Bioinformatics, 33(15), 2017, 2372–2378
doi: 10.1093/bioinformatics/btx174
Advance Access Publication Date: 24 March 2017
Original Paper

Data and text mining

Mining heterogeneous causal effects for personalized cancer treatment

Weijia Zhang[1,*], Thuc Duy Le[1,2], Lin Liu[1], Zhi-Hua Zhou[3] and Jiuyong Li[1]

[1]School of Information Technology and Mathematical Sciences, [2]Centre for Cancer Biology, University of South Australia, Adelaide 5000, Australia and [3]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on November 4, 2016; revised on March 19, 2017; editorial decision on March 20, 2017; accepted on March 20, 2017

Abstract

Motivation: Cancer is not a single disease and involves different subtypes characterized by different sets of molecules. Patients with different subtypes of cancer often react heterogeneously towards the same treatment. Currently, clinical diagnoses rather than molecular profiles are used to determine the most suitable treatment. A molecular level approach will allow a more precise and informed way for making treatment decisions, leading to a better survival chance and less suffering of patients. Although many computational methods have been proposed to identify cancer subtypes at molecular level, to the best of our knowledge none of them are designed to discover subtypes with heterogeneous treatment responses.

Results: In this article we propose the Survival Causal Tree (SCT) method. SCT is designed to discover patient subgroups with heterogeneous treatment effects from censored observational data. Results on TCGA breast invasive carcinoma and glioma datasets have shown that for each subtype identified by SCT, the patients treated with radiotherapy exhibit significantly different relapse free survival pattern when compared to patients without the treatment. With the capability to identify cancer subtypes with heterogeneous treatment responses, SCT is useful in helping to choose the most suitable treatment for individual patients.

Availability and Implementation: Data and code are available at https://github.com/WeijiaZhang24/SurvivalCausalTree.

Contact: weijia.zhang@mymail.uinsa.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Choosing the most appropriate treatment is of great importance in the battle against cancer. Although many advanced techniques have been developed to treat the dreaded disease, there has been no consensus about which treatment is most suitable when it comes to a particular patient with a specific type of cancer (Hayden, 2009).

Recent research has shown that rather than being a single disease, cancer involves different subtypes characterized by different sets of molecules (Perou *et al.*, 2000; The Cancer Genome Atlas Network, 2012), and different subtypes often respond heterogeneously towards the same treatment (Goldhirsch *et al.*, 2011). For example, estrogen receptor (ER) positive breast cancer subtype responds to hormone therapy, and the human epidermal growth factor receptor 2 (HER2) positive subtype will most likely respond to chemotherapy.

Unfortunately, our current understanding of cancer subtypes at the molecular level is far from complete. Decisions for cancer treatments are almost entirely based on clinical factors, disease stages, morphology based pathological indicators and types of surgery rather than expression profiles.

Treating cancer patients based on their molecule subtypes has important clinical impact. In breast cancer, more than 50% of the

patients have received radiotherapy (RT) as treatment, equating to over half a million patients worldwide each year. Although RT is effective for many patients, not all patients have benefited from the treatment as evidenced by distant metastatic spread and local recurrence (Bellon, 2015). Prediction of individual responses will allow a stratified approach of applying the treatment, saving those unsuitable patients from the associated iatrogenesis.

Many computational methods have been proposed to identify molecular cancer subtypes. Efron (1988), Goeman (2009), and Park and Hastie (2007) proposed techniques based on L1-regularization and COX proportional hazard model (Cox, 1972) to identify important genes that are related to patient survival time. The Cancer Genome Atlas Network (2012), Monti (2003), Wilkerson and Hayes (2010) and Shen *et al.* (2009) exploited the idea of clustering to discover disease subgroups at a molecule level. Bair and Tibshirani (2004) and Koestler *et al.* (2010) combine Cox regression with recursive partitioned mixture model (RPMM) to form a semi-supervised approach for identifying disease subtypes.

However, these methods do not answer the critical question of whether the identified subtypes show heterogeneous responses toward a treatment. In other words, the survival outcome of treated and untreated patients may not be significantly different for each of their identified subtypes.

Identifying subtypes with heterogeneous treatment effects is a *causal* problem. In order to estimate the effect of a treatment, one has to answer the *counterfactual* question: what would the survival outcome of a treated patient be, if he had not accepted the treatment; and what would the outcome of an untreated patient be, if he had been treated? The fundamental challenge is that for each patient only one of the two potential outcomes can be observed.

Heterogeneous treatment effect analysis has attracted increasing attention (Athey and Imbens, 2016; Doove *et al.*, 2013; Imai and Ratkovic, 2013; Kang *et al.*, 2012; Su *et al.*, 2009). These approaches utilize recursive partitioning to discover the desired subgroups. However, two limitations prevent these methods from being applied to our task. Firstly, existing methods are only applicable to data without censoring, unfortunately the outcomes in medical studies are seldom complete but almost always censored. Secondly, many of the existing methods are designed to analyze data with randomized treatment assignment, directly applying them to observational data will cause estimation bias since the treatment assignment not randomized (Imbens and Rubin, 2015).

In this article we extend the causal tree (Athey and Imbens, 2016) method to censored survival data and propose the Survival Causal Tree (SCT) method. Utilizing gene expression profiles and censored survival outcomes, SCT is able to identify molecular cancer subtypes with heterogeneous treatment effects towards the treatment of interest.

