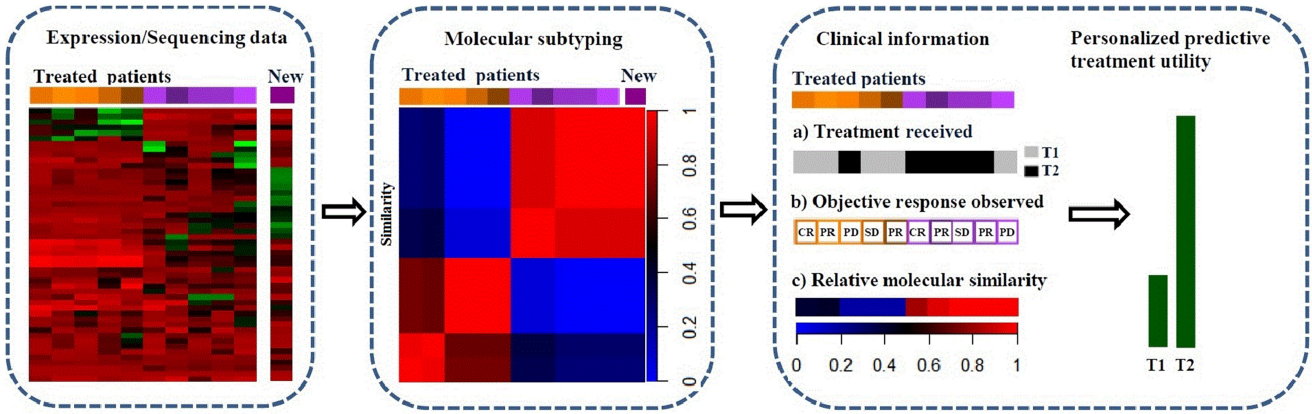

Figure 1. The proposed process for personalizing treatment based on genomic data using Bayesian predictive modeling. Left box, quantifying the relative molecular similarities among all patients by applying unsupervised clustering methods to genomic data; middle box, characterizing the extent to which each historical patient should be considered partially exchangeable with the new patient based on the measure of relative molecular similarity; right box, treatment selection using a Bayesian prediction model that integrates the genomic similarity measures with the clinical information, which in this case consists of ordinal outcomes based on the Response Evaluation Criteria In Solid Tumors (RECIST): progressive disease (PD), stable disease (SD), partial response (PR), and complete response (CR). Treatment 2 (T2) obtains higher predictive utility value, and therefore is recommended for the new patient.

sis for using the clinical and molecular information acquired from each historical patient to predict each new patient's personalized treatment utility under each candidate therapy. Our approach relies on a heuristic measure of molecular similarity, which provides robustness to the underlying stochastic data model and feasibility in the presence of a small sample.

The ideas in this article are conveyed in the following sequence. In Section 2, we present the Bayesian predictive method for personalized treatment selection. In Section 3, we use simulations to evaluate the method's capacity to learn from accumulating information in the presence of complex predictive relationships between molecular quantities and treatment response. In Section 4, we demonstrate the potential clinical utility of our personalized treatment strategy when applied to a data set of lower grade glioma obtained from The Cancer Genome Atlas (TCGA). We provide concluding remarks in Section 5.

2. Method

We assume that patients who have tumors that are molecularly similar are partially exchangeable, and hence avoid conducting inference on a potentially complex but misspecified model for the treatment-covariate interactions. Our modeling approach involves two steps. In the first step, we apply unsupervised clustering to generate similarity measures based on the genomic covariates. In the second step, we integrate the similarity measures into a Bayesian model that we use to compute the personalized predictive utility of each candidate treatment. There is a very rich body of literature on the use of clustering algorithms and techniques in different scientific disciplines (Jain, 2010). We briefly introduce two methods commonly used in oncology in Section 2.1, and present the Bayesian predictive model in Section 2.2. In Section 2.3, we demonstrate how to compute a personalized predictive treatment utility.

2.1. *Quantifying Tumor Similarity between Patients Using Unsupervised Clustering*

Using the available methods for unsupervised clustering, we can quantify molecular similarities among tumors in relation to the available genomic information provided by previously treated patients. Several studies have shown that clustering approaches based on multiple genes can successfully predict clinical outcomes (Fan et al., 2006), whereas methods based on a small set of genes may not fully characterize tumor heterogeneity (Kelloff and Sigman, 2012). Importantly, other recent studies have confirmed that molecular subtypes defined by clustering methods were associated with clinical endpoints (Sadanandam et al., 2013). However, because different clustering algorithms may result in different partitions of the same data set (Jain, 2010), these methods may yield different patient-cluster assignments. This undesirable scenario occurs frequently when the relevant characteristics of some patients lie near the boundaries between two or more groups (Lai et al., 2013). Consequently, subgroup analyses that use groups defined by clusters are limited in this context because they ignore intra-group heterogeneity.

To overcome this issue, our prediction model incorporates a continuous measure of similarity that assumes a value on the unit interval $[0, 1]$. The similarity measure quantifies the extent to which the tumors from any two patients may be considered similar based on the current understanding of the extent of heterogeneity exhibited by the disease. In Section 2.2, we propose a statistical model for predicting the treatment response of each new patient that assigns a higher weight to information contributed by historical patients whose tumors exhibit greater similarity. At the boundaries, patients who attain similarity 1 are assumed to be statistically exchange-

able, while the treatment histories for patients with similarity measure 0 fail to inform the prediction.

We consider a data set consisting of l biomarkers in n patients, denoted by $\boldsymbol{X}_{l\times n}$. Without loss of generality, we treat these biomarkers as continuous variables, such as gene or protein expression levels. We derive the pairwise similarities from clustering methods based on $\boldsymbol{X}_{l\times n}$ and store them in the $n \times n$ symmetric similarity matrix $\boldsymbol{S}$. Two widely applied approaches, consensus clustering (CC) (Monti et al., 2003) and nonnegative matrix factorization (NMF) (Brunet et al., 2004), are explored in our simulation and case studies. More details of these two methods are provided in Section A of the Supplementary Materials.

