2015

Prognostic and predictive directions for clinical trials

Debashis Ghosh, *University of Colorado Denver*

Prognostic and predictive directions for clinical trials

Debashis Ghosh

Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO, 80207, U.S.A.

debashis.ghosh@ucdenver.edu

Summary

In many clinical trials, treatment effects can be quite heterogeneous across subgroups so that individuals in different subgroups can receive different benefits of the treatment. This can be quite important for the purposes of clinical decision-making purposes. In this article, we introduce a general concept of risk score that is motivated by potential outcomes considerations. The concepts of prognostic and predictive directions for outcome data are defined. Their basis is in the dimension reduction (DR) literature and can also be motivated by conditional independence assumptions. Under some conditions, one can use existing methods from the DR literature to estimate the directions assuming a complete data structure. We show how to adapt the procedure with data that come from a randomized clinical trial. The methodology is illustrated with application to a set of colorectal cancer clinical trials.

Keywords: Causal effect; effect decomposition; link-free inference; model misspecification; personalized medicine; risk score.

1 Introduction

In many clinical trial settings, the overall treatment effect is the estimand of primary scientific interest, but it may not be appropriate for all the populations considered in the study. As a concrete example, we consider infection with hepatitis C (HCV), a viral infection that is estimated to be present in a chronic form in 3% of the world-wide population (Mohd Hanafiah et al., 2013) and whose incidence has risen due to increasing intravenous drug use and poor sterilization practices in hospital. There have been six subclassifications for HCV developed based on genotypic profiles, and recently, clinical trials have shown the benefit of using treatments involving sofosbuvir, both individually and in combination with other drugs, for treatment of HCV for genotypes 2 and 3 (Gane et al., 2013).

With the example in mind, developing methods for identification of appropriate patient subgroups for which the treatment might be of major benefit has become a topic of intense interest in the statistical literature. A related problem is that of determining qualitative treatment covariate interactions (Gail and Simon, 1985). For example, Follman and Proschan (1999) have developed a likelihood ratio testing method for this problem. More recently, Bonetti and Gelber (2004) have developed a graphical descriptive summary for identification of patient subgroups that is termed Subpopulation Treatment Effect Pattern Plot (STEPP) in which treatment/covariate interactions for survival data are discovered via permutation test-based procedures. More recently, Cai et al. (2011) developed a modelling strategy to identify subgroups of patients who would benefit from the treatment. Tree-based and related approaches (e.g., Kehl and Ulm, 2006; Su et al., 2007, 2008; Foster et al., 2011) for finding treatment subgroups have also been proposed.

In this work, inspired by ideas from causal inference and its links with dimension reduction methods (Ghosh, 2011), we develop a general risk modelling framework in which the concepts of prognostic and predictive directions are proposed. The idea is to posit potential outcomes for the subject under each of the possible treatments and to then model functionals of them. The prognostic and predictive directions represent two such quantities. In the case where the complete potential outcomes are available, we can then exploit dimension reduction methods in order to estimate the directions. One of the appealing features of such procedures is that the estimated direction is a risk score, i.e., a linear combination of the predictor variables. A

risk score is very intuitive and can be easily applied in practice. While we describe the concept within the potential outcomes framework in Section 2, in practice, the counterfactuals are not observed. Thus, we impute the outcomes using random forests (Breiman, 2001), a step that was also applied in the 'virtual twins' method of Foster et al. (2011).

The essence behind the procedure is that the treatment ignorability assumption of Rosenbaum and Rubin (1983) can be thought of as a conditional independence assumption involving the potential outcomes, the treatment and the covariates. Ghosh (2011) has shown that these types of conditional independence assumptions are naturally compatible with the conditional independence assumptions present in dimension reduction methods (Li, 1991; Cook, 1998). This implies that one can use dimension reduction procedures to estimate causal inferential estimands if the so-called strongly ignorable treatment assumption, defined in §2.1, holds. We use this fact in the current article in order to find the prognostic and predictive directions. In addition, we develop a multivariate regression framework in which the two sets of directions can be jointly estimated using a multivariate version of partial least squares (PLS, Dayal and MacGregor, 1998), which is an algorithm that is commonly used in chemometrics. The structure of the paper is as follows. In Section 2, we outline the background material on the potential outcomes framework as well as computation of the predictive and prognostic directions using DR methodology. Section 3 describes the generalization of results from Naik and Tsai (2000) that allow for the application of multivariate PLS algorithms for computation of the directions simultaneously. In Section 4, we provide methods for validating the estimated prognostic and predictive directions. Section 5 features an illustration of the techniques to data from 12 colorectal cancer studies we have previously analyzed (Ghosh et al., 2012). Some discussion concludes Section 6.

2 Proposed Framework

2.1 Potential outcomes framework and applications to risk modelling

We work within the potential outcomes framework of Rubin (1974) and Holland (1986). Assume that $(Y_i(0), Y_i(1), T_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, a random sample from the triple $(Y(0), Y(1), T, \mathbf{Z})$, where $(Y(0), Y(1))$ represents the counterfactuals, T denotes the treatment group, and Z is a p-dimensional vector of covariates, is observed for all subjects. Let T take the values $\{0, 1\}$

so that the treatment is binary. In practice, we cannot observe both potential outcomes; we merely use the setup to be able to define the prognostic and predictive directions.

