Issues in developing mulitvariable models for treatment selection

Michael C Sachs and Lisa M McShane

February 19, 2016

Introduction

Omics technologies that generate a large amount of molecular data about biospecimens have the potential to provide information about a patient's disease characteristics above and beyond standard clinical and pathological features. By combining the information from a large amount of molecular features into a multivariable model, hereafter referred to as a *biomarker signature*, there is the opportunity to identify distinct subgroups of patients for whom treatment decisions can be personalized. A multivariable biomarker can guide the decisions to treat or not to treat and help identify the patients who are most likely to survive. The key challenge we address in this paper is to estimate the precise combination of features from a high dimensional molecular assay suited to the clinical context.

Terminology and Notation

A **biomarker signature** is a transformation of multiple individual features, typically molecular characterstics measured on a multiplex assay, to a one-dimensional space. Specifically, let X denote the set of p features under consideration. The signature is an unknown function $f(X) : \mathbb{R}^p \mapsto \mathbb{R}^1$. The signature may be continuous, take multiple discrete values, or be dichotomous.

Let S denote the development dataset, which includes X, an outcome Y, a treatment Z, and possibly other variables. S is a sample of size n from distribution $\mathcal{P}$ with domain $\mathcal{X}$. Let $\mathcal{F}$ be a mapping from $\mathcal{X}$ to the space of continuous functions with domain $\mathbb{R}^p$ and range $\mathbb{R}$, $\mathcal{D}$. Thus $\mathcal{F} : \mathcal{X} \mapsto \mathcal{D}$ denotes the process or algorithm through which a particular f is estimated. We do not place any other restrictions on $\mathcal{F}$, it could be a clustering approach, a regression approach, a combination of both, or something else entirely. We will use $\mathcal{F}$ to denote the manner in which f is estimated and will write $f \in \mathcal{F}$ to denote that f is estimated with the class of methods $\mathcal{F}$.

Let $\phi : \mathcal{D} \times \mathcal{X} \mapsto \mathbb{R}$ denote the statistic that quantifies the performance of the function f, such as predictive accuracy, mean squared error, or area under the receiver operating characteristic (ROC) curve (AUC). This is a function of both f and S. We are interested in estimating $E_{\mathcal{P}}[\phi_f(S)]$, which is the expected error under the data generation mechanism, for a particular $f \in \mathcal{F}$. This allows us to understand how the signature will perform on future observations generated from $\mathcal{P}$. We may also be interested in estimating $E_{\mathcal{P}}[\phi_f(S)]$ *for all* $f \in \mathcal{F}$, which is the generalization error for f generated using mechanism $\mathcal{F}$. This doesn't guide outside researchers as to which specific f to use, yet it is useful for development because it tells us how much signal is in the data. As shorthand we will write this as $E_{\mathcal{P}}[\phi_{\mathcal{F}}(S)]$

A signature that **reliably predicts** an outcome Y is one that has generalization error small enough for the clinical context. Such a signature may be useful for treatment selection, prognosis, or other type of clinical management.

Overview of biomarker signature development

The goal of the development phase of a biomarker signature is to provide a valid estimate of the performance of $f \in \mathcal{F}$. Optionally, one can provide a specification of f for others to use, and we want that particular f to be estimated as precisely as possible. Typically, a specific f is estimated using $\mathcal{F}$ based on some training data. This can be done using a variety of different methods. In recent years there have been an explosion in the literature of computational approaches to classification and prediction, and we do not intend to summarize them all here. Some excellent reviews of such approaches are Hastie, Friedman, and Tibshirani (2009) and Moons, Kengne, Woodward, et al. (2012). The main considerations in signature estimation are identifying the features to include, deciding what transformations to apply, determining how to combine the features, and whether to apply thresholds/cutoffs to the resulting signature.

Signatures can be estimated by identifying naturally occuring clusters, or intrinsic subtypes using X alone without regard to the outcome Y. This was the case with the PAM-50 gene signature [reference needed], and later, the intrinsic subtypes were shown to be strongly associated with clinical outcomes. Supervised learning techniques can also be used to identify signatures, as was the case with Oncotype DX (Paik et al. 2004). In this case, regression-based methods are used to estimate a model that is highly predictive for the outcome Y. In the case of Oncotype DX, the outcome in question was recurrence of breast cancer.