2.2. *Bayesian Predictive Model for Partially Exchangeable Data*

Let $j = 1, \ldots, J$ index candidate therapies in a training data set consisting of data from $n = \sum_{j=1}^{J} n_j$ patients, where n_j denotes the number of patients treated with therapy j. We assume that the therapeutic response to the treatment (e.g., tumor response), evaluated via an ordinal-valued criteria (such as the commonly utilized response criteria described in the next section) has been ascertained and that baseline molecular characteristics have been measured for each patient. Let y_i denote the random variable of the ith patient's response to therapy $A_i = j$ among K possible levels of increasing treatment benefit, where $y_i = k$, for $i = 1, \ldots, n$, and $k = 0, 1, \ldots, K-1$. In addition, let $\boldsymbol{\theta}_j = \{\pi_{j,k} : k = 0, 1, 2, \ldots, K-1\}$, denote the model parameters associated with the jth therapy, where $\pi_{j,k}$ characterizes the probability of observing outcome k under treatment j. Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_j : j = 0, 1, 2, \ldots, J\}$ denote the collection of all model parameters. As the treatment response is represented by an ordinal value, y_i follows a multinomial distribution $y_i|A_i = j \sim \textit{Multinomial}(\boldsymbol{\theta}_j)$. We devise a personalized treatment rule for the purpose of selecting a therapy for a newly enrolled, untreated patient on the basis of the treatment histories of previously treated patients. Thus, statistically, the selection rules should derive from statements of predictive probability which accounts for both the uncertainty arising from estimation of the model parameters using the historical information and the extent of variability in ascertaining the responses of future patients. Let $\tilde{i}$ index a new, heretofore untreated patient. Given the observed responses for the n_j patients previously treated with therapy j, which we denote by $\boldsymbol{Y}_j$, we represent the predictive probability of response level k under treatment j as

$$Pr(y_{\tilde{i}} = k|\boldsymbol{Y}_j, A_{\tilde{i}} = j, \boldsymbol{x}_{\tilde{i}}), \tag{1}$$

where $\boldsymbol{x}_{\tilde{i}}$ denotes the l-dimensional vector containing measurements from l molecular biomarkers for the $\tilde{i}$th patient. Note that (1) also depends on the biomarker measures $\boldsymbol{X}_{l\times n}$ for all previously treated patients, which we omit to simplify the notation.

We proceed to derive (1) using a Bayesian model, whereby the extent to which two patients are considered exchangeable is determined by the similarity measure. This can be achieved using a power prior model. A power prior represents a "weighted" Bayesian update of an "initial prior" $p(\boldsymbol{\theta})$ using a supplemental data source, $\boldsymbol{D}_0$, for inference on a primary data source $\boldsymbol{D}$ (Ibrahim et al., 2003). In our setting, $\boldsymbol{D}_0 = \{\boldsymbol{Y}, \boldsymbol{X}\}$ represents the data for all historically treated patients, and the primary data source is $\boldsymbol{D} = \{y_{\tilde{i}}, \boldsymbol{x}_{\tilde{i}}\}$. Let $L(\boldsymbol{\theta}|\cdot)$ denote the likelihood function, assumed to be common for both data sources. Formally under the power prior formulation, the posterior distribution of $\boldsymbol{\theta}|\boldsymbol{D}, \boldsymbol{D}_0$ arises as $p(\boldsymbol{\theta}|\boldsymbol{D}, \boldsymbol{D}_0) \propto L(\boldsymbol{\theta}|\boldsymbol{D})L(\boldsymbol{\theta}|\boldsymbol{D}_0)^{a_0}\, p(\boldsymbol{\theta})$, where the exponent, $a_0 \in [0, 1]$, controls the extent to which the supplemental data influence the posterior. Because the likelihood functions for the supplemental and primary data sources assume identical model parameters, $\boldsymbol{\theta}$, formally a_0 can be interpreted as a measure of the extent to which $\boldsymbol{D}_0$ is "exchangeable" with $\boldsymbol{D}$.

Recall that the (r, i)th entry of the similarity matrix, $\boldsymbol{S}$, represents the pairwise similarity measure between patients r and i. In order to identify the best therapy for the new patient, we use the power prior framework to combine the observed treatment outcomes with the pairwise similarity measures for the new patient, $\{\boldsymbol{S}(\tilde{i}, 1),\ y_1\}, \ldots, \{\boldsymbol{S}(\tilde{i}, n),\ y_n\}$, to compute the predictive probability of treatment success under each candidate therapy. Specifically, we use the pairwise similarity measures as exponents, a_0, and thereby assume that the similarity measures characterize the extent to which each historical patient is statistically exchangeable with the new patient. Assuming independent $\textit{Dirichlet}(\boldsymbol{\alpha}_j)$ "initial priors" for each therapy, $j = 1, ..., J$, where we choose hyperparameters $\boldsymbol{\alpha}_j = \{\alpha_{j,k} : k = 0, 1, 2, \ldots, K-1\}$ for model parameters $\boldsymbol{\theta}_j = \{\pi_{j,k} : k = 0, 1, 2, \ldots, K-1\}$, the posterior distribution is

$$p(\boldsymbol{\theta}_j|\boldsymbol{Y}_j, y_{\tilde{i}}) \propto L(\boldsymbol{\theta}_j|y_{\tilde{i}}) \left\{ \prod_{i=1}^{n_j} L(\boldsymbol{\theta}_j|y_i)^{\boldsymbol{S}(\tilde{i},i)} \right\} p(\boldsymbol{\theta}_j). \tag{2}$$

The index j indicates that data from all patients previously treated with therapy j is utilized when constructing the treatment-specific power priors. To simplify the notation, we assume that data for patients assigned treatment j are of the order $\boldsymbol{S}(\tilde{i}, 1), \ldots, \boldsymbol{S}(\tilde{i}, n_j)$.