By analyzing the causal relationships between gene expressions and treatment responses, the subtypes identified by SCT can be used to predict the potential treatment effects of unseen patients. Our results on both TCGA breast invasive carcinoma and glioma datasets (The Cancer Genome Atlas Network, 2012) have shown that not only the subgroups identified from the training data have heterogeneous treatment effects, but also the survival patterns are similar in the test data.

Since the output of SCT is a tree model, the identified disease subgroups are readily interpretable. Each subgroup is defined by only a handful of genes, which is convenient for future clinical applications. The method can be used to help oncologists in determining the best treatment strategy for each individual cancer patient. Figure 1 presents the work flow of how SCT can be applied.

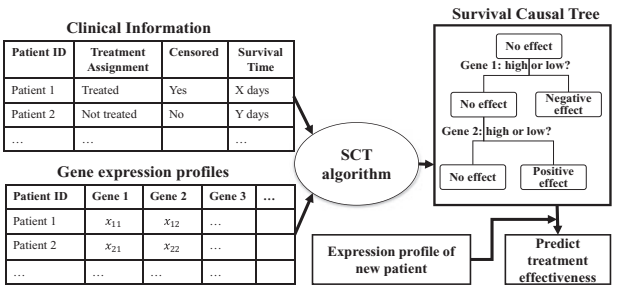

Fig. 1. Workflow of the application of Survival Causal Tree (SCT). SCT utilizes matched clinical information and gene expression profiles to train a causal tree model. The trained model can be used to predict whether a treatment should be applied to unseen patients

2 Materials and methods

2.1 Estimating treatment effects from data with censoring

First we introduce the necessary preliminaries for causal studies with fully observed outcomes, then we extend the discussion to censored outcomes.

Let $W_i \in \{0,1\}$ denote the treatment assignment, with $W_i = 1$ indicating the ith unit is treated and $W_i = 0$ indicating the opposite. Let Y_i be the observed survival outcome of interest, and $\mathbf{X}_i = \{X_{i1}, \ldots, X_{ip}\}$ be a vector describing the patient's gene expressions. The observed dataset consists of i.i.d. samples $(Y_i, W_i, \mathbf{X}_i)$, for $i = 1, \ldots, N$. For the sake of simplicity, the subscript i is dropped when the context is clear.

Let $Y^{(W)}$ denote the potential survival time of a patient if he had received treatment W, the observed survival time can be described as $Y = WY^{(1)} + (1-W)Y^{(0)}$. Note that each patient Y is associated with two potential outcome $Y^{(1)}$ and $Y^{(0)}$, but only one of them can be observed as Y. The average treatment effect of the population is defined as the expected survival time of all patients if they were treated minus their expected potential survival time if they were not treated:

$$\tau = \mathbb{E}\left[Y^{(1)}\right] - \mathbb{E}\left[Y^{(0)}\right]. \tag{1}$$

Since each patient can only receive or not receive the treatment, Equation 1 is counterfactual and thus cannot be directly estimated. If the treatment assignment is completely randomized, i.e. $\left(Y^{(0)}, Y^{(1)}\right) \perp\!\!\!\perp W$), the average treatment effect can be estimated using $\tau = \mathbb{E}(Y|W=1) - \mathbb{E}(Y|W=0)$. Therefore an unbiased estimator of average treatment effect for data with randomized treatment assignment can be given as:

$$\widehat{\tau}_{rct} = \frac{\sum_{i=1}^{N} W_i \cdot Y_i}{\sum_{i=1}^{N} W_i} - \frac{\sum_{i=1}^{N} (1-W_i) \cdot Y_i}{\sum_{i=1}^{N} (1-W_i)} \tag{2}$$

For observational data, the treatment assignment is usually not completely randomized therefore treated patients may not be comparable with untreated patients. To estimate treatment effects from observational data, Imbens and Rubin (2015) introduces the unconfoundedness assumption:

Assumption 1. (Unconfoundedness) $W \perp\!\!\!\perp \left(Y^{(0)}, Y^{(1)}\right) | \mathbf{X}$.

The assumption ensures that for all samples the treatment assignment W is independent of the outcome Y when the expression profiles $\mathbf{X}$ are considered. With this assumption, propensity score (Rosenbaum and Rubin, 1983) can be used with inverse probability

weighting (Seaman and White, 2013; Zhang and Zhou, 2014) to obtain an unbiased estimation of average treatment effect. The propensity score is defined as the probability of treatment assignment conditional on the covariates:

$$\pi(\mathbf{X}) = Pr(W = 1|\mathbf{X}) \tag{3}$$

Utilizing Assumption 1 and the fact that $W(1- W) = 0$, we have

$$\begin{aligned}\mathbb{E}[W \cdot Y/\pi(\mathbf{X})] &= \mathbb{E}\left[\frac{I(W=1)\cdot Y^{(1)}}{\pi(\mathbf{X})}\right] \\ &= \mathbb{E}\left\{\mathbb{E}\left[\frac{I(W=1)\cdot Y^{(1)}}{\pi(\mathbf{X})}|Y^{(1)},\mathbf{X}\right]\right\} \\ &= \mathbb{E}\left\{\frac{Y^{(1)}}{\pi(\mathbf{X})}\cdot\mathbb{E}\left[I(W=1)|Y^{(1)},\mathbf{X}\right]\right\} \\ &= \mathbb{E}\left[Y^{(1)}\right].\end{aligned} \tag{4}$$

