A Selective Overview of Statistical Methods for Identification of the Treatment-Sensitive Subsets of Patients

Xinyi Ge, Yingwei Peng, and Dongsheng Tu

Abstract Identification of a subset of patients who may benefit from or be sensitive to a specific type of treatment has become a very important research topic in clinical trials and other types of clinical research. Statistical methods are essential in helping clinical researchers to identify the subset. In this article, we provide a selective overview of statistical methods developed in recent years in this research areas. Specifically, we consider first the cases where the outcome of the clinical studies is time-to-event or survival time and the subset is defined by one continuous covariate, such as the expression level of a gene, or by multiple covariates which can be continuous or categorical, such as mutation statuses of multiple genes. The cases where the outcomes of the clinical studies are longitudinal or repeated measurements, such as patient reported quality of life scores before, during, and after a treatment, are considered next. Gaps between the needs in clinical research and the methods available in statistical literature are identified and future research topics to bridge these gaps are discussed based on this overview.

Keywords Censored survival times · Clinical trials · Interaction · Longitudinal data · Predictive function

1 Introduction

For many diseases, such as cancer, it is often difficult to find a treatment that benefits all patents. There is an interest to identify a subset of patients, defined by individual characteristics, such as age, gender, blood test results, or gene expression levels,

X. Ge
Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada
e-mail: xinyi.ge@queensu.ca

Y. Peng · D. Tu (✉)
Departments of Public Health Sciences and Mathematics and Statistics, Queen's University, Kingston, ON, Canada
e-mail: yingwei.peng@queensu.ca; dtu@ctg.queensu.ca

W. He et al. (eds.), *Advances and Innovations in Statistics and Data Science*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-031-08329-7_15

who may be more sensitive to a specific treatment and have a larger treatment effect in comparison with a standard treatment. Conversely, if a treatment is costly or has potential negative side effects, there is also an interest to look for subsets of patients for which the treatment has less side effects. Therefore, identification of treatment-sensitive subsets of patients for a specific treatment has become a very important topic in clinical research. For example, in a recent secondary analysis of data from CO.17 and CO.20 trials conducted by the Canadian Cancer Trials Group (CCTG), the investigators were interested to know whether older patients with advanced colorectal cancer treated by, respectively, cetuximab alone or cetuximab plus brivanib had a less benefit, in comparison with younger patients, in terms of various outcomes including overall survival and quality of life (Wells et al. 2008).

Subset analysis, which includes (1) identification of the subsets, (2) estimation of treatment effects in the subsets, and (3) tests for the significance of the differences in the treatment effects in these subsets, is a main statistical tool to assess the heterogeneity in treatment effects in subsets defined by certain characteristics of patients. For example, in the analyses of CO.17 and CO.20 data mentioned above, patients were divided into two age subsets based on whether their age was 70 years or older and differential treatment effects in these two age subsets were assessed through a test of interaction between the subset and treatment. However, it is unclear whether 70 years is an optimal cutpoint to define the age subsets when assessing the heterogeneity of treatment effects by age. This issue arises in many studies where the variable to define subsets is continuous but a pre-specified cutpoint is not available from previous studies or clinical experience, and a statistical approach is often needed to determine the optimal cutpoints based on data.

When the outcomes for the subgroup analyses are times to an event or survival times, such as progression-free or overall survivals, several approaches have been proposed for the determination of cutpoints in the definition of subsets. For example, Jiang et al. (2007) proposed a biomarker-adaptive threshold design, which combines a test for overall treatment effect in all patients with the determination and validation of a cutpoint for a biomarker which is used to define a sensitive subset. Chen et al. (2014) developed a hierarchical Bayesian procedure to estimate simultaneously the interaction parameter and cutpoint in a threshold Cox proportional hazards model. He et al. (2018) proposed a single-index threshold Cox proportional hazard model, which includes a smoothly clipped absolute deviation (SCAD) penalty function, to select and linearly combine multiple biomarkers in identification of treatment-sensitive subsets. Su et al. (2008) developed an interaction tree procedure, which recursively partitions the patients into two subsets based on the greatest interaction between the subset and treatment, to obtain treatment-sensitive subsets.

When the outcomes are longitudinal measurements, Moineddin et al. (2008) used multilevel models including patient-specific random effects to identify subsets of patients with differential treatment effects of gabapentin versus placebo on longitudinal measurements of hot flashes based on the baseline measurements in a double-blind randomized controlled trial for treatment of hot flashes in women who enter menopause naturally but a median was used as the cutpoint in defining subsets. Andrews et al. (2017) considered a random effects linear model for longitudinal

outcomes to determine whether a patient had a positive response to the treatment and supervised learning algorithms were proposed to estimate a predictive function for the positive response but 0.5 was used as an ad hoc cutpoint for the predictive function to assign patients into subsets. Recently, Ge et al. (2020) introduced a threshold linear mixed model for the identification of treatment-sensitive subsets of patients based on longitudinal outcomes.