Another approach to identifying treatment-selection signatures is to use regression techniques to estimate a signature that has a strong interaction with a particular treatment. It is possible, and quite common in high-dimensional settings, to combine multiple approaches to estimating f. For instance, a data-reduction step by variable selection or clustering may be performed before doing regression analysis on the resulting components.

No matter what the particular model building method is, our main concern and focus of this paper is with obtaining a valid estimate of its performance, that is, a good estimate of $E_{\mathcal{P}}[\phi_{\mathcal{F}}(S)]$. This depends on the true signal in the data and the specific algorithm $\mathcal{F}$ used. An optional component of the development phase is to provide a specification of f for others to use on independent data or in clinical practice.

Biomarker signatures in clinical practice

A biomarker signature can inform clinical practice in a number of ways, regardless of how the signature was developed. A highly prognostic signature may identify a subpopulation that has such a good chance of long term survival, that they do not need to undergo treatment that carries risks and side effects such as chemotherapy. In the context of a specific therapy that targets a particular molecular pathway, a signature may identify a subpopulation that does not benefit from that therapy, thereby guiding the decision to treat or not. A signature that was developed to identify intrinsic molecular subtypes could be both prognostic and informative about the effectiveness of a therapy.

The key point is that signatures are useful if they can correctly and reliably classify patients into distinct subgroups for which different treatment decisions would be made. There are two distinct but related statistical concepts involved here: calibration or accuracy, and discrimination. Signatures are often optimized to be well-calibrated, that is, highly accurate for predicting outcomes. However if the signature does not separate a population into distinct subgroups, then it is unlikely to be informative enough to change clinical practice.

In the development process it is important to evaluate both of these statistical concepts. Furthermore, it is not trivial to assess each of these in a valid manner when the data are used to define the signature itself. We illustrate the potential for bias, and remedies, in our examples.

Janes et al. (2011), Janes et al. (2014), McShane and Polley (2013), Polley et al. (2013)

Data analysis example

Throughout this paper, we reanalyze data from Zhu et al. (2010). Briefly, the data of interest are from the JBR.10 trial, which was a randomized controlled trial of adjuvant vinorelbine/cisplatin (ACT) versus observation alone (OBS) in 482 participants with non small cell lung cancer (NSCLC). Of those 482 participants, 169 had frozen tissue collected, and of those samples, 133 (71 in ACT and 62 in OBS) had gene-expression profiling done using U133A oligonucleotide microarrays (Affymetrix, Santa Clara, CA).

The goal of the Zhu et al. (2010) paper was to identify a multi-gene signature that strongly predicts prognosis, and the hypothesis was that the poor prognosis subgroup would benefit more from ACT compared to the good prognosis subgroup. The signature was trained to predict disease specific survival. The annotated gene expression data and clinical information are available from the Gene Expression Omnibus (identifier: GSE14814, Edgar, Domrachev, and Lash (2002)). Batch effects were removed using the ComBat function in the sva R package (Leek et al. 2016) and then the gene expression values were centered by their means and scaled by their standard deviations. We reproduce their analysis while illustrating the key issues under discussion.

The authors of Zhu et al. (2010) present results that mainly focus on the discrimination ability of their estimated signature. They do that by demonstrating that the two risk subgroups predicted by their signature (high risk and low risk) have separation in their survival curves and that the hazard ratio for their signature is large and significant even when adjusting for other risk factors. They do not directly address calibration, that is, whether their signature accurately predicts survival times.

Issues

Recall that the main goal is to estimate $E_{\mathcal{P}}[\phi_{\mathcal{F}}(S)]$, the expected value of a given statistic on future observations for $f \in \mathcal{F}$. This can be estimated with the in-sample empirical estimate: $\hat{E}[\phi_f(S)] = \frac{1}{n}\sum_{i=1}^{n}\phi_f(s_i)$ for a particular f. However, if S is used to estimate f then the estimate will be biased due to overfitting, that is, $|E_{\mathcal{P}}[\phi_f(S)] - \hat{E}[\phi_f(S)]|$ will be large. This is because ϕ depends on f, and thus the statistic ϕ is being adaptively defined based on the observed data S, hence causing the overfitting.

In many cases during signature development, the statistic ϕ in question is a statistic that relates to calibration, such as classification accuracy, correlation or mean squared error. While a biomarker signature may accurately predict a clinical outcome, that does not neccessarily imply that the signature is clinically useful. As we mentioned before, two statistical critera are necessary for determining clinical usefulness: calibration and discrimination. To assess discrimination, a different statistic ϕ may be used, such as an odds ratio, hazard ratio, or difference in survival probabilities. Evaluation of ϕ is also subject to bias due to overfitting, a fact that is commonly overlooked in the medical literature.