Similarly $\mathbb{E}[(1 - W) \cdot Y/(1 - \pi(\mathbf{X}))] = \mathbb{E}\left[Y^{(0)}\right]$. The average treatment effect for observational data can be estimated as (Lunceford and Davidian, 2004):

$$\widehat{\tau}_{ob} = \frac{\sum_{i=1}^{N}\frac{W_i\cdot Y_i}{\pi(\mathbf{X}_i)}}{\sum_{i=1}^{N}\frac{W_i}{\pi(\mathbf{X}_i)}} - \frac{\sum_{i=1}^{N}\frac{(1-W_i)\cdot Y_i}{(1-\pi(\mathbf{X}_i))}}{\sum_{i=1}^{N}\frac{1-W_i}{(1-\pi(\mathbf{X}_i))}}. \tag{5}$$

The denominators of Equation 5 come from the fact that $\mathbb{E}[W/\pi(\mathbf{X})] = 1$ and $\mathbb{E}[(1 - W)/(1 - \pi(\mathbf{X}))] = 1$.

Now we extend our discussion towards data with censoring. In medical studies the observation of outcomes are almost always not complete because the limited time of the follow-up period. For example, the relapse free survival time of a cancer patient is only completely observed if the event of interest (i.e. relapse of cancer) occurs within the follow-up period, otherwise the outcome is considered as censored.

Formally, let W denote the treatment indicator, C denote the censoring time. Let Y denote the realized survival outcome and $Y^{(j)}$ denote the potential survival time. Instead of observing Y, one observes $Q = WQ^{(1)} + (1 - W)Q^{(0)}$ where $Q^{(j)} = \min\{Y^{(j)}, C\}$, as well as the complete case indicator $\delta = W\delta^{(1)} + (1 - W)\delta^{(0)}$, where $\delta^{(j)} = I(C \geq Y^{(j)})$. The censored survival data can be described as i.i.d. random vectors $(Q, \delta, \delta Y, W, \mathbf{X})$. The focus is using this data to estimate the average treatment effect $\tau_{censor} = \mathbb{E}\left[Y^{(1)}\right] - \mathbb{E}\left[Y^{(0)}\right]$.

Similar to data without censoring, we assume the treatment assignment is independent of censored and uncensored outcomes given the expression profiles, the unconfoundedness assumption is extended to:

Assumption 2. $\left(Y^{(0)}, Y^{(1)}, Q^{(0)}, Q^{(1)}\right) \perp\!\!\!\perp W|\mathbf{X}$.

In addition, we assume that censoring is independent of the outcomes and covariates when treatment assignment is considered:

Assumption 3. $\left(Y^{(0)}, Y^{(1)}, Q^{(0)}, Q^{(1)}, \mathbf{X}\right) \perp\!\!\!\perp C|W$.

Let $K_p(u) = Pr(C \geq u|W)$ denote the treatment specific censoring distribution, for treated samples we have:

$$\begin{aligned}\mathbb{E}\left[\frac{W\delta Y}{\pi(\mathbf{X})K_1(Q)}\right] &= \mathbb{E}\left\{\mathbb{E}\left[\frac{W\delta Y^{(1)}}{\pi(\mathbf{X})K_1(Q)}|Q^{(0)},Q^{(1)},Y^{(1)},\mathbf{X},W\right]\right\} \\ &= \mathbb{E}\left\{\frac{WY^{(1)}}{\pi(\mathbf{X})\cdot K_1(Q)}\mathbb{E}\left[I(C\geq Q)|Q^{(0)},Q^{(1)},Y^{(1)},\mathbf{X},W\right]\right\} \\ &= \mathbb{E}\left[\frac{WY^{(1)}}{\pi(\mathbf{X})}\right] = \mathbb{E}\left[Y^{(1)}\right].\end{aligned} \tag{6}$$

The first equation uses $W^2 = W$, $W(1 - W) = 1$ and the law of total expectation. The second equation is obtained by utilizing the assumptions. The inner expectation of the third equation is given as $K_1\left(Q^{(1)}\right)I(W = 1) + K_0\left(Q^{(1)}\right)I(W = 0)$, and equals to $K_1\left(Q^{(1)}\right)I(W = 1)$ when multiplied by W. The derivation of the last equation is as same as that of observational data without censoring.

Similarly, for untreated samples $\mathbb{E}\left[Y^{(0)}\right] = \mathbb{E}\{[\delta \cdot Y \cdot (1 - W))]/[(1 - \pi(X)) \cdot K_0(Q)]\}$. Therefore the average treatment effect for censored survival data can be estimated by (Anstrom and Tsiatis, 2001):

$$\widehat{\tau}_{censor} = \frac{\sum_{i=1}^{n}\frac{W_i\cdot\delta_i\cdot Y_i}{\pi(\mathbf{X}_i)\cdot\widehat{K_1}(Q_i)}}{\sum_{i=1}^{n}\frac{W_i\cdot\delta_i}{\pi(\mathbf{X}_i)\cdot\widehat{K_1}(Q_i)}} - \frac{\sum_{i=1}^{n}\frac{(1-W_i)\cdot\delta_i\cdot Y_i}{(1-\pi(\mathbf{X}_i))\cdot\widehat{K_0}(Q_i)}}{\sum_{i=1}^{n}\frac{(1-W_i)\cdot\delta_i}{(1-\pi(\mathbf{X}_i))\cdot\widehat{K_0}(Q_i)}}, \tag{7}$$

where $\widehat{K_p}(Q)$ is the Kaplan–Meier estimation (Kaplan and Meier, 1958) of the censoring distribution.