As described in Rosenbaum and Rubin (1983), the standard assumption needed for causal inference is that

$$T \perp \{Y(0), Y(1)\}|\mathbf{Z}, \tag{1}$$

i.e. treatment assignment is conditionally independent of the set of potential outcomes given covariates. Rosenbaum and Rubin (1983) refer to (1) as the strongly ignorable treatment assumption; it allows for the estimation of causal effects.

Rosenbaum and Rubin (1983) proposed the use of the propensity score for estimation of causal effects in observational studies. The propensity score is defined as

$$e(\mathbf{Z}) = P(T = 1|\mathbf{Z}) \tag{2}$$

and represents the probability of receiving treatment as a function of covariates. Use of the propensity score leads to balance in covariates between the groups with $T = 0$ and $T = 1$. Statistically, this corresponds to the conditional independence of T and $\mathbf{Z}$ conditional on $e(\mathbf{Z})$ and is summarized in Theorem 1 of Rosenbaum and Rubin (1983). Given the treatment ignorability assumption in (1), it also follows by Theorem 3 of Rosenbaum and Rubin (1983) that treatment is strongly ignorable given the propensity score, i.e.

$$\mathbf{Z} \perp \{Y(0), Y(1)\}|e(\mathbf{Z}).$$

We now exploit the work of Ghosh (2011) and use further conditional independence assumptions from the dimension reduction literature. Assume that there exists a $p \times q$ matrix $\mathbf{A}$, $q \leq p$, such that treatment is conditionally independent of $\mathbf{Z}$, given $\mathbf{A}'\mathbf{Z}$. This can be expressed notationally as

$$T \perp \mathbf{Z}|\mathbf{A}'\mathbf{Z} \tag{3}$$

Assumption (3) is a crucial one for defining the estimand targetted by most dimension reduction methods. In particular, if $S(\mathbf{A})$ represents the subspace generated by the columns of $\mathbf{A}$, then the smallest subspace containing all possible spaces is known as the central subspace (Cook, 1998) and typically exists in most problems.

Combining assumptions (3) and (1), we have

$$T \perp \{Y(0), Y(1)\} | \mathbf{A}'\mathbf{Z} \tag{4}$$

so that the columns of $\mathbf{A}$ capture the essential information about the potential outcomes. These columns are what we term the *directions* in the outcome data. Note that (4) implies that

$$T \perp g(\{Y(0), Y(1)\}) | \mathbf{A}'\mathbf{Z} \tag{5}$$

for any function $g(y, z)$ whose domain is R^2 and whose range is R. We now define two such functions:

1. $g_1(y, z) = y + z$;
2. $g_2(y, z) = y - z$.

Of course, many other functions are possible, but we such focus on these two. One can view the first as representing the shared component in each potential outcome, while the second focuses on the contrast between the two treatments. We define the columns of $\mathbf{A}$ corresponding to g_1 as prognostic directions, while those corresponding to g_2 are termed the predictive directions.

2.2 Computation of Directions

As noted by Ghosh (2011), with the sequence of conditional assumptions being invoked in §2.1., one can then employ dimension reduction procedures in order to compute the prognostic and predictive directions. We first consider estimation of them in a univariate manner. The following high-level algorithm uses sliced inverse regression (Li, 1991), although other methods could also be used, such as SAVE (Cook and Weisberg, 1991) and MAVE (Xia et al., 2002):

A. Compute $Y_i^* \equiv g\{Y_i(1), Y_i(0)\}$ for subject i, $i = 1, \ldots, n$. Examples include g_1 and g_2 from the previous section.

B. Perform sliced inverse regression of Y_i^* on $\mathbf{Z}_i$ $(i = 1, \ldots, n)$ in order to estimate the directions (i.e., the columns of $\mathbf{A}$).

We recall that SIR, as initially introduced by Li (1991) required two assumptions for its validity. The first is termed the linearity condition and can be mathematically expressed as $E(\mathbf{Z}|\mathbf{A}'\mathbf{Z}) = \mathbf{A}'\mathbf{Z}$. The second is that the variance of $\mathbf{Z}$ given $\mathbf{A}'\mathbf{Z}$ satisfies a constant variance condition. The linearity condition is viewed as restrictive, as it is effectively satisfied by elliptically symmetric distributions. We do note that there have been efforts to lessen this assumption in several directions (e.g., Chiaromonte et al., 2002; Li et al., 2011; Lee et al., 2013).