In Zhu et al. (2010), ϕ was the hazard ratio comparing the high risk and low risk groups as defined by the JBL.10 signature f. The risk groups were determined by the signature f, which was estimated using S the same data that is then used to estimate the hazard ratio which is estimated to be 15. They go on to assess the discrimination of the JBL.10 signature in a series of independent data sets, in which they find hazard ratios around 2. We will reanalyze the dataset and illustrate some remedies for avoiding bias in determining the discrimination (an additionally the calibration) of this signature.

Altman and Royston (2000), Buyse (2007), Moons, Kengne, Woodward, et al. (2012), Moons, Kengne, Grobbee, et al. (2012)

Remedies to Overfitting

```
## [1] "Insert simulation results"
```

A traditional remedy to this problem of overfitting is the split sample approach. First, randomly partition S into the training sample S_t and the holdout sample S_h with sample sizes n_t and n_h, respectively. Then, S_h is hidden from the analyst while $\mathcal{F}$ is applied to S_t to estimate the signature function f_t. Then, with f_t, and therefore ϕ fixed, $\hat{E}[\phi_{f_t}(S_h)]$ is an unbiased estimator of both $E_{\mathcal{P}}[\phi_{f_t}(S)]$ and $E_{\mathcal{P}}[\phi_{\mathcal{F}}(S)]$. As a side-effect, the specific form of f_t that is fixed using the S_t partition can be reported as the function for others to use, therefore the aforementioned estimator is also an estimate of the error for that specific f_t. The drawback of the split-sample approach is that f is not estimated as precisely as it would be compared to using the entire dataset S and for the same reason neither is $E_{\mathcal{P}}[\phi_{\mathcal{F}}(S)]$. Dobbin and Simon (2011) investigate how to optimally split a dataset into training and holdout partitions.

Another approach to avoid overfitting is cross-validation, which is a resampling based approach. For a fixed integer k, which can be between 1 and n, we randomly select a partition of k observations from S, denoted S_k. Then f_{-k} is estimated and fixed by appling $\mathcal{F}$ on S_{-k} which is the the subset of S that is disjoint from S_k. Then, we get an estimate $\hat{E}[\phi_{f_{-k}}(S_k)]$ which is an unbiased estimate of $E_{\mathcal{P}}[\phi_{\mathcal{F}}(S)]$. This process is repeated K times to yield K estimates. Each of these estimates is unbiased, but noisy, because typically k is very small relative to n. Thus, we average over K to get a less noisy estimate. This process is called "leave k out" cross-validation. Note that for each partition that is selected, we obtain a new estimate of f, therefore the estimator is an estimator only for $E_{\mathcal{P}}[\phi_{\mathcal{F}}(S)]$. Typically, if a specific form for f is desired, it would be estimated using the entire dataset S.

A variation on the cross-validation approach is bootstrapping. In that case, a sample S_b of size n is sampled *with replacement* from S. Then f_b is estimated and fixed by applying $\mathcal{F}$ to S_b. The

error ϕ is estimated on the subset of S that is disjoint from S_b: S_{-b}. This process is repeated K times to yield K estimates. These K estimates are averaged to obtain the mean over the bootstrap replicates. Efron and Tibshirani (1997) suggest a variation, the 0.632 estimate:

$$\hat{E}^*[\phi_{\mathcal{F}}(S)] = .368\hat{E}[\phi_f(S)] + 0.632\hat{E}[\phi_{f_b}(S_{-b})],$$

where $\hat{E}[\phi_f(S)]$ is the naive estimate of ϕ_f using the entire dataset.

To illustrate the different properties of these estimates and how they deal with overfitting, we conduct a simulation study. Data were generated with n samples, each with a binary outcome Y with prevalence 0.3, and d features sampled from the standard normal distribution. This is the null case where no features are associated with Y. The signature estimation procedure entails a feature selection step, in which each feature is regressed against Y in a logistic regression model. The 25 features with the smallest p-values are selected for inclusion in a multivariable logistic regression model which defines our final signature.

We compare each of the three methods described above along with the biased approaches of using the full sample to select the features, followed by either fitting the model on a split sample, or using fitting the full model inside the cross validation step. Note that Zhu et al. (2010) used the latter approach, which explains the differences in our results. Our main interest is in comparing the bias and variance of the resulting estimates of $E_{\mathcal{P}}[\phi_{\mathcal{F}}(S)]$. In our simulation, we look at two different statistics, classification accuracy to assess calibration and the odds ratio for the outcome comparing the signature groups to assess discrimination.