For treatment selection, we are interested in the predictive probability of $y_{\tilde{i}} = k$, (1), which derives from the integration of (2) with respect to $\boldsymbol{\theta}_j$. The posterior predictive distribution of $(y_{\tilde{i}}|\boldsymbol{Y}_j, A_{\tilde{i}} = j, \boldsymbol{x}_{\tilde{i}})$ is given as

$$p(y_{\tilde{i}}|\boldsymbol{Y}_j, A_{\tilde{i}} = j, \boldsymbol{x}_{\tilde{i}}) = \int L(\boldsymbol{\theta}_j|y_{\tilde{i}}, A_{\tilde{i}} = j) \times \left\{ \prod L(\boldsymbol{\theta}_j|y_i)^{\boldsymbol{S}(\tilde{i},i)} \right\} p(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j. \tag{3}$$

Defining the indicator $I_k(y_i) = 1$ if $y_i = k$ and 0 otherwise, we can simplify the power prior as: $\left\{\prod_{i=1}^{n_j} L(\boldsymbol{\theta}_j|y_i)^{\boldsymbol{S}(\tilde{i},i)}\right\} p(\boldsymbol{\theta}_j) \propto \prod_{k=0}^{K-1} (\pi_{j,k})^{\left\{\sum_{i=1}^{n_j} \boldsymbol{S}(\tilde{i},i) I_k(y_i) + \alpha_{jk} - 1\right\}}$. Recognizing that $\prod_{k=0}^{K-1} (\pi_{j,k})^{\left\{\sum_{i=1}^{n_j} \boldsymbol{S}(\tilde{i},i) I_k(y_i) + \alpha_{jk} - 1\right\}}$ is the kernel of a $\textit{Dirichlet}(\tilde{\boldsymbol{\alpha}}_j)$ distribution, with $\tilde{\boldsymbol{\alpha}}_j = \{\tilde{\alpha}_{j,k} : k = 0, 1, 2, \ldots, K-1\}$ and $\tilde{\alpha}_{jk} = \sum_{i=1}^{n_j} \boldsymbol{S}(\tilde{i}, i) I_k(y_i) + \alpha_{jk}$, the power prior is

$$\left\{ \prod_{i=1}^{n_j} L(\boldsymbol{\theta}_j|y_i)^{\boldsymbol{S}(\tilde{i},i)} \right\} p(\boldsymbol{\theta}_j) = \frac{\Gamma(\sum_{k=0}^{K-1} \tilde{\alpha}_{jk})}{\prod_{k=0}^{K-1} \Gamma(\tilde{\alpha}_{jk})} \prod_{k=0}^{K-1} (\pi_{j,k})^{\left\{\sum_{i=1}^{n_j} \boldsymbol{S}(\tilde{i},i) I_k(y_i) + \alpha_{jk} - 1\right\}}. \tag{4}$$

Thus, we can express the predictive density (3) as follows:

$$p(y_{\tilde{i}}|\boldsymbol{Y}_j, A_{\tilde{i}} = j, \boldsymbol{x}_{\tilde{i}}) = \frac{\Gamma(\sum_{k=0}^{K-1} \tilde{\alpha}_{jk})}{\prod_{k=0}^{K-1} \Gamma(\tilde{\alpha}_{jk})} \int \int \dots \int L(\pi_{j,1}, \pi_{j,2}, \dots, \pi_{j,K}|y_{\tilde{i}}, A_{\tilde{i}} = j) \times \prod_{k=0}^{K-1} (\pi_{j,k})^{\tilde{\alpha}_{jk}} \, d\pi_{j,1}, d\pi_{j,2}, \dots, d\pi_{j,K-1}.$$

Therefore, the power prior predictive probability of $y_{\tilde{i}} = k$ (1) is

$$p(y_{\tilde{i}} = k|\boldsymbol{Y}_j, A_{\tilde{i}}=j, \boldsymbol{x}_{\tilde{i}})=E(\pi_{j,k}|\tilde{\alpha}_{j1}, \tilde{\alpha}_{j2}, \dots, \tilde{\alpha}_{jK}) = \frac{\tilde{\alpha}_{jk}}{\sum_{k=0}^{K-1} \tilde{\alpha}_{jk}}. \quad (5)$$

Note that the total "prior effective sample size" (or amount of effective information contained in the prior when characterized as a sample size) for predicting the new patient's treatment outcome is $\sum_{k=0}^{K-1} \tilde{\alpha}_{jk}$, and thus depends on the extent to which the new patient's tumor exhibits similarity with those of the previously treated patients.

In addition, (5) provides the predictive probability of treatment success for the special case of binary outcomes. Thus, among the J candidate therapies, the one with the maximum predictive probability would be recommended. However, this strategy is limited when used for treatment selection in settings with multinomial response categories (e.g., short-term tumor response).

2.3. *Personalized Predictive Treatment Utility*

In this section, we consider how to proceed with treatment selection in a multinomial setting, where we use utilities to obtain an one-dimensional selection criterion. In clinical oncology, often there are multiple therapeutic objectives that pertain to the extent of loco-regional control of a tumor, distant tumor progression, and the occurrence of toxicity. Short-term "surrogate" endpoints, such as "objective tumor response," are most commonly used in lieu of long-term progression-free or overall survival durations, which represent the primary therapeutic endpoints for demonstrating clinical benefit (FDA, 2007). Tumor response can be observed shortly after treatment, for example, using RECIST (http://www.recist.com/), which considers both changes in tumor size as well as the occurrence of new lesions and/or distant migration (Michaelis and Ratain, 2006). According to RECIST, the four *ordinal* tumor response categories are progressive disease (PD), stable disease (SD), partial response (PR), and complete response (CR). Decisions pertaining to treatment selection in this multinomial setting need consider the relative importance of each level of the ordinal response. Section B of the Supplementary Materials describes other types of ordinal-valued response criteria that are commonly utilized in oncology.