2.2 Recursive partitioning for heterogeneous treatment effects

The goal of SCT is not only estimating the average treatment effect with censored data. More importantly, it aims to find the patient subgroups with heterogeneous treatment effects. Therefore, instead of estimating the average treatment effect on the whole population level, we want to find subgroups with heterogeneous conditional treatment effect (Athey and Imbens, 2016):

$$\tau_c(\mathbf{X}) = \mathbb{E}[Y(1) - Y(0)|\mathbf{X}]. \tag{8}$$

Recursive partitioning is an ideal way for finding such subgroups. Starting from the root node containing the entire population, a tree model is constructed by recursively splitting the node into two disjoint child nodes until a stopping criterion is met. By the end of this construction, each sub-populations is naturally presented by a terminal node of the tree.

We follow the most popular recursive partitioning approach, CART (Breiman *et al.*, 1984) to construct the survival causal tree. The tree construction consists of three major components: (i) growing a large initial tree; (ii) a pruning strategy; (iii) a cross validation method to determine the best tree size.

To grow the initial tree, we want to find the splitting variable and the threshold that maximizes the sum of squared average treatment effects of the two children nodes (Athey and Imbens, 2016):

$$Q_{split}(\widehat{\tau}_c) = \left(\widehat{\tau}_c^L\right)^2 + \left(\widehat{\tau}_c^R\right)^2, \tag{9}$$

where $\widehat{\tau}_c^L$ is the conditional treatment effect of the left child node estimated with Equation 7 using the samples within the node, and $\widehat{\tau}_c^R$ is the conditional treatment effect of the right child node.

The splitting process is repeated in each child node until one of the stopping criteria is met, usually the maximum depth of the tree. The procedure results in a large initial tree.

To prune the tree, we adopt the standard cost complexity pruning strategy. Specifically, for a pre-specified complexity parameter α, we penalize the splitting criterion proportional to the complexity of the tree model

$$Q_{prune}(\widehat{\tau}_c) = Q_{split}(\widehat{\tau}_c) - \alpha \cdot K, \tag{10}$$

where K is the number of leaves in the tree. The best α value is selected using cross validation as in the original CART algorithm (Breiman *et al.*, 1984).

In practice the size of the tree can also be moderated by setting the minimum number of samples in terminal nodes or the minimal number of samples in a node to consider a split.

We summarize the SCT algorithm in the following procedure.

Procedure. Survival Causal Tree
Input: n training examples $(Y_i, \delta_i, W_i, \mathbf{X_i})$, where (Y_i^{obs}, δ_i) are the censored survival time, $\mathbf{X}_i$ are the covariates, W_i is the treatment.

1. Construct a causal tree using the splitting criterion in Equation 9 with fixed α (usually $\alpha = 0$).
2. Find the optimal α with cross validation.
3. Prune the tree with α.

Output: The pruned tree model, where the subgroups are defined by the leaf nodes of the tree.

3 Results

In this section, we compare SCT with two existing methods on TCGA cancer datasets to study their effectiveness for discovering heterogeneous treatment effects.

3.1 Breast cancer

This dataset contains breast invasive carcinoma (BRCA) samples obtained from TCGA, which includes both expression profiles and the corresponding clinical information.

The data is preprocessed by removing genes with mean expression levels in the lower quartile. The processed dataset contains expression levels of 11 535 genes across 964 patients.

The radiotherapy (RT) status of each patient is used as the treatment indicator, and the relapse free survival (RFS) time is considered as the outcome of interest.

The dataset is divided into a *training set* containing half the samples and a *test set* with the remaining samples. The number of treated and untreated samples are forced to be similar in both sets.

At whole population level, the impact of RT on RFS is not significant. The treated and untreated RFS curves in compared in Figure 2 (left). The result is agreed with the findings from clinical research (Bellon, 2015), that no study has shown a significant survival benefit of RT at the entire population level.

SCT identifies four subgroups of patients from the training data. The corresponding tree model is illustrated in Figure 3 (left), and the RFS curves of treated and untreated patients in each subgroup are shown in Figure 4. The first group of patients, defined by low expressions of AGR2 and MFAP3L, is found to have a non-significant response towards RT; the second group of patients, defined by low

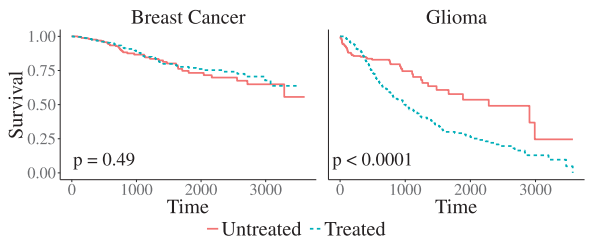

Fig. 2. The Kaplan–Meier curve of the relapse free survival for breast cancer (left) and glioma (right) patients with and without radiotherapy treatment. The *P*-value is obtained by log-rank test (Schoenfeld, 1981). The unit of time is day

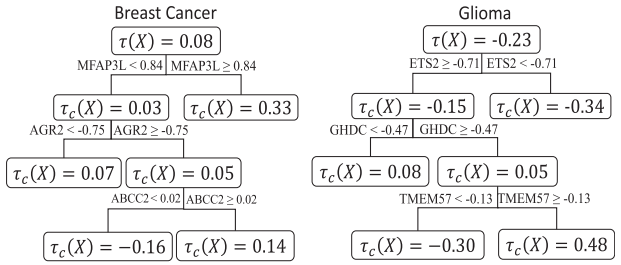

Fig. 3. The survival causal tree (SCT) constructed from the training data BRCA (left) and Glioma (right) datasets, respectively. $\tau(X)$ and $\tau_c(X)$ are the average treatment effect and conditional average treatment effect at a node, respectively

expressions of the MFAP3L and ABCC2 genes but high expression of AGR2, is found to receive negative effect from RT; the third group which is defined by low expression of MFAP3L and high expressions of both AGR2 and ABCC2, and the fourth group which is defined by high expression of MFAP3L, are both found to benefit significantly from RT.