As pointed out before, in practice, we cannot implement the high-level algorithm in the previous paragraph due to the inability to observe both potential outcomes. Instead of $\{Y_i(0), Y_i(1)\}$, we observe $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$. We thus modify the algorithm by including an imputation step:

1. Fit a random forests model (Breiman, 2001) for Y_i as a function of $T_i, \mathbf{Z}_i, T_i\mathbf{Z}_i$ and $(1 - T_i)\mathbf{Z}_i$, $i = 1, \ldots, n$. Such an algorithm will allow for computation of $(\hat{Y}_i(1), \hat{Y}_i(0))$ based on the observed covariates $\mathbf{Z}_i$, $i = 1, \ldots, n$.

2. Compute the variable $\tilde{Y}_i = g\{Y_i(1), Y_i(0)\}$ for subject i, $i = 1, \ldots, n$.

3. Sort $\tilde{Y}_1, \ldots, \tilde{Y}_n$ into increasing order and group them into d slices, termed $S_1, \ldots, S_d$.

4. Standardize the predictor observations as

$$\tilde{\mathbf{Z}}_i = \hat{\Sigma}^{-1/2}(\mathbf{Z}_i - \hat{\mu}), (i = 1, \ldots, n),$$

   where $\hat{\mu}$ and $\hat{\Sigma}$ are the sample mean and covariance matrices of $Z_1, \ldots, Z_n$.

5. Calculate within-slice stimates of sample mean $\bar{\mathbf{Z}}_j = \frac{1}{n_j}\sum_{i=1}^{n} I(\tilde{Y}_i \in S_j)\tilde{\mathbf{Z}}_i$, where $n_j = \sum_{i=1}^{n} I(\tilde{Y}_i \in S_j)$, $j = 1, \ldots, d$.

6. Estimate the population covariance matrix of $\mathbf{Z}$ as $\hat{\Theta} = \sum_{j=1}^{d} \frac{n_j}{n}\bar{\mathbf{Z}}_d\bar{\mathbf{Z}}_d'$.

7. Calculate the eigenvalues of $\hat{\Theta}$. The estimated directions are the eigenvectors corresponding to the largest eigenvalue.

We make several remarks about this algorithm. First, any imputation step can be used in the first step of the algorithm; this includes methods such as IVEWARE (Raghunathan et al.,

2001) and multiply imputed chained equations (van Buuren, 2012). The algorithm we use, random forests, belongs to the category of so-called ensemble methods. In random forests, many trees are generated, and predictions are averaged over all the trees. Second, as noted in Ghosh (2011), an alternative approach to SIR for estimating directions is partial least squares (PLS; Wold, 1975; Helland, 1988). Justification for PLS comes from the work of Naik and Tsai (2000), who showed that its performance was competitive with SIR in many instances. Third, step 1 corresponds to the imputation step that is need in algorithms such as in Foster et al. (2011). However, their subsequent steps are different from ours. Fourth, an implicit parameter in the algorithm is the number of slices we need to use in step 2. As Li (1991) argues, SIR is relatively insensitive to the number of slices used in the algorithm. Fifth, we should mention that a broad set of outcome variables can be handled by random forests in the first step. These include variables that are continuous, binary or unordered categorical. For censored variables, we use the suggestion of Keles and Segal (2002) and compute a first-stage martingale residual from a null model (i.e., one with no covariates). We then treat the residual as a continuous variable to be input into the random forests algorithm. Finally, we note that this algorithm, along with the other ones proposed in the paper, provide consistent estimates of the *relative* directions assuming the true underlying model holds. Mathematically, estimates are up to a constant of proportionality. This leads naturally to the use of rank-invariant measures for evaluation; one such measure is the C-index, whose use we study more in §4.

3 Simultaneous direction estimation

In this section, we develop an estimation procedure for joint estimation for prognostic and predictive directions based on a multivariate extension of the results of Naik and Tsai (2000). To do this, we consider a multivariate response for each subject that can be summarized in an $n \times m$ matrix $\mathbf{U}$ with a $n \times p$ matrix of covariates $\mathbf{X}$. Note that each column corresponds to a different outcome variable; to be concrete, we can take $m = 2$ where the first column is the prognostic function g_1 and the second column is the predictive function g_2. In particular,

we now formulate the following multivariate regression model:

$$\begin{pmatrix} U_1 \\ \vdots \\ U_m \end{pmatrix} = \begin{pmatrix} g_1(\beta_1' X, \epsilon_1) \\ \vdots \\ g_m(\beta_m' X, \epsilon_m), \end{pmatrix} \tag{6}$$

where g_j $(j = 1, \ldots, m)$ are monotonic functions in both arguments and $\epsilon_1, \ldots, \epsilon_m$ are random vectors representing the error distributions for the models. In (6), the parameters $\beta_1, \ldots, \beta_m$ specify the directions of interest. We also distinguish between X in (6), which specifies the covariate in the regression model, from $\mathbf{X}$, which specifies the data matrix which will be used for estimation. In the case where $m = 1$, model (6) reduces a generalized single-index model. We make the following assumptions.

Assumption 1. $\mathbf{X}$ has a multivariate normal distribution.

Assumption 2. The covariance matrices $n^{-1}\mathbf{X'X}$ and $n^{-1}\mathbf{X'U}$ converge in probability to limits Σ_{xx} and Σ_{xu}, respectively.