We can establish utility weights by considering the relationship between tumor response and the primary therapeutic goal (e.g., overall survival), whereby the utilities define the extent to which each short-term response level achieves the clinical objectives. Let w_k denote the utility assigned to tumor response level k. Section D.2 of the Supplementary Materials describes strategies for specifying the utilities in the presence of historical data. Given these weights, we can calculate the mean "predictive utility" of treatment j for patient $\tilde{i}$ as

$$\mu_j(\tilde{i}) = \sum_{k=0}^{K-1} w_k \times p(y_{\tilde{i}} = k|\boldsymbol{Y}_j, A_{\tilde{i}} = j, \boldsymbol{x}_{\tilde{i}}). \quad (6)$$

We assign to patient $\tilde{i}$ the treatment with the largest predictive utility, which may be considered as "optimal" among the available therapies given the current information.

3. Simulation Study

We used simulation to evaluate the performance of the proposed method for personalized treatment selection. Our simulation study was based on actual gene expression data from a well-known data set of leukemia, which we describe in Section C.1 of the Supplementary Materials. We simulated the application of NMF as well as CC with hierarchical (HC), k-means (KM) and partitioning around medoids (PAM) algorithms (Gaujoux and Seoighe, 2010; Wilkerson and Hayes, 2010). Hereafter, we refer to these methods as NMF, HC, KM, and PAM. We compared the performance of our methods to that of commensurate treatment selection rules obtained from three competing approaches: an unweighted version of our Bayesian predictive approach that assumes all patients are exchangeable, the L_1 penalized continuation ratio model (or Lasso), and ridge regression (Archer and Williams, 2012). After fitting separate models for each treatment group, personalized treatments were selected to maximize the resulting predicted utility. Identical response utility weights were used for all methods.

3.1. *Simulation Approach*

Simulation scenarios. We considered two treatments under three scenarios. In scenario 1, treatment 1 was generally more beneficial for ALL (patients 1–27), while treatment 2 targeted AML (patients 28–38). We determined the response probabilities for scenario 1 from the linear combination of two metagenes obtained from NMF with rank 2. Scenario 2 mimicked scenario 1; however, we used the first principal component obtained from PCA to determine the response probabilities. In scenario 2, the true response probabilities for patients 20–27 were slightly better than those for patients 1–19; whereas the true response probabilities in scenario 1 were roughly similar for patients 1–27. In scenario 3, we used a linear combination of the first two principal components. For this scenario, treatment 1 was effective for all patients, with the exception of patients 20–27, who benefited from treatment 2. Details pertaining to the simulation design are provided in Section C.2 of the Supplementary Materials.

For all methods and scenarios, we fixed the response utility weights at $w_0 = 0$, $w_1 = 20$, $w_2 = 60$, and $w_3 = 100$ to reflect conventional tumor response criteria based on PD, SD, PR, and CR, respectively. After specifying the true parameters of the generating model for each simulation scenario, we used the resulting response probabilities $P(y_i = k|\boldsymbol{x}_i, A_i = j)$ to calculate the true mean treatment utility (MTU) for patient i with treatment j, which we denote hereafter by $\text{MTU}(i, j) = \sum_k w_k P(y_i = k|\boldsymbol{x}_i, A_i = j)$. Figure S1 in Section C.3 of the Supplementary Materials depicts the true differ-

ences in the MTUs, $\Delta\text{MTU}(i)=\text{MTU}(i,1)-\text{MTU}(i,2)$, that result for each patient, $i = 1, \ldots, 38$, under each of the three simulation scenarios. The scenarios reflect the following inevitabilities inherent in personalized medicine in oncology: 1) patients whose cancers exhibit molecular similarity may benefit (or not) from the same treatment; 2) the benefits may vary largely even among patients classified into the same broad subgroups; 3) the targeted treatment may yield only minimal benefit, especially for those "boundary" patients who possess genomic characteristics intrinsic to both subgroups (e.g., note that in scenario 2 the differences in utilities for ALL-B patients 6 and 11 are close to zero). In the original study, two patients were misclassified in terms of cancer subtype, while patient 29 was misclassified in our simulation. Possible explanations are that these samples were near the boundaries between clusters or these patients were misdiagnosed (Brunet et al., 2004). Because we used gene expressions (as opposed to leukemia subtype) to determine the response probabilities, patients who were originally misclassified were retained in our simulation study to evaluate the extent to which the performance of the proposed genomic-based method for personalized treatment selection would vary for these patients.

Treatment selection. For each patient, we predicted the mean utility (6) of each treatment using the observed responses acquired for the other $n = 37$ patients. We assigned patients to the treatment that yielded the highest predicted mean utility. For each iteration of the simulation, this process was repeated until treatments were assigned for all $n = 38$ patients for each of the considered methods. The proposed Bayesian predictive model assumed a flat prior, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \sim \textit{Dirichlet}(0.25, 0.25, 0.25, 0.25)$, for the response probability vector corresponding to each of the two treatments. For the Lasso and ridge regression methods, we used Akaike's information criterion (AIC) to select the best model among those within a specified range of the penalty parameter λ (from 0.01 to 1 times $\lambda_{\max}$, which is the smallest value such that the coefficients are zero) (Friedman et al., 2010).