The subgroups identified from the training set generalize well to the patients from the test set. From Figure 4 (second row), each subgroup in the test set show similar RFS curves as those in the training set. These results demonstrate that SCT can be used to predict treatment responses of unseen patients.

All three genes selected by SCT have been biologically proven to be closely related to cancer development. ABCC2 is shown to be closely related to the relapse free survival of breast cancer patients (Maciejczyk *et al.*, 2011); AGR2 has been considered as a potential drug target and biomarker for breast cancer patients (Salmans *et al.*, 2013); and MFAP3L has been studied in colorectal cancer and is shown to be able to promote cell invasion and metastasis (Lou *et al.*, 2014). We have also validated these genes on an independent collection of 3951 breast cancer patients (Gyrffy *et al.*, 2009), the results show that the expressions of these genes are significantly related to the RFS time of the patients ($P < 0.00001$) (the details is included in the Supplementary Material).

3.2 Glioma

The glioma dataset is also obtained from TCGA. The data is processed with the same procedure as the BRCA dataset. The processed dataset contains 632 samples and 11 543 genes.

At the whole population level, the effectiveness of RT on RFS is complicated. The treated and untreated RFS curves are illustrated in Figure 2 (right). It is clear that during initial weeks, RT improves the survival significantly (Valduvieco *et al.*, 2012). However, later on the survival probability of treated patients drops dramatically and becomes significantly lower than the untreated patients. One possible explanation is that radiotherapy is known to have different effects on glioma patients based on the grade and location of the tumor (Chao and Suh, 2006).

SCT has identified four subgroups in this dataset, and the constructed tree is shown in Figure 3 (right). The RFS curves of treated and untreated patients in each subgroup are shown in Figure 5. For the first subgroup (high expression of ETS 2 but low expression of GHDC), no significant difference in survival time between treated and untreated patients has been found. However, for the second (low expression of ETS2) and the third group (high expression of ETS2 but low expression of TMEM57), the untreated survival probability is significantly higher than the treated survival probability. For the fourth group of patients, defined by high expression of ETS2

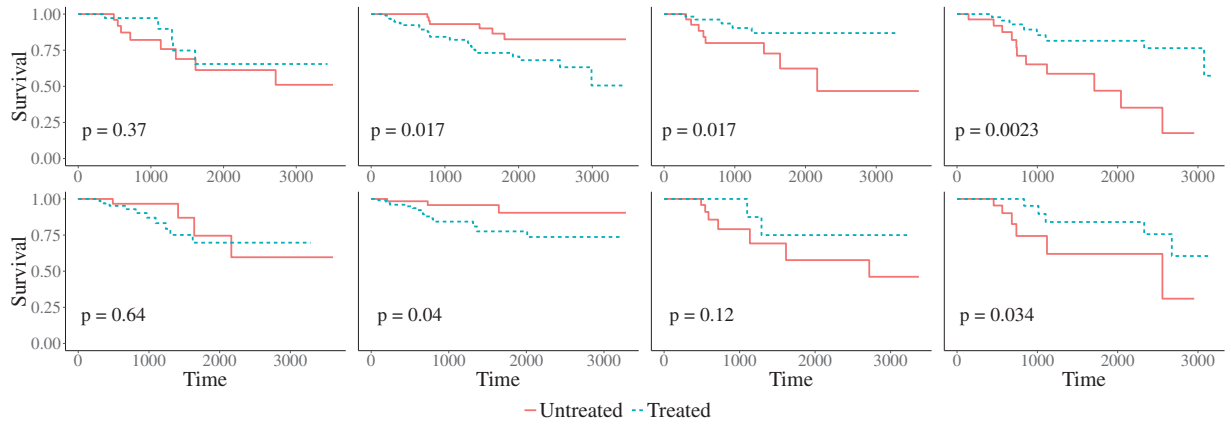

Fig. 4. The RFS curves of treated and untreated BRCA patients of each subgroups identified by SCT. First row shows the results on the training data, second row shows the result on the test data. The unit of time is day