Evaluating Calibration

Evaluating Discrimination and Clinical Benefit

Developing and validating a predictive biomarker for an effective treatment

Identify the non-responders and demonstrate that they do not benefit (hard)

Developing and validating a predictive biomarker for an ineffective treatment

Identify the responders and demonstrate that they do benefit (no quite as hard)

Conclusion

What I said and how I said it.

References

Altman, Douglas G, and Patrick Royston. 2000. "What Do We Mean by Validating a Prognostic Model?" *Statistics in Medicine* 19 (4). John Wiley & Sons, Ltd.: 453–73.

Buyse, Marc. 2007. "Towards Validation of Statistically Reliable Biomarkers." *European Journal of Cancer Supplements* 5 (5). Elsevier: 89–95.

Dobbin, Kevin K, and Richard M Simon. 2011. "Optimally Splitting Cases for Training and Testing High Dimensional Classifiers." *BMC Medical Genomics* 4 (1). BioMed Central Ltd: 31.

Edgar, Ron, Michael Domrachev, and Alex E Lash. 2002. "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository." *Nucleic Acids Research* 30 (1). Oxford Univ Press: 207–10.

Efron, Bradley, and Robert Tibshirani. 1997. "Improvements on Cross-Validation: The 632+ Bootstrap Method." *Journal of the American Statistical Association* 92 (438). Taylor & Francis: 548–60.

Hastie, T, J Friedman, and R Tibshirani. 2009. *The Elements of Statistical Learning*. Vol. 2. 1. Springer.

Janes, Holly, Marshall D Brown, Ying Huang, and Margaret S Pepe. 2014. "An Approach to Evaluating and Comparing Biomarkers for Patient Treatment Selection." *The International Journal of Biostatistics* 10 (1): 99–121.

Janes, Holly, Margaret S Pepe, Patrick M Bossuyt, and William E Barlow. 2011. "Measuring the Performance of Markers for Guiding Treatment Decisions." *Ann. Intern. Med.* 154 (4). Am Coll Physicians: 253–59.

Leek, JT, WE Johnson, HS Parker, EJ Fertig, AE Jaffe, and JD Storey. 2016. "Sva: Surrogate Variable Analysis. R Package Version 3.18.0."

McShane, Lisa M, and Mei-Yin C Polley. 2013. "Development of Omics-Based Clinical Tests for Prognosis and Therapy Selection: The Challenge of Achieving Statistical Robustness and Clinical Utility." *Clin. Trials* 10 (5). SAGE Publications: 653–65.

Moons, Karel GM, Andre Pascal Kengne, Diederick E Grobbee, Patrick Royston, Yvonne Vergouwe, Douglas G Altman, and Mark Woodward. 2012. "Risk Prediction Models: II. External Validation, Model Updating, and Impact Assessment." *Heart*. BMJ Publishing Group Ltd; British Cardiovascular Society, heartjnl–2011.

Moons, Karel GM, Andre Pascal Kengne, Mark Woodward, Patrick Royston, Yvonne Vergouwe, Douglas G Altman, and Diederick E Grobbee. 2012. "Risk Prediction Models: I. Development, Internal Validation, and Assessing the Incremental Value of a New (Bio) Marker." *Heart* 98 (9). BMJ Publishing Group Ltd; British Cardiovascular Society: 683–90.

Paik, Soonmyung, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L Baehner, et al. 2004. "A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer." *New England Journal of Medicine* 351 (27). Massachusetts Medical Society: 2817–26.

Polley, Mei-Yin C, Boris Freidlin, Edward L Korn, Barbara A Conley, Jeffrey S Abrams, and Lisa M McShane. 2013. "Statistical and Practical Considerations for Clinical Evaluation of Predictive

Biomarkers." *J. Natl. Cancer Inst.* 105 (22). Oxford University Press: 1677–83.

Zhu, Chang-Qi, Keyue Ding, Dan Strumpf, Barbara A Weir, Matthew Meyerson, Nathan Pennell, Roman K Thomas, et al. 2010. "Prognostic and Predictive Gene Signature for Adjuvant Chemotherapy in Resected Non–small-Cell Lung Cancer." *Journal of Clinical Oncology* 28 (29). American Society of Clinical Oncology: 4417–24.