Performance evaluation. Because the treatment-specific mean utilities are known for each patient, we can compare the simulated treatment assignments to the true optimal treatment determined by the simulation scenario for each patient. The average number of patients who were assigned non-optimal treatments (referred to as CT) provides a simple measure for comparing the methods. In addition, because patients are inherently heterogeneous, the extent to which the optimal treatment provides an improvement (in relation to the non-optimal treatment) varies for each patient. Intuitively, treatment selection should be compared in consideration of the extent to which each assignment yields differential treatment benefit, which is determined for each patient by $\Delta\text{MTU}(i)$. Therefore, we also compared the various methods on the basis of the extent of gain achieved by each simulated treatment assignment as measured by the difference in the true mean treatment utilities (the assigned minus the non-assigned). To facilitate this comparison, we needed to consider a new summary statistic. Let $j^*(i) = \arg\max_j \left\{\sum_{k=0}^{K-1} w_k \times P(y_i = k|\boldsymbol{x}_i, A_i = j)\right\}$, represent the true optimal treatment for patient i and let $j_g(i)$ denote the treatment recommended by selection rule g. Letting indicator $A^*\left\{j_g(i)\right\} = 1$ if $j_g(i) = j^*(i)$ and -1 otherwise, the *total MTU* that is obtained for any selection rule can be defined as the sum of true gains (or losses) achieved for each treatment assignment, $\Delta\text{MTU}_g = \sum_i^n A^*\left\{j_g(i)\right\}|\Delta\text{MTU}(i)|$. For each simulation scenario, the maximum possible total MTU, which we denote by $\Delta\text{MTU}_{\text{opt}} = \sum_i^n |\Delta\text{MTU}(i)|$, is achieved when all patients are assigned to their respective optimal treatments. We compared the total MTU resulting from each competing method for treatment selection in proportion to the maximum possible total MTU, which we denote by $\%\Delta\text{MTU}_g = \Delta\text{MTU}_g/\Delta\text{MTU}_{\text{opt}}$. Thus, $\%\Delta\text{MTU}_g$ is bounded above by 1, which represents the best possible simulation result. Larger values of $\%\Delta\text{MTU}_g$ imply that rule g yielded better overall performance for treatment selection. We generated a total of 100 duplicated data sets to calculate CT and $\%\Delta\text{MTU}_g$ for each of the methods.

3.2. *Simulation Results*

The proposed Bayesian predictive methods generally outperformed the methods based on penalized regression (ridge regression and Lasso), with the exception of PAM in scenarios 1 and 2 (Table 1). Among the Bayesian predictive implementations, there seems to be no uniformly superior approach for all scenarios. However, NMF maintained relatively accurate predictive performance for treatment selection in all scenarios, with HC achieving a close second best performance. In contrast, PAM worked very well only in scenario 3. Assignments based on the elastic net (Zou and Hastie, 2005) were also simulated, but effectuated only slight improvement when compared to Lasso and ride regression (results are provided in Section C.4 of the Supplementary Materials).

We found consistency between the numbers of patients assigned to non-optimal treatments (CT) and the results obtained for $\%\Delta\text{MTU}_g$. For example, PAM (in scenarios 1 and 2), ridge regression and the Lasso method assigned more patients, on average, to the non-optimal treatment. Consequently, $\%\Delta\text{MTU}_g$ obtained for these approaches were significantly lower than those of the competing methods. For instance, in scenario 1 with the top 100 genes, NMF3 achieved the best possible result, with 0 loss in the maximum mean utility ($\%\Delta\text{MTU}_g$=1); whereas Lasso resulted in a 23% loss ($\%\Delta\text{MTU}_g$=0.77) on average. This represents a 30% increase in $\%\Delta\text{MTU}_g$ for NMF3 when compared to Lasso.

To clarify the relative performance of each method, we provide more detailed results for a subset of patients in Table 2. The top panel in Table 2 displays the results for 10 patients for whom the smallest differences in the true mean treatment utilities were found. Thus, for these patients, the respective optimal treatments should be more difficult to determine from the data. We found that both ridge and Lasso method frequently assigned patients to the less beneficial treatment among the cohort of patients for whom the optimal treatment provides limited clinical benefit. In contrast, the NMF-based approaches assigned the most patients to the optimal treatments almost uniformly, with the exception of patient 17, for whom the less beneficial treatment was chosen more than 80 times (out of 100). The results achieved by HC3-5 were similar, with diminished performance for patients 6 and 29. For comparison, NMF3 assigned all of these patients (except patient 17) to the less beneficial treatments fewer than 4 times (out of 100); whereas Lasso assigned all patients in

Table 1

Simulation results for the scenarios portrayed in Figure 2 based on the top 100 and 200 varied genes. $\%\Delta_g$=$\%\Delta MTU_g$ represents the average proportion of the maximum possible gain in total MTU. CT represents the average number of patients who were assigned to non-optimal treatments. We implemented the Bayesian prediction method using four clustering methods: NMF3–NMF5, HC3–HC5, KM3–KM5, and PAM3–PAM5, which represent the non-negative matrix factorization method as well as consensus clustering approaches using hierarchical, k-means, and partitioning around medoids algorithms, with ranks of 3–5, respectively.