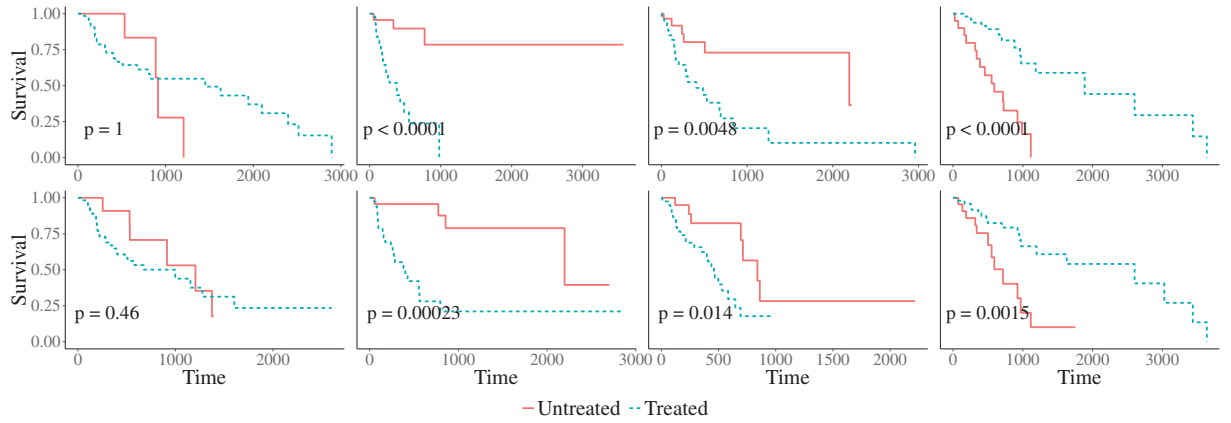

Fig. 5. The RFS curves of treated and untreated Glioma patients of each subgroup identified by SCT. First row shows the results on the training data, second row shows the result the test data. The unit of time is day

and high expression of TMEM57, the RFS curve of treated patient is significantly better than that of the untreated patient.

Both EST2 and TMEM57 are known to be related to cancer metastasis from biologic experiments. ETS2 is related to multiple cancers, including breast cancer, lung cancer, and prostate cancer (Carbone, 2003). TMEM57 encodes transmembrane proteins, and the dysregulation of transmembrane proteins is related to multiple cancers (Kampen, 2011; Zhang *et al.*, 2016). However, as mentioned earlier, the effectiveness of RT depends on many factors. The type of glioma, the grade and the location of the tumor should all be considered when deciding whether radiotherapy should be used as a treatment. Currently the limited amount of samples from public available datasets does not support an analysis considering all the factors. However, the results demonstrate that SCT can serve as a promising way for discovering the genes responsible for the heterogeneous responses to cancer treatment.

3.3 Comparison to existing methods

In this section we investigate whether existing methods can be used to find patients subgroups with heterogeneous treatment effects. Two representative methods are examined for this purpose: semi-supervised clustering (Bair and Tibshirani, 2004) and L1-regularized COX proportional hazard model (Goeman, 2009).

Clustering is one of the most widely used methods for identifying cancer subtypes (Bair and Tibshirani, 2004; Koestler *et al.*, 2010; Liu *et al.*, 2014; Monti, 2003; Shen *et al.*, 2009; Wilkerson and Hayes, 2010). Instead of utilizing all genes, semi-supervised clustering (SS-Clust) selects genes that are most relevant to the survival outcome, then uses k-means clustering to identify the disease subgroups. In order to utilize treatment information, genes related to the RFS of treated and untreated patients are selected separately. Then the union of two sets of selected genes is used for the clustering procedure. The number of cluster k is determined by the silhouette method (Rousseeuw, 1987). As shown in the Supplementary Material, different k-values do not change the results.

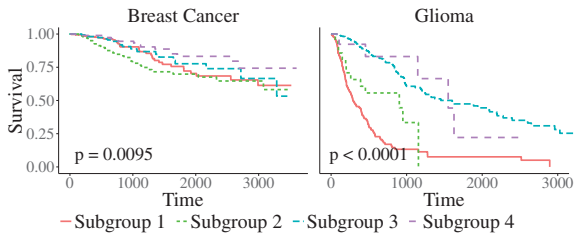

Fig. 6. The Kaplan–Meier curve of the RFS for different subtypes found by SS-Clust on both datasets. The unit of time is day

The subgroups found by SS-Clust show different RFS curves (Fig. 6). However, for each subgroup identified by SS-Clust, the RFS curves between treated and untreated patients are not significantly

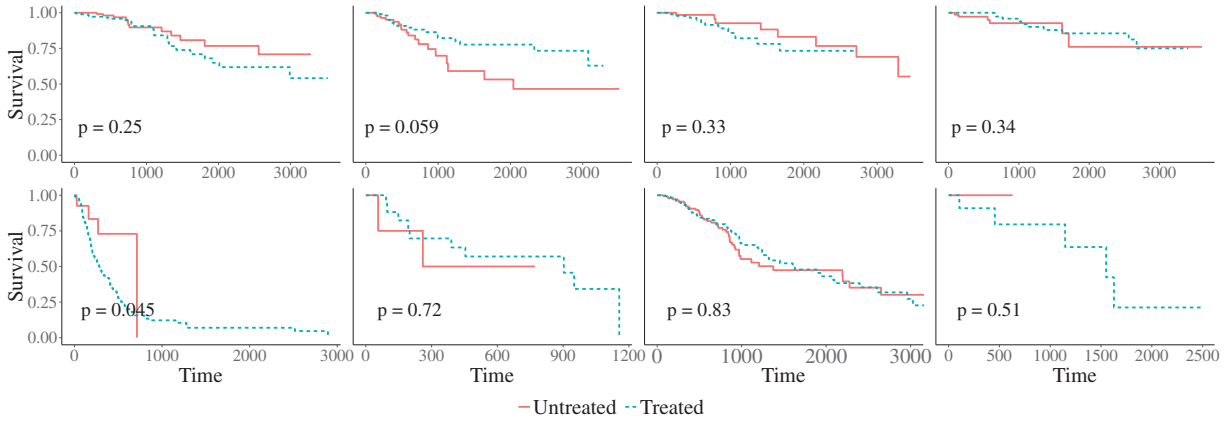

Fig. 7. Patients subgroups discovered by SS-Clust do not differentiate patients response towards RT. First row: results for BRCA dataset. Second row: results for Glioma dataset. The unit of time is day

