Chapter 22
Targeted Methods for Biomarker Discovery

Catherine Tuglus, Mark J. van der Laan

The use of biomarkers in disease diagnosis and treatment has grown rapidly in recent years, as microarray and sequencing technologies capable of detecting biological signatures have become more effective research tools. In an attempt to create a level of quality assurance with respect to biological and more specifically biomarker research, the FDA has called for the development of a standard protocol for biomarker qualification (Food and Drug Administration 2006). Such a protocol would define “evidentiary” standards for biomarker usage in areas of drug development and disease treatment and provide a standardized assessment of a biomarker’s significance and biological interpretation. This is especially relevant for RCTs, where the protocol would prohibit the use of unauthenticated biomarkers to determine treatment regime, resulting in safer and more reliable treatment decisions (Food and Drug Administration 2006). Consequentially, identifying accurate and flexible analysis tools to assess biomarker importance is essential. In this chapter, we present a measure of variable importance based on a flexible semiparametric model as a standardized measure for biomarker importance. We estimate this measure with the TMLE.

Many biomarker discovery methods only measure the association between the marker and the biological outcome. However, a significant association is often difficult to interpret and does not guarantee that the biomarker will be a suitable and reliable drug candidate or diagnostic surrogate. This is especially true with genomic data, where genes are often present in multiple pathways and can be highly correlated amongst themselves. Applying association-based methods to these data will often lead to a long and ambiguous listing of biomarkers, which can be expensive to analyze.

Ideally, biomarker discovery analyses should identify markers that systematically affect the outcome through a biological pathway or mechanism, in other words, markers causally related to the outcome of interest. Once these markers are identified, they can be further analyzed and eventually applied as potential drug targets or...
Catherine Tuglus, Mark J. van der Laan

prognostic markers. Due to the complex nature of the human genome, this is not a straightforward task, and certain assumptions are required to identify a causal effect.

In general, causal effects are often difficult if not impossible to estimate correctly, especially based on high-dimensional and highly correlated genomic data structures. The required identifiability assumptions such as the time-ordering assumption, the randomization assumption, and the positivity assumption are often only fully realized in RCTs, making their utility in a standard protocol limited. However, measures that are causally interpretable in RCTs can still be biologically interpretable based on observational data as measures of importance.

Here, we present the typical representation of a causal effect as a potential measure of importance for a biomarker $A$:

$$\Psi(P_0)(a) = E_0[E_0(Y | A = a, W) - E_0(Y | A = 0, W)].$$

Given the observed data structure $O = (W, A, Y) \sim P_0$, this measure corresponds to the effect of a biomarker ($A$) on the outcome ($Y$), adjusting for confounders ($W$). Here, $A$ can represent a single biomarker or set of biomarkers. This chapter will focus on the univariate case. This measure can be estimated in semiparametric models for $P_0$, and with formal inference, using the TMLE.

In this chapter, we present the TMLE of the variable importance measure (VIM) above under a semiparametric regression model, which can accommodate continuous treatment or exposure variables often seen in biomarker analyses. We will primarily focus on its application to biomarker discovery. However, this method also has important applications to clinical trial data when the treatment is binary or continuous, and when one wishes to test for possible effect modification by baseline variables, for instance, treatment modified by biomarkers measured at baseline.

We demonstrate the efficacy and functionality of this VIM and its TMLE in a simulation study. The simulations provide a performance assessment of our estimated measure under increasing levels of correlation of $A$ with $W$. We show the accuracy with which the TMLE of the VIM can detect “true” variables from amongst increasingly correlated “decoy” variables. Additionally, we also evaluate the accuracy of three commonly used methods for biomarker discovery under the same conditions: univariate linear regression, lasso regression (Efron et al. 2004), and random forest (Breiman 1999, 2001a). We also apply the method in an application to a leukemia data set (Golub et al. 1999).

22.1 Semiparametric-Model-Based Variable Importance

Previous chapters have focused on the TMLE of the above VIM in a nonparametric model for variables $A$ that are discrete; for instance, $A$ might be an indicator for