	Scenario 1		Scenario 2		Scenario 3	
Treatment selection method	100 Genes $\%\Delta_g$(CT)	200 Genes $\%\Delta_g$(CT)	100 Genes $\%\Delta_g$(CT)	200 Genes $\%\Delta_g$(CT)	100 Genes $\%\Delta_g$(CT)	200 Genes $\%\Delta_g$(CT)
Bayesian predictive approach without similarities						
	0.47 (10.22)		0.15 (13.78)		0.60 (8.00)	
Bayesian predictive approach with similarities						
NMF3	1.00 (0.00)	1.00 (0.00)	0.93 (1.78)	0.93 (1.86)	0.95 (1.19)	0.95 (1.20)
NMF4	1.00 (0.04)	1.00 (0.01)	0.92 (2.17)	0.92 (2.01)	0.95 (1.22)	0.95 (1.21)
NMF5	0.98 (0.48)	0.96 (1.09)	0.91 (2.48)	0.92 (1.93)	0.93 (1.77)	0.93 (1.66)
HC3	0.91 (3.13)	0.91 (3.01)	0.93 (2.82)	0.92 (2.86)	0.95 (1.19)	0.95 (1.19)
HC4	0.91 (3.08)	0.92 (2,72)	0.93 (2.85)	0.92 (2.97)	0.95 (1.19)	0.95 (1.19)
HC5	0.91 (3.05)	0.93 (2.50)	0.92 (2.93)	0.92 (3.04)	0.95 (1.19)	0.92 (1.75)
KM3	0.74 (5.71)	0.74 (5.80)	0.78 (5.26)	0.76 (5.66)	0.81 (3.74)	0.70 (6.02)
KM4	0.87 (2.88)	0.79 (4.63)	0.88 (3.15)	0.82 (4.38)	0.96 (0.89)	0.91 (2.03)
KM5	0.93 (1.54)	0.93 (1.63)	0.90 (2.61)	0.88 (3.15)	0.97 (0.75)	0.94 (1.38)
PAM3	0.67 (6.88)	0.66 (7.06)	0.61 (8.83)	0.56 (9.51)	0.94 (1.49)	0.92 (1.63)
PAM4	0.73 (5.93)	0.70 (6.40)	0.71 (6.84)	0.65 (7.88)	0.95 (1.24)	0.96 (0.64)
PAM5	0.75 (5.63)	0.75 (5.42)	0.74 (6.30)	0.77 (5.41)	0.95 (1.28)	0.96 (0.64)
Penalized regression						
Ridge	0.86 (3.03)	0.88 (3.12)	0.87 (4.02)	0.85 (4.66)	0.88 (3.02)	0.87 (3.13)
Lasso	0.77 (4.94)	0.82 (4.08)	0.66 (8.77)	0.71 (7.40)	0.74 (5.74)	0.73 (5.54)

the top panel to the inferior treatment at least 21 times (with a range of 21–47). The middle panel of Table 2 presents the 10 patients for whom the largest differences in mean treatment utilities were obtained, and thus for whom the respective optimal treatment should be easier to identify. We observed improved treatment selection for these patients under all the competing approaches. However, ridge and Lasso continued to show diminished performance when compared to the Bayesian predictive approaches. For example, NMF3-5 and HC3-5 correctly assigned patient 33 to the optimal treatment for every assignment, while ridge and Lasso assigned this patient to the less beneficial treatment 7 and 13 times (out of 100), respectively.

The bottom panel of Table 2 displays the results for the patient cohort targeted by treatment 2 in scenario 3, with 200 genes. Again, ridge and Lasso assigned patients to less beneficial treatments more often than the Bayesian predictive approaches. For example, patient 26 was assigned to the non-optimal treatment 64 and 41 times, respectively. In contrast, our proposed Bayesian predictive method using NMF and HC resulted in only 1 incorrect assignment for this patient. The poor results for penalized regression approaches may be explained in part by sensitivity to the small sample size, which involved only 8 patients for this subgroup. However, as demonstrated in Section C.5 of the Supplementary Materials, they did not yield improved results when applied in an expanded simulation study involving 152 patients, with 40 patients benefiting from treatment 1.

While the Bayesian predictive approach is useful for treatment selection, its performance does depend on the chosen clustering algorithm, rank and genomic features. Section C.3 of the Supplementary Materials presents results obtained from our full simulation study, wherein we evaluated performance for ranks 2–10 for all clustering methods under consideration. The full simulation results are presented in Table S2. Figures S2 and S3 in Section C.3 of the Supplementary Materials summarize the distributions of these summary statistics. In all simulation scenarios, results for NMF3-9 and HC3-8 (Table S2) are shown to deviate less from their respective means when compared to results obtained for the penalized regression approaches. In contrast, the Bayesian predictive approach based on PAM worked well only in scenario 3. In addition, the results obtained from the top 200 genes were quite close to those resulting from the top 100 genes for every scenario. Additionally, as demonstrated in Section C.5 of the Supplementary Materials, predictive accuracy diminishes as more noise is introduced into the genomic features. Overall, our proposed approach is most beneficial in contexts wherein patient similarity is measured in relation to a set of pre-specified candidate molecular features that have

Table 2

Number of non-optimal treatment assignments among 100 replications of the simulation under scenarios 2 and 3, with 200 genes for a subset of selected patients. NMF3–NMF5 and HC3–HC5 represent the non-negative matrix factorization method and consensus clustering approach using a hierarchical algorithm, with ranks of 3–5, respectively. ΔMTU provides the patient-specific true difference in mean treatment utility, where positive (negative) values favor treatment 1 (treatment 2).

		Bayesian predictive						Penalized regression	
Subject	ΔMTU	NMF3	NMF4	NMF5	HC3	HC4	HC5	Ridge	Lasso
Scenario 2 with 200 genes									
Top 10 patients with least differences in mean treatment utility									
20	1.4	0	0	0	0	0	0	40	40
11	3.2	2	5	7	4	5	10	33	28
6	4.2	4	2	9	98	99	97	43	37
32	−9.5	2	1	4	0	2	1	27	27
10	9.6	0	0	0	0	0	0	19	26
17	−13.3	96	86	11	4	10	14	49	47
29	13.3	4	6	17	100	99	99	23	37
3	13.7	4	6	7	5	4	3	6	21
13	15.9	3	4	6	5	4	4	25	29
7	17.3	3	5	8	4	4	4	13	36
Top 10 patients with largest differences in mean treatment utility									
33	−37.3	0	0	0	0	0	0	7	13
22	40.8	2	2	2	2	2	2	4	7
23	43.5	0	0	0	0	0	0	5	15
25	43.9	0	0	0	0	0	0	1	7
21	45.7	0	0	0	0	0	0	1	6
27	46.3	0	0	0	0	0	0	1	5
38	−54.5	2	2	1	2	2	3	3	12
35	−56.8	4	4	2	1	3	3	1	5
31	−59.4	3	3	1	1	2	2	9	5
30	−63.0	3	3	1	2	2	2	1	1
Scenario 3 with 200 genes									
All patients who benefit from treatment 2									
20	−12.5	1	1	1	1	1	1	81	45
21	−43.7	2	2	2	2	2	2	3	15
22	−40.5	4	4	4	4	4	4	14	15
23	−42.1	3	3	3	3	3	3	10	19
24	−38.3	7	7	7	7	7	7	20	26
25	−39.5	2	2	2	2	2	2	19	18
26	−26.3	1	1	1	1	1	1	64	41
27	−40.7	1	1	1	1	1	1	28	16

been identified as intrinsic to disease pathogenesis in prior study.