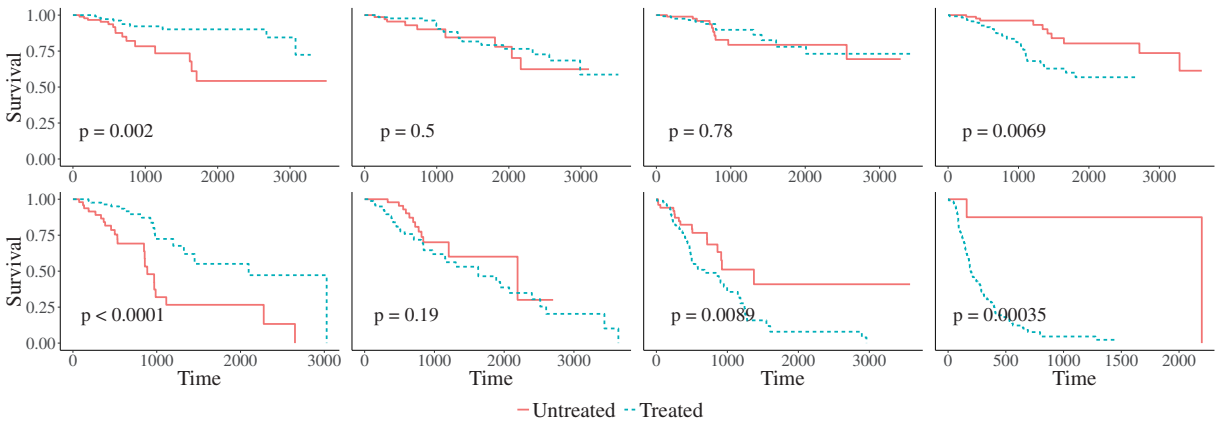

Fig. 8. The RFS curves of treated and untreated patients of each subgroup identified by L1-Cox. First row: BRCA. Second row: Glioma. The unit of time is day

separated for either dataset (Fig. 7). These results indicate that although SS-Clust is effective for finding subgroups with different survival patterns, it is not effective for discovering subgroups with heterogeneous treatment effects.

Proportional hazards (PH) model (Cox, 1972) is one of the most widely used survival analysis methods. L1-regularized Cox (Goeman, 2009) improves the high dimensional performance of PH model by utilizing L1 regularization. In this comparison we use L1-regularized Cox model (L1-Cox) with the following settings. The regressors $\mathbf{X}_{reg}$ consist of treatment variable W, gene expression levels $\mathbf{X}$ and the interaction term between the treatment and the expression levels $W \cdot \mathbf{X}$, i.e. $\mathbf{X}_{reg} = (W, \mathbf{X}, W \cdot \mathbf{X})$. The shrinkage parameter is selected by 5-fold cross validation. Once the regression coefficients β are estimated, the patients are divided into four subgroups according to the quartiles of value $I = \beta \mathbf{X}_{reg|W=1} - \beta \mathbf{X}_{reg|W=0}$, where I is the difference between treated prognostic index (PI) and the untreated PI (Bovelstad *et al.*, 2007).

For L1-Cox, the first and the last subgroups show different treatment effects for both datasets (Fig. 8). Specifically, patients in the first quartile show positive treatment effect, and those in the last quartile have negative effect. Although L1-Cox can be used to identify subgroups with different treatment effects, it is different from SCT. First, L1-Cox is necessarily a linear model whereas SCT is a tree-based model thus more general. The subgroups found by L1-Cox is reciprocal in a sense that the differences will almost always occur between the first and the last quartiles; however, such limitation does not apply to SCT. In addition, L1-Cox uses much more genes to define the subgroups (60 genes) than SCT (3 genes), which makes SCT more friendly for potential clinical implementation.

4 Conclusions

Identifying patient subgroups with heterogeneous treatment effects is of great importance for personalized cancer treatment. Recent research has shown that due to the genetic differences in people and their tumors, widely used cancer treatments are not suitable for every patient.

Computational methods are needed to find the genes responsible for heterogeneous treatment effects. However there are no commonly accepted criteria for deciding whether a treatment is applicable for each individual patient, largely because of the large number of genes in human genome.

Existing methods for finding disease subtypes are not suitable for the task. As shown in the experiments, even with treatment information considered the subtypes identified by existing methods do not differentiate heterogeneous treatment effects.

In this article we propose the SCT method, a causal approach for discovering patient subgroups with heterogeneous treatment effect from censored survival data. To the best of our knowledge, this is the first method designed for such a task. Results on two TCGA

datasets demonstrate that SCT is effective for identifying patient subgroups with different responses to RT.

The method can be used for personalized treatment. As demonstrated in the experiments, the models derived from training data generalize well to the test sets on both datasets. This would enable medical institutes to build models on existing patient data, and use the model to help medical practitioners in selecting the most suitable treatment strategy for each individual patient.