The objectives of this article are to provide a detailed review of the methods mentioned above and, based on this review, to discuss some future directions in this interesting and important area of research.

The remainder of this article is organized as follows. Sections 2 and 3 present a detailed review of statistical methods developed when, respectively, survival times and longitudinal measurements are the outcomes of the clinical research. Discussions on the future research directions are presented in the last section.

2 Statistical Methods for Treatment-Sensitive Subset Identification with Survival Times

Time to an event, which is denoted as F in this article and usually called as the survival time with overall survival or progression-free survival as examples, is usually a primary endpoint in a cancer clinical trial. Before we give detailed descriptions on the approaches proposed to identify treatment-sensitive subsets of patients based on survival times, some conventional notations, and a commonly used statistical model for the survival times are introduced below.

Denote F_i and C_i as, respectively, the potential survival and censoring times of a patient i ($i = 1, 2, \cdots, n$). The observed survival times T_i and survival status indicator δ_i are defined, respectively, as

$$\begin{cases} T_i &= \min(F_i, C_i), \\ \delta_i &= I_{(F_i < C_i).} \end{cases} \tag{1}$$

Let $h(t|\mathbf{W_i})$ be the hazard function of survival time F_i for a patient with a vector of covariates $\mathbf{W_i}$, which may include treatment indicators $\mathbf{X_i}$ and biomarkers of interest $\mathbf{Z_i}$. In the survival analysis, Cox's proportional hazards model (Cox 1972, 1975) is usually used to model the relationship between $h(t|\mathbf{W_i})$ and $\mathbf{W_i}$ as follows:

$$h(t|\mathbf{W_i}) = h_0(t)g(\mathbf{W_i}, \boldsymbol{\beta}),$$

where $g(\cdot)$ is a given link function, $h_0(t)$ is an unknown baseline hazard function, and $\boldsymbol{\beta}$ is an unknown vector of regression coefficients. A non-informative censoring is assumed, which implies that, given the covariates W_i, F_i, and C_i are independent.

2.1 An Approach Based on a Biomarker-Adaptive Threshold Design

We first review an approach based on a biomarker-adaptive threshold design proposed by Jiang et al. (2007), which tests first for an overall treatment effect in all patients and, if the overall treatment effect is not significant, proceeds to the next step to determine a cutpoint for a biomarker to identify a potential treatment-sensitive subset of patients.

Specifically, consider the following threshold Cox's proportional hazards model:

$$\log\{h(t|\mathbf{W_i})\} = \log h_0(t) + \beta_1 X_{1i} + \beta_2 I_{(Z_{1i}>c)} + \beta_3 X_{1i} I_{(Z_{1i}>c)}, \tag{2}$$

where, for $i = 1, 2, \cdots, n$, $\mathbf{W_i} = (X_{1i}, Z_{1i})$ with X_{1i} an treatment indicator equal to 1 if patient i is assigned into a treatment group or 0 if into a control group and Z_{1i} the value of a continuous biomarker which is used to define the treatment-sensitive subset, c is an unknown threshold parameter for the definition of the sensitive subset, β_1 is the main treatment effect, β_2 is the main biomarker effect, and β_3 is the treatment by biomarker interaction effect. Without loss of generality, c and Z_{1i} are assumed to take values in the interval (0, 1).

In the first step of their procedure, the effect of treatment over all patients is assessed, which can be achieved by taking $\beta_2 = \beta_3 = 0$ in model (2) and testing the null hypothesis that $\beta_1 = 0$ in the reduced model

$$\log h(t|\mathbf{W_i}) = \log h_0(t) + \beta_1 X_{1i}$$

by a likelihood ratio test. If the test rejects the null hypothesis of no treatment effect over all patients, the procedure stops and one can conclude that the treatment will benefit all patients. Otherwise, the procedure will continue to assess whether there is a subset of patients defined by a biomarker who may benefit from the treatment by testing the null hypothesis that $\beta_3 = 0$ in the full model (2).

Since the threshold parameter c is unknown, the following procedure is proposed to test the null hypothesis that $\beta_3 = 0$ under the assumption that $\beta_1 = 0$: For each candidate biomarker threshold in the range (0, 1), a reduced model (2) with $\beta_1 = 0$ is fitted on the subset of patients with biomarker values over c to obtain a log-likelihood ratio statistic $S(c)$ for testing the null hypothesis $\beta_3 = 0$ under the given c. Maximizing $S(c)$ over a range of possible cutpoint values would give a test statistic for testing null hypothesis $\beta_3 = 0$ with c unspecified. In order to obtain a reasonable power, a test statistic T is defined as $\max((S(0)+R), \max_{0<c<1} S(c))$, where R is a positive constant which was suggested to be 2.2 by Jiang et al. (2007). The p-value of this test statistic can be calculated from a resampling-based approach by randomly permutating treatment labels. If the test rejects the null hypothesis $\beta_3 = 0$, the optimal threshold c_0 can be estimated as

$$\hat{c}_0 = \arg\max_{c_0} l(c_0),$$

where $l(c_0)$ is the partial log-likelihood function based on model (2):

$$l(c_0) = \max_{\beta_1,\beta_2,\beta_3} l(\beta_1, \beta_2, \beta_3, c_0).$$

Therefore, the treatment-sensitive subset of patients can be defined by $\{i : I(Z_{1i} > \hat{c}_0)\}$, that is, a patient will be sensitive to the treatment if the observed value of the biomarker from this patients is over $\hat{c}_0$.