4. Case Study of Lower Grade Glioma

4.1. *TCGA Data*

We applied the proposed methods to the publicly available data of lower grade glioma from the TCGA data portal. We downloaded both the clinical and level 3 RNASeqV2 mRNA expression data from https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp. Baseline and follow-up information was collected on 240 patients; 156 (65%) and 136 (57%) patients who received adjuvant radiation and targeted therapies, respectively. We excluded from the analysis the data for three patients who had missing values of responses and one who had no genomic information. We focused on the subset of patients treated with adjuvant radiation and investigated two treatment regimes, targeted therapies ($n = 113$) compared to non-targeted therapies ($n = 41$). Among the 154 patients in this analysis, 52 were classified as having grade II glioma (34%) and 102 as having grade III glioma (66%). We categorized tumor responses into the four standard categories, CR, PR, SD, and PD. As only 5 patients achieved PR, we combined PR and CR into a new category of treatment responders, which we abbreviate as CP. The three levels of the ordinal outcomes are therefore CP, SD, and PD. Since the TCGA data were collected from observational studies, we matched 41 pairs of patients to avoid potential bias. More details pertaining to the matching procedure and resulting covariate balance can be found in Section D.1 of the Supplementary Materials.

4.2. *Data Analysis*

We calculated the maximum absolute deviation (the maximum value minus the minimum value) for each gene, selected the top 200 most varying genes for the analysis, and standardized the gene expression values. As the RNASeqV2 data are read counts, which are naturally non-negative, we applied NMF to the original expression data. As explained in Section D.2 of the Supplementary Materials, we fixed the utility weights as 0, 25, and 100 for PD, SD, and CP, respectively. Hereafter, we discuss an empirical summary measure that we use to evaluate the potential clinical impact of our proposed personalized treatment selection method in this setting.

Summary measures. Following published approaches (Song and Pepe, 2004; Kang et al., 2014), we assessed model performance according to the following summary measure, which can be interpreted as the increased clinical benefit derived from biomarker-driven treatment selection with respect to randomized treatment assignment. Let $\tilde{R}$ denote the random treatment response, and define $\Delta(\boldsymbol{X}) = P(\tilde{R} = 1|A = 1, \boldsymbol{X}) - P(\tilde{R} = 1|A = 0, \boldsymbol{X})$ where $A = 0, 1$ represents the targeted and non-targeted treatments, respectively. The relative increase in the population response rate under a biomarker-based treatment selection method in relation to randomized assignment is defined as

$$\{P(\tilde{R} = 1|A = 1, \Delta(\boldsymbol{X}) > 0) \times P(\Delta(\boldsymbol{X}) > 0) + P(\tilde{R} = 1|A = 0, \Delta(\boldsymbol{X}) < 0) \times P(\Delta(\boldsymbol{X}) < 0)\} - P(\tilde{R} = 1|A),$$

which can be estimated from the data (Kang et al., 2014), i.e., the overall response rate calculated as the number of responders (CP) divided by the total sample size. We used the response rates for the subset of patients assigned by the proposed method (e.g., $\Delta(\boldsymbol{X}) > 0$) to the treatment actually received (e.g., $A = 1$) to estimate the extent of clinical benefit that is attributable to the proposed method (Song and Pepe, 2004). The weight $P(\Delta(\boldsymbol{X}) > 0)$ can be empirically estimated as the number of patients who are recommended for targeted treatment divided by the total number of patients. Quantities involving $A = 0$ and $\Delta(\boldsymbol{X}) < 0$ can be similarly calculated. The difference in the two quantities therefore measures the increased clinical benefit obtained under a specific treatment selection method, which we reference hereafter as the "summary measure." Note that we based this summary measure on only two response categories, (CP) and non-responders (PD + SD); whereas, we used all three levels of the ordinal outcome in the data analysis and to implement personalized treatment selection.

4.3. *Results*

We investigated the performance of the clustering approaches, implementing cluster ranks of 2–10. The results are presented in detail in Section D.3 of the Supplementary Materials. The summary measures for NM2, HC7, KM2, PAM7 were 0.103, 0.187, 0.055, 0.091, respectively. Results obtained from all implementations of HC were noticeably better than those obtained from any of the other approaches we considered. We also observed that model performances varied by cluster rank. This was likely caused by discrepancies between cluster assignments that result from varying the cluster rank, which suggests that some implementations fail to identify potential hidden clusters in the data (Li et al., 2009).

We further conducted leave-one-out cross-validation (LOOCV) analyses using the summary measure as the criteria for model (rank) selection (Foster et al., 2011; DeRubeis et al., 2014). We followed the same procedure for the ridge and Lasso methods, but used AIC for model selection. The LOOCV based summary measures for NMF, HC, KM, PAM, ridge regression and Lasso were 0.037, 0.158, −0.002, 0.02, 0.017, and −0.001, respectively. Again, HC yielded the highest value of the summary measure, and we also found that rank=7 was selected most of the time (35/82). Based on these results, we would recommend that investigators use the Bayesian predictive method with HC7 when selecting treatments for future patients with low grade glioma. Additional discussion pertaining to cluster and rank selection can be found in Section D.4 of the Supplementary Materials.