Assumption 3. Taken as an operator, the range of Σ_{xx} coincides with the range of Σ_{xu}.

Based on these three assumptions, we have the following result.

Theorem: Under Assumptions 1–3, the multivariate PLS estimator converges in probability to a constant times $(\beta_1, \ldots, \beta_m)$.

Proof: The multivariate PLS estimator can be expressed as $\hat{R}(\hat{R}' n^{-1}\mathbf{X'X}\hat{R})^{-1}\hat{R}' n^{-1}\mathbf{X'U}$, where $\hat{R}$ is the matrix derived from the Kryvlov sequence of matrices of $n^{-1}\mathbf{X'X}$ and $n^{-1}\mathbf{X'U}$. By assumption 2, this will converge to $R(R'\Sigma_{xx}R)^{-1}R'\Sigma_{xu}\beta^*$. By assumption 3, β^* will the in the space spanned by R so that $\Sigma_{xx}^{1/2}\beta^*$ will be in the space spanned by $\Sigma_{xx}^{1/2}R$. The statement of the theorem then follows.

The algorithm for multivariate PLS that we will use is the kernel algorithm that has been recently described in Dayal and MacGregor (1998). We assume that the columns of $\mathbf{U}$ and $\mathbf{X}$ are centered and scaled. Recall again that the PLS model formulation is given by

$$\begin{pmatrix} \mathbf{X} = \mathbf{TP'} + \mathbf{E} \\ \mathbf{U} = \mathbf{VQ'} + \mathbf{F} \end{pmatrix}, \tag{7}$$

where $\mathbf{T}$ and $\mathbf{V}$ are $n \times l$ matrices, and $\mathbf{P}$ and $\mathbf{Q}$ are the so-called locaing matrices corresponding to $\mathbf{X}$ and $\mathbf{U}$, respectively. The dimension of $\mathbf{P}$ is $p \times l$, while for $\mathbf{Q}$, it is $m \times l$. The matrices $\mathbf{E}$ and $\mathbf{F}$ are the error terms with entries being independent and identically distributed normal random variables with mean zero and variance σ^2.

The kernel algorithm proceeds as follows:

1. Compute the matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{U}$.

2. Let $b = 1$. Compute $\mathbf{q}_1$ as the eigenvector corresponding to the largest eigenvalue of $\mathbf{U}'\mathbf{X}'\mathbf{X}\mathbf{U}$. Then set $\mathbf{w}_b' = (\mathbf{X}'\mathbf{U})_b\mathbf{q}_b$ and rescale the entries of $\mathbf{w}_b$ to have unit norm. For $b = 1$, $(\mathbf{X}'\mathbf{U})_b = \mathbf{X}'\mathbf{U}$; for $b > 1$, its definition will be given in (8).

3. Compute $\mathbf{r}_b$. For $b = 1$, $\mathbf{r}_1 = \mathbf{w}_1$, while for $b > 1$,
$$\mathbf{r}_b = \mathbf{w}_b - \sum_{a=1}^{t-1} \mathbf{p}_a'\mathbf{w}_b\mathbf{r}_b.$$

4. Compute $\mathbf{t}_b = \mathbf{X}\mathbf{r}_b$, $\mathbf{p}_b = \mathbf{t}_b'\mathbf{X}/\mathbf{t}_b'\mathbf{t}_b$ and $\mathbf{q}_b' = \mathbf{r}_b'(\mathbf{X}'\mathbf{U})/\mathbf{t}_b'\mathbf{t}_b$.

5. Compute
$$(\mathbf{X}'\mathbf{U})_{b+1} = (\mathbf{X}'\mathbf{U})_b - \mathbf{p}_b\mathbf{q}_b'(\mathbf{t}_b'\mathbf{t}_b) \tag{8}$$

6. Repeat steps (2)-(5).

At the end, the regression coefficients are given by the outer product of the matrix consisting of $\mathbf{r}_b$ and that consisting of $\mathbf{q}_b$. This kernel algorithm has been implemented in the kernelpls.fit function that is available in the *pls* package (Mevik and Wehrens, 2007).

4 Evaluation of estimated directions

4.1 Prognostic Directions

Based on the estimated directions, we now have candidate risk scoring schemes that can be evaluated for their accuracy. Many standard methods of evaluation as described in Harrell (2001) can be used for evaluation of the rules that are generated here. In practical settings, we will wish to evaluate the utility of the estimated directions for defining appropriate rules. Again, one intuitively appealing feature of the directions is that they correspond to risk score estimators so that one can use the estimated directions to define a risk score $\hat{\beta}'X$. Based on such a computed score, one can then threshold the score to define a clinically meaningful rule.