There are multiple research directions for future exploration. First, a more comprehensive study considering more clinical factors should be conducted when more samples are available. Second, estimating the survival distribution is time consuming and a more efficient approach may significantly reduce the running time of the algorithm (see Supplementary Material for a brief comparison). Alternative principles for choosing the splitting gene are also worth considering, such as maximizing the homogeneity within each child node. In addition, ways to relax the independent assumptions should also be explored.

Acknowledgement

The authors would like to thank the anonymous reviewers for their highly constructive comments and suggestions.

Funding

This work is supported by Australian Research Council Discovery Project (DP140103617). Thuc Duy Le is supported by NHMRC Grant (ID 1123042). Zhi-Hua Zhou is supported by: National Science Foundation of China (61333014).

Conflict of Interest: none declared.

References

Anstrom,K.J. and Tsiatis,A.A. (2001) Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data. *Biometrics*, **57**, 1207–1218.

Athey,S. and Imbens,G. (2016) Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, 7353–7360.

Bair,E. and Tibshirani,R. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.*, **2**, e108.

Bellon,J.R. (2015) Personalized radiation oncology for breast cancer: the new frontier. *J. Clin. Oncol.*, **33**, 1998–2000.

Bovelstad,H. *et al.* (2007) Predicting survival from microarray data a comparative study. *Bioinformatics*, **23**, 2080–2087.

Breiman,L. *et al.* (1984) *Classification and Regression Trees*. Wadsworth Publishing Company, Belmont, CA.

Carbone,G.M. (2003) Selective inhibition of transcription of the ets2 gene in prostate cancer cells by a triplex-forming oligonucleotide. *Nucleic Acids Res.*, **31**, 833–843.

Chao,S.T. and Suh,J.H. (2006) When should radiotherapy for low-grade glioma be given—immediately after surgery or at the time of progression?. *Nat. Clin. Pract. Oncol.*, **3**, 136–137.

Cox,D.R. (1972) Regression models and life-tables. *J. R. Stat. Soc.*, **34**, 187–220.

Doove,L.L. *et al.* (2013) A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. *Adv. Data Anal. Classif.*, **8**, 403–425.

Efron,B. (1988) Logistic regression, survival analysis, and the Kaplan–Meier curve. *J. Am. Stat. Assoc.*, **83**, 414–425.

Goeman,J.J. (2009) L1 penalized estimation in the cox proportional hazards model. *Biom. J*, **52**, 70–84.

Goldhirsch,A. *et al.* (2011) Strategies for subtypes–dealing with the diversity of breast cancer: highlights of the St Gallen international expert consensus on the primary therapy of early breast cancer 2011. *Ann. Oncol.*, **22**, 1736–1747.

Gyrffy,B. *et al.* (2009) An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res. Treat.*, **123**, 725–731.

Hayden,E.C. (2009) Personalized cancer therapy gets closer. *Nature*, **458**, 131–132.

Imai,K. and Ratkovic,M. (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.*, **7**, 443–470.

Imbens,G. and Rubin,D. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge, UK.

Kampen,K.R. (2011) Membrane proteins: The key players of a cancer cell. *J. Membr. Biol.*, **242**, 69–74.

Kang,J. *et al.* (2012) Tree-structured analysis of treatment effects with large observational data. *J. Appl. Stat.*, **39**, 513–529.

Kaplan,E.L. and Meier,P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Stat. Ass.*, **53**, 457–481.

Koestler,D.C. *et al.* (2010) Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. *Bioinformatics*, **26**, 2578–2585.

Liu,Y. *et al.* (2014) A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC Bioinformatics*, **15**, 37.

Lou,X. *et al.* (2014) MFAP3l activation promotes colorectal cancer cell invasion and metastasis. *Biochim. Biophys. Acta (BBA) Mol. Basis Dis.*, **1842**, 1423–1432.

Lunceford,J.K. and Davidian,M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statist. Med.*, **23**, 2937–2960.

Maciejczyk,A. *et al.* (2011) ABCC2 (MRP2, cMOAT) localized in the nuclear envelope of breast carcinoma cells correlates with poor clinical outcome. *Pathol. Oncol. Res.*, **18**, 331–342.

Monti,S. (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*, **52**, 91–118.

Park,M.Y. and Hastie,T. (2007) L1-regularization path algorithm for generalize data linear models. *J. R. Stat. Soc. Ser. B*, **69**, 659–677.

Perou,C.M. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

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.

Rousseeuw,P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Salmans,M.L. *et al.* (2013) The estrogen-regulated anterior gradient 2 (AGR2) protein in breast cancer: a potential drug target and biomarker. *Breast Cancer Res.*, **15**, 204.

Schoenfeld,D. (1981) The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, **68**, 316–319.

Seaman,S.R. and White,I.R. (2013) Review of inverse probability weighting for dealing with missing data. *Stat. Methods Med. Res.*, **22**, 278–295.

Shen,R. *et al.* (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.

Su,X. *et al.* (2009) Subgroup analysis via recursive partitioning. *J. Mach. Learn. Res.*, **10**, 141–158.

The Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Valduvieco,I. *et al.* (2012) Impact of radiotherapy delay on survival in glioblastoma. *Clin. Transl. Oncol.*, **15**, 278–282.

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

Zhang,W. *et al.* (2016) Predicting miRNA targets by integrating gene regulatory knowledge with expression profiles. *Plos One*, **11**, e0152860.

Zhang,W.-J. and Zhou,Z.-H. (2014). Multi-instance learning with distribution change. In: *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pp. 2184–2190.