5. Discussion

In this article, we propose a utility-based Bayesian predictive modeling approach to inform treatment selection. This method circumvents the need to assume that patients are fully "exchangeable," and hence is suitable for highly heterogeneous diseases such as cancer. We conducted simulation studies to evaluate the method's performance and analyzed a data set of low grade glioma to illustrate the utility of the proposed approach. Our analysis of the low grade glioma data showed that the overall response rate could be, on average, increased by 15.8% when compared to a default strategy of randomly allocating patients treatments, if the treatment selection recommendations derived from the Bayesian predictive method with HC7 were followed. Moreover, the proposed Bayesian predictive approach was shown to outperform the penalized regression approaches in both the simulation and case studies. Additional comparisons with selection methods devised for low-dimensional settings, such as those proposed by Zhang et al. (2012), could have been considered following application of dimension reduction techniques (e.g., principal component analysis).

Our work represents one of the first attempts to define personalized treatment assignment rules based on large-scale genomic data. The following criteria should be considered carefully before implementing the approach in practice: specimen quality, assay procedures, model evaluation, response criteria, and clinical trial designs (McShane et al., 2013). In this article, we used data from existing observational studies. While our simulation study utilized the actual genomic data to generate patient responses, actual treatment outcomes were used in the analysis of our case study, wherein analysis considered a matched subset to attenuate potential selection bias. Specifically, we matched patients by age, gender, and tumor grade. However, because bias due to unmeasured confounding factors cannot be removed, one should use caution when interpreting the results. The methodology presented here is perhaps most useful when implemented with training data that was acquired from randomized study, and thus devoid of selection bias.

Our model can be easily implemented and is computationally efficient. While the computation time required to obtain a

similarity measure depends on the data size, clustering methods, and ranks, we found that the computation cost was minimal. For instance, it took up to 8.5 minutes for 100 runs with the rank of 3 for the leukemia data with 38 patients and 5000 genes (Gaujoux and Seoighe, 2010). We plan to pursue generalizations of this approach, including extensions of our power prior approach to accommodate other settings, such as survival endpoints and the inclusion of clinical covariates in the treatment selection rule. Another future avenue we plan to pursue is to define an integrative probabilistic framework in which model-based clustering and treatment assignments are jointly estimated from the data. This substantial methodological development will provide a more flexible and statistically sound approach as well as circumvent the need to pre-specify a given clustering approach and rank.

6. Supplementary Materials

Web supplementary sections of A, B, C, and D referenced in Sections 2–4 as well as R code to implement the proposed methods are available at the *Biometrics* Website on Wiley Online Library.

Acknowledgements

The second and third authors were supported by the Cancer Center Support Grant (CCSG) (P30 CA016672). The authors thank LeeAnn Chastain for editing assistance.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., et al. (2002). *Molecular Biology of the Cell*. New York: Garland Science, 4th edition.

Archer, K. and Williams, A. (2012). L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine* **31**, 1464–1474.

Bedard, P. L., Hansen, A. R., Ratain, M. J., and Siu, L. L. (2013). Tumour heterogeneity in the clinic. *Nature* **501**, 355–364.

Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences* **101**, 4164–4169.

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., and Lorenzo-Luaces, L. (2014). The personalized advantage index: Translating research on prediction into individualized treatment recommendations. a demonstration. *PloS One* **9**, e83875.

Deschoolmeester, V., Baay, M., Specenier, P., Lardon, F., and Vermorken, J. B. (2010). A review of the most promising biomarkers in colorectal cancer: one step closer to targeted therapy. *The Oncologist* **15**, 699–731.

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., Nobel, A. B., et al. (2006). Concordance among gene-expression–based predictors for breast cancer. *The New England Journal of Medicine* **355**, 560–569.

FDA. (2007). Guidance for industry: Clinical trial endpoints for the approval of cancer drugs and biologics. *Washington, DC, US Food and Drug Administration* pages 1–19.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine* **30**, 2867–2880.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Gaujoux, R. and Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367.

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., and Rosenthal, J. (2014). Clinical development success rates for investigational drugs. *Nature Biotechnology* **32**, 40–51.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2003). On optimality properties of the power prior. *Journal of the American Statistical Association* **98,** 204–213.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters* **31**, 651–666.

Kang, C., Janes, H., and Huang, Y. (2014). Combining biomarkers to optimize patient treatment recommendations. *Biometrics* **70**, 695–707.

Kelloff, G. J. and Sigman, C. C. (2012). Cancer biomarkers: Selecting the right drug for the right patient. *Nature Reviews Drug Discovery* **11**, 201–214.

Kim, E. S., Herbst, R. S., Wistuba, I. I., Lee, J. J., Blumenschein, G. R., Tsao, A., et al. (2011). The BATTLE trial: Personalizing therapy for lung cancer. *Cancer Discovery* **1**, 44–53.

Knox, S. S. (2010). From 'omics' to complex disease: A systems biology approach to gene-environment interactions in cancer. *Cancer Cell International* **10**, 355–364.

Lai, Y., Hayashida, M., and Akutsu, T. (2013). Survival analysis by penalized regression and matrix factorization. *The Scientific World Journal* **2013**, 1–11.

Lee, J. J., Gu, X., and Liu, S. (2010). Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials* **7**, 584–596.

Li, A., Walling, J., Ahn, S., Kotliarov, Y., Su, Q., Quezado, M., et al. (2009). Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. *Cancer Research* **69**, 2091–2099.

Maitournam, A. and Simon, R. (2005). On the efficiency of targeted clinical trials. *Statistics in Medicine* **24**, 329–339.

McShane, L. M., Cavenagh, M. M., Lively, T. G., Eberhard, D. A., Bigbee, W. L., Williams, P. M., et al. (2013). Criteria for the use of omics-based predictors in clinical trials. *Nature* **502**, 317–320.

Michaelis, L. C. and Ratain, M. J. (2006). Measuring response in a post-recist world: From black and white to shades of grey. *Nature Reviews Cancer* **6**, 409–414.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* **52**, 91–118.

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., et al. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Medicine* **19**, 619–625.

Song, X. and Pepe, M. S. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics* **60**, 874–883.

Wilkerson, M. D. and Hayes, D. N. (2010). Consensusclusterplus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68**, 1010–1018.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320.

Received February 2015. *Revised August* 2015.
Accepted October 2015.