In this article, we consider the use of the C-index (Pencina and D'Agostino, 2008) in order to evaluate the directions. It has interpretations both in terms of the area under the receiver

operating characteristic curve as well as a relationship to the Mann-Whitney statistic. Let S_i denote the predicted risk score using a direction, and let Y_i denote the associated outcome for that subject. We can estimate the C-index nonparametrically by

$$\hat{C} = \frac{1}{2}\left[\frac{\hat{\theta}_Z}{\hat{\theta}_Z + 1} + 1\right] \tag{9}$$

where

$$\hat{\theta} = \frac{\sum_{i<j} I\{(S_i - S_j)(Y_i - Y_j) > 0\}}{\sum_{i<j} I\{(S_i - S_j)(Y_i - Y_j) < 0\}}.$$

In words, (9) estimates the proportion of concordance between S and Y. The consistency and asymptotic normality of $\hat{C}$ follows from the theory of rank statistics.

4.2 Predictive Directions

Recall that the goal of the predictive direction is to define risk scores for who should be receiving treatment. What matters is whether the treatment assignment to the patient benefits the patient. To evaluate the predictive direction as a scoring rule, we need a training and testing set in which both studies are randomized and consist of the same treatments. In addition, outcome variables need to be measured in both studies. The proposal is related to one discussed in Vickers et al. (2007). To simplify the discussion, we will deal with the case of two treatment groups. The procedure works as follows:

(a). Estimate the predictive direction using the training dataset.

(b). Using the estimated direction, compute scores for all subjects in the test set.

(c). Based on the scores, determine which treatment each subject should receive in the test set.

(d). For the subjects whose predicted treatment match their randomized treatment in the test set, compare the outcomes between the two treatment groups.

We mention some points at this stage. First, we note that for step (b)., the outcome information in the test set is not used at all. Only the covariate information is used to compute the scores. The outcome information is needed in step (d). in order to compute the measure of treatment effect between the two groups. Note also that the fact that the test set also comes from a clinical trial is a necessary feature here. In step (d)., we will be excluding

two types of subjects in the test set: those who were predicted to have greatest benefit from one treatment group but were observed to receive the other one. Thus, we are performing a subgroup analysis in step (d). based on subjects in the test set whose predicted and actual treatment assignments are concordant. The randomization of treatment is necessary in order to ensure that the subgroup analysis will also be the same as the overall treatment effect.

5 Numerical example: Meta-analysis of colorectal cancer datasets

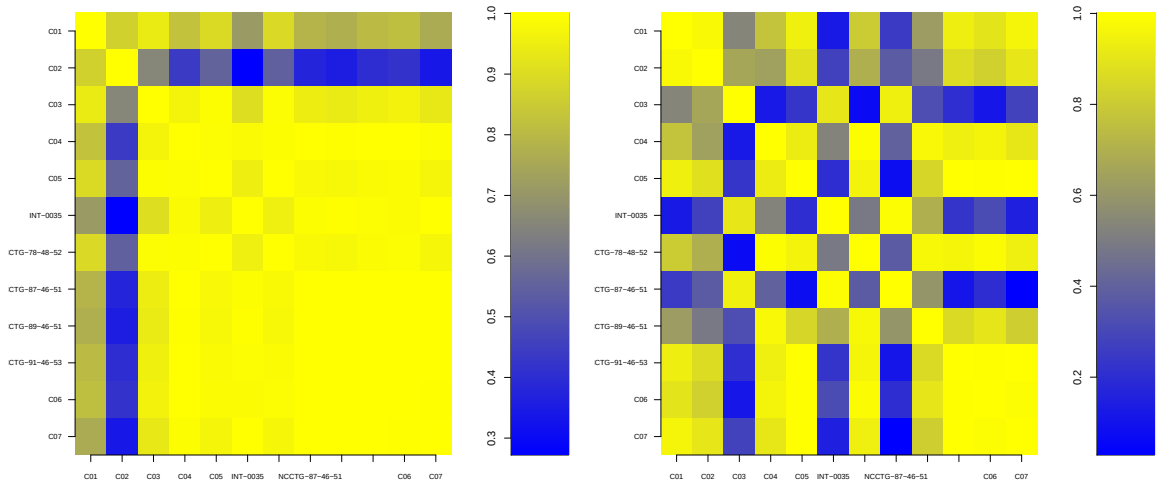
In this section, we will apply the proposed methods to data from a series of 12 adjuvant colon cancer studies that were evaluated for surrogate endpoints in Ghosh et al. (2012). Here, we will use data on treatment, age at baseline, stage and gender to explore prognostic and predictive directions with respect to survival time. Because we have data on 12 studies, we can furthermore explore the issue of whether or not the estimated directions show concordance across studies. Thus, we will estimate the predictive and prognostic directions on a study-specific basis to determine concordance.

Using SIR, we compute the prognostic directions and predictive directions and assess their concordance across the 12 studies. The results are shown in Figure 1. Numerical results can be One point to make here is that as alluded to in §2, the relative directions are estimated consistently, meaning that there is an equivalence up to the sign of the coefficients. To correct for this, we show heatmaps based on the absolute correlation coefficient in the estimated directions between studies. Note that we find strong agreement in the prognostic directions across studies. This is reflected in the prevalence of yellow in Figure 1a, which indicates high correlation. By contrast, there is less evidence of agreement in the predictive directions across studies, which is shown in Figure 1b.

In the Appendix, we show the analogous figure using the multivariate PLS algorithm described in §3. The same conclusions can be drawn there as well. The numerical estimates of the directions using both SIR and PLS can be found in the Appendix.

Next, we attempt to evaluate the classification ability of the estimated prognostic directions. To do this, we estimate directions using SIR on data from one study and use the remaining studies as the validation set. The results are given in Table 1 below.

Note that there is remarkable concordance between the prognostic directions estimated



(a) Prognostic directions

(b) Predictive directions

Figure 1: Plots of estimated prognostic directions (Figure 1a) and predictive directions (Figure 1b) using sliced inverse regression for the 12 colorectal cancer studies.

Training Data	C-index	C-index 95% CI
C01	0.593	(0.577,0.610)
C02	0.575	(0.559,0.592)
C03	0.608	(0.593,0.624)
C04	0.612	(0.596,0.627)
C05	0.605	(0.589,0.621)
INT-0035	0.604	(0.589,0.620)
NCCTG-78-48-52	0.604	(0.589,0.619)
NCCTG-87-46-51	0.612	(0.597,0.627)
NCCTG-89-46-51	0.609	(0.594,0.625)
NCCTG-91-46-53	0.612	(0.596,0.627)
C06	0.607	(0.591,0.622)
C07	0.611	(0.595,0.626)

Table 1: Results from computing C-indices for estimated prognostic directions. The column titled 'Training Data' denotes the colorectal cancer study that was used to estimate the prognostic directions. The second column denotes the C-index computed using the remaining 11 studies as a test dataset. The third column denotes a 95% confidence interval for the C-index computed using the normal approximation.

using the different studies in terms of the C-index. While the C-index values in Table 1 are dependent due to overlapping test datasets, the estimates and confidence intervals are very similar. The one exception is study C02, which has a lower C-index relative to the other eleven studies. The other point to note is that the C-indices, which can be interpreted as area under a receiver operating characteristic curve, are fairly moderate in nature. In the Appendix, we also applied the multivariate PLS algorithm and found that the C indices were quite concordant but systematically lower that what is given in Table 1.

Next, we evaluated the identified predictive directions using the proposed methodology. The results from using the sliced inverse regression-based procedure is shown in Table 2. While the studies suggest that the predictive directions lead to some benefit of selecting

Training Data	HR	95% CI
C01	0.84	(0.69, 1.02)
C02	0.64	(0.45,0.91)
C04	0.94	(0.81,1.10)
NCCTG-78-48-52	1.00	(0.93,1.09)
NCCTG-89-46-51	0.71	(0.53,0.95)
NCCTG-91-46-53	0.90	(0.80,1.00)
C06	0.90	(0.65,1.23)
C07	0.75	(0.68,0.82)

Table 2: Results from computing hazard ratios for estimated predictive directions. The column titled 'Training Data' denotes the colorectal cancer study that was used to estimate the prognostic directions. The second column denotes the hazard ratio computed using the remaining 11 studies as a test dataset based on the procedure outlined in §The third column denotes a 95% confidence interval for the hazard ratio computed using the normal approximation.

patients in a consistent, we mention two things at this point. First, for four studies (C03, C05, INT-0035, NCCTG-87-46-51), we were unable to compute a hazard ratio. This was due to the fact that the estimated predictive directions did not lead to predicted treatment assignments that were concordant with the observed treatment assignments in the test datasets. The joint PLS algorithm of Section 3 actually performed worse, with only two studies yielding estimates of the treatment effect (data not shown). These analyses suggest that there is not a predictive direction that is tremendously generalizable from the series of 12 colorectal cancer trials.

6 Discussion

In this article, we have developed the concept of risk modelling using prognostic and predictive directions. These concepts have their roots in two fields of statistics: dimension reduction methodology and causal inference. Conditional independence is the key foundational concept that forms the basis of these directions.

As with the virtual twins method of Foster et al. (2011), the procedure requires imputing potential outcomes for each subject under the possible treatment assignments. In the parlance of Cai et al. (2011), this is a working model that is used in the direction estimation algorithm. Because our final evaluations are performed on a test set, we can argue as in Cai et al. (2011) that the ultimate estimands of interest do not rely on proper specification of the working model.

Section 3 provided a multivariate extension of the results of Naik and Tsai (2000). Theorem 1 requires the stringent assumption of multivariate normality. As alluded to in the paper, there have been attempts to relax this type of assumption in the literature on DR (e.g., Li et al., 2011; Lee et al., 2013). We are currently working on generalizations of these results to the multivariate outcome setting.

Acknowledgments

This research is supported by National Institutes of Health grant CA129102.

References

Bonetti, M. and Gelber, R. D. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* **5**, 465 –81. PubMed PMID: 15208206.

Breiman, L. (2001). Random Forests. *Machine Learning* **45**, 5 – 32.

Byar, D. P. (1985). Assessing apparent treatment-covariate interactions in randomized clinical trials. *Statistics in Medicine* **4**, 255-263.

Cai, T., Tian, L., Wong, P. H. and Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* **12**, 270 – 282.

Chiaromonte, F., Cook, R. D., and Li, B. (2002). Sufficient dimensions reduction in regressions with categorical predictors. *Annals of Statistics* **30**, 475–497. doi:10.1214/aos/1021379862. http://projecteuclid.org/euclid.aos/1021379862.

Cook, R. D. (1998). *Regression Graphics*. Wiley,New York.

Cook, R. D. and Weisberg, S. (1991). Discussion of 'Sliced inverse regression' by Li. Journal of the American Statistical Association **86**, 328 – 332.

Dayal, B. S. and MacGregor, J. F. (1997), Improved PLS algorithms. J. Chemometrics, 11: 7385.

Follman, D. A. and Proschan, M. A. (1999). A multivariate test of interaction for use in clinical trials. *Biometrics* **55**, 1151 – 1155.

Foster, J. C., Taylor, J. M. and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine* **30**, 2867 – 80. doi: 10.1002/sim.4322. Epub 2011 Aug 4. PubMed PMID: 21815180; PubMed Central PMCID: PMC3880775.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* **41**, 361 – 372.

Gane, E. J., Stedman, C. A., Hyland, R. H., Ding, X., Svarovskaia, E., Symonds, W. T., Hindes, R. G., Berrey, M. M. (2013). Nucleotide polymerase inhibitor sofosbuvir plus ribavirin for hepatitis C. *New England Journal of Medicine* **368**, 34 –44.

Ghosh, D. (2008). Semiparametric inference for surrogate endpoints with bivariate censored data. *Biometrics* **64**, 149 – 156.

Ghosh, D. (2011). Propensity score modelling in observational studies using dimension reduction methods. Statistics and Probability Letters 81, 813 – 820.

Ghosh, D., Taylor, J. M. G. and Sargent, D. J. (2012). Meta-analysis for surrogacy: accelerated failure time modelling and semi-competing risks (with discussion), *Biometrics* **68**, 226 – 247.

Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in Statistics – Simulation and Computation* **17**, 581 – 607.

Holland, P. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81**, 945 – 970.

Kehl, V. and Ulm, K. (2006). Responder identification in clinical trials with censored data. *Comput Stat Data Anal* **50**, 1338 – 1355.

Keles, S. and Segal, M. R. (2002). Residual-based tree-structured survival analysis. *Statistics in Medicine* **21**, 313 – 326.

Lee, K.-Y.;,Li, B., and Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Annals of Statistics* **41**, 221–249.

Li, B., Artemiou, A., Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Annals of Statistics* **39**, 3182–3210.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86**, 316 – 342.

Mevik, B.-H. and Wehrens, R. (2007); The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software* **18**, 1 – 24.

Mohd Hanafiah, K., Groeger, J., Flaxman, A. D. and Wiersma, S. T. (2013), Global epidemiology of hepatitis C virus infection: New estimates of age-specific antibody to HCV seroprevalence. Hepatology, 57: 13331342.

Naik, P. A. and Tsai, C. L. (2000). Partial least squares estimator for single-index models. *J. R. Statist. Soc. B* **62**, 763 – 771.

Pencina, M. J., and D'Agostino, R. B. (2004). Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine* **23**, 2109 – 2123.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. V. and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology* **27**, 85 – 95.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41 – 55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688 – 701.

Su, X., Tsai, C. L., Wang, H., Nickerson, D. M. and Bogong, L. (2009). Subgroup analysis via recursive partitioning. *J Machine Learning Research* **10**, 141 – 158.

Su X, Zhou, T., Yan, X., Fan, J. and Yang, S. (2008). Interaction trees with censored survival data. *Intl J Biostatistics* **4**, article 2.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data.* Boca Raton, FL: Chapman & Hall/CRC Press.

Vickers, A. J., Kattan, M. W. and Sargent, D. (2007). Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials* **8**, 14.

Wold, H. (1975). Path models with latent variables: The NIPALS approach. In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, and V. Capecchi (Eds.), Quantitative sociology: International perspectives on mathematical and statistical modeling (pp. 307 – 357). New York, NY: Academic.

Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space (with discussion). *Journal of the Royal Statistical Society Series B* **64**, 363 – 410.

Numerical Results for the Colorectal Cancer Data

	tx	age	stage	gender
C01	-0.020	0.057	0.800	0.597
C02	-0.012	-0.068	-0.603	-0.795
C03	0.008	0.017	0.952	0.307
C04	0.004	0.012	0.993	0.121
C05	-0.004	-0.015	-0.965	-0.260
INT-0035	0.003	0.006	0.997	-0.076
NCCTG-78-48-52	-0.017	0.005	0.972	0.234
NCCTG-87-46-51	-0.016	0.001	0.999	0.050
NCCTG-89-46-51	-0.013	0.012	1.000	-0.010
NCCTG-91-46-53	-0.005	0.005	0.998	0.063
C06	-0.002	0.017	0.995	0.096
C07	0.006	-0.013	-1.000	0.003

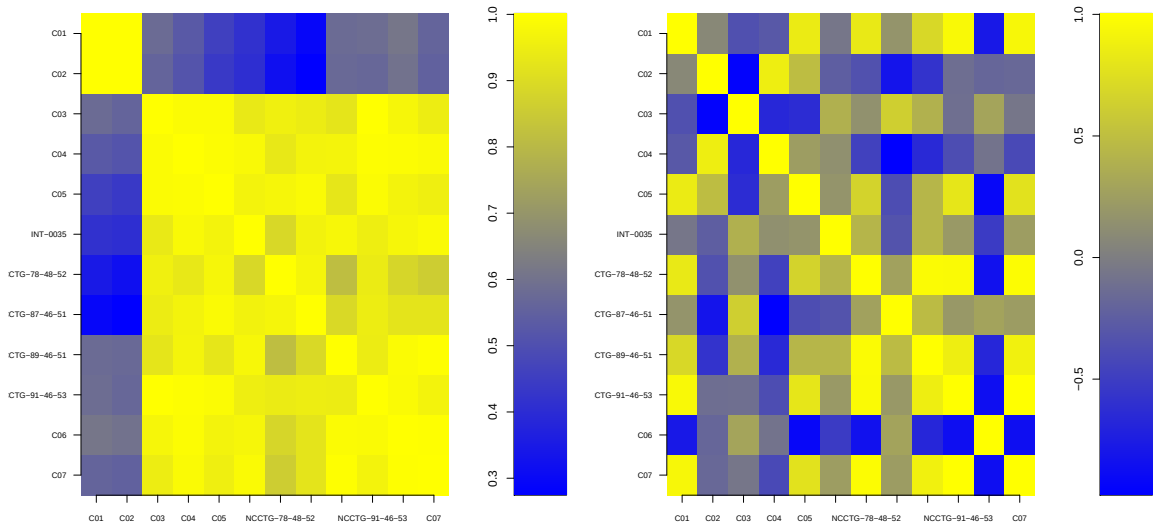
	tx	age	stage	gender
C01	-0.052	0.026	-0.238	0.969
C02	0.061	-0.039	-0.157	0.985
C03	-0.025	-0.010	-0.626	-0.780
C04	-0.065	0.121	-0.689	0.712
C05	-0.013	-0.005	-0.603	0.798
INT-0035	0.017	-0.021	-0.854	-0.519
NCCTG-78-48-52	-0.019	-0.026	0.740	-0.672
NCCTG-87-46-51	-0.197	0.125	0.968	0.095
NCCTG-89-46-51	0.019	0.010	-0.980	0.195
NCCTG-91-46-53	-0.010	-0.026	0.621	-0.784
C06	-0.038	0.001	0.659	-0.752
C07	-0.017	-0.010	0.552	-0.834

	tx	age	stage	gender
C01	-0.006	0.250	0.169	0.120
C02	-0.001	0.280	0.176	0.128
C03	0.004	0.082	0.185	0.062
C04	0.001	0.088	0.220	0.038
C05	0.001	0.067	0.209	0.047
INT-0035	-0.000	0.064	0.230	-0.012
NCCTG-78-48-52	-0.000	0.069	0.306	0.120
NCCTG-87-46-51	-0.002	0.010	0.121	0.012
NCCTG-89-46-51	-0.005	0.101	0.211	-0.007
NCCTG-91-46-53	-0.004	0.076	0.178	0.048
C06	-0.004	0.095	0.189	0.025
C07	-0.001	0.052	0.111	0.003

	tx	age	stage	gender
C01	-0.002	0.030	-0.013	0.061
C02	0.000	-0.004	0.008	0.009
C03	-0.003	0.004	-0.025	-0.047
C04	-0.000	-0.022	-0.004	-0.001
C05	0.001	0.000	-0.011	0.020
INT-0035	0.002	-0.036	-0.051	-0.033
NCCTG-78-48-52	0.003	0.046	-0.070	0.044
NCCTG-87-46-51	-0.007	0.087	0.020	-0.006
NCCTG-89-46-51	0.003	0.041	-0.061	0.016
NCCTG-91-46-53	0.001	0.016	-0.014	0.024
C06	-0.003	0.003	0.016	-0.015
C07	0.000	0.014	-0.015	0.018

	cindex2	V2	V3
C01	0.554	0.537	0.571
C02	0.551	0.534	0.567
C03	0.563	0.546	0.580
C04	0.564	0.546	0.581
C05	0.565	0.548	0.583
INT-0035	0.568	0.551	0.584
NCCTG-78-48-52	0.574	0.558	0.590
NCCTG-87-46-51	0.587	0.571	0.603
NCCTG-89-46-51	0.559	0.542	0.576
NCCTG-91-46-53	0.563	0.547	0.580
C06	0.560	0.543	0.576
C07	0.562	0.545	0.578

Table 3: Results from computing C-indices for estimated prognostic directions. Same as Table 1 except for the fact that the prognostic directions were computed using the multivariate PLS algorithm given in §3.



(a) 1a

(b) 1b

Figure 2: plots of....

Plots of concordance across studies for prognostic and predictive directions