2.2 A Hierarchical Bayesian Method

Chen et al. (2014) proposed a hierarchical Bayesian method to estimate all unknown parameters, including the threshold c, in model (2) simultaneously without assumption $\beta_1 = 0$.

For simplicity of presentation, denote $[X_{1i}, I(Z_{1i} > c), X_{1i}I(Z_{1i} > c)]'$ as $\mathbf{W_i}(c)$ and $[\beta_1, \beta_2, \beta_3]'$ as $\boldsymbol{\beta}$. With these notations, model (2) can be rewritten as

$$h(t|\mathbf{W_i}(c)) = h_0(t)\exp\{\mathbf{W_i'}(c)\boldsymbol{\beta}\}. \tag{3}$$

Chen et al. (2014) assumed that the threshold parameter c has a prior Beta distribution Beta(2,q) for a given hyper-parameter $q > 1$, which can be written as

$$p_1(c|q) \propto q(q+1)c(1-c)^{q-1}.$$

This prior is flexible enough to accommodate any prior distribution in a family with its mode taking any specific value in the interval (0, 1). In order to assign a specific prior distribution of c, instead of taking an arbitrary value for q, it is considered that q has a hyper-prior distribution with the following density function form

$$p_2(q) \propto \frac{q-1}{q(q+1)}, \quad q > 1.$$

At the same time, $\boldsymbol{\beta}$ is assumed to has a uniform improper prior distribution $p(\boldsymbol{\beta}) \propto 1$. For every given $0 < c < 1$, the corresponding partial likelihood function of $\boldsymbol{\beta}$ in model (3) is given by

$$p_3(\boldsymbol{\beta}|c) = \prod_{i=1}^{n}\left[\frac{\exp\{\mathbf{W_i'}(c)\boldsymbol{\beta}\}}{\sum_{j\in R(T_i)}\exp\{\mathbf{W_j'}(c)\boldsymbol{\beta}\}}\right]^{\delta_i},$$

where the risk set $R(t)$ is the index set of patients who are at risk of experiencing an event at time t. Consequently, given the observed data, the joint posterior distribution of $\boldsymbol{\beta}, c, q$ can be written as

$$p(\boldsymbol{\beta}, c, q|data) \propto p_1(c|q)p_2(q)p_3(\boldsymbol{\beta}|c)$$
$$= \prod_{i=1}^{n} \left[\frac{\exp\{\mathbf{W}_\mathbf{i}'(c)\boldsymbol{\beta}\}}{\sum_{j \in R(T_i)} \exp\{\mathbf{W}_\mathbf{j}'(c)\boldsymbol{\beta}\}} \right]^{\delta_i} c(1-c)^{q-1}(q-1).$$

Therefore, the marginal posterior distributions of $\boldsymbol{\beta}$ and c can be calculated, respectively, as

$$p(\boldsymbol{\beta}) = \int_{c,q} p(\boldsymbol{\beta}, c, q|data)dcdq$$

$$p(c) = \int_{\boldsymbol{\beta},q} p(\boldsymbol{\beta}, c, q|data)d\boldsymbol{\beta}dq.$$

Statistical inferences, such as point estimation, confidence interval and hypothesis testing, on the threshold parameter c and the regression coefficient $\boldsymbol{\beta}$ can be obtained based on these marginal distributions. After obtaining the estimation of the threshold c, the treatment-sensitive subset of patients consequently can be defined if β_3 is significantly different from 0.

2.3 A Procedure Based on a Single-index Threshold Cox Model

In some clinical trials, it may be difficult to identify a treatment-sensitive subset of patients based on a single biomarker, but a combination of multiple biomarkers may have a potential to identify a treatment-sensitive subset. For example, in a randomized control trial PA.3 conducted by NCIC Clinical Trials Group, 35 key proteins were selected from a global genetic analysis of pancreatic cancers with the purpose of identifying a subset of patients with locally advanced or metastatic pancreatic cancer who will be sensitive to the treatment of erlotinib in addition to gemcitabine (Shultz et al. 2016). However, no significant interaction was found between the treatment and any of these biomarkers, which implies that it is impossible to identify a treatment-sensitive subset according to a single biomarker. He et al. (2018) found that a combination of some of these biomarkers (CA 19-9 and Axl) had the potential to define a treatment-sensitive subset of patients with pancreatic cancer. It is more complicated to identify a treatment-sensitive subset

based on multiple biomarkers, compared to the cases where there is only a single biomarker.

Several approaches have been proposed in subgroup analysis based on multiple biomarkers. He et al. (2018) proposed a single-index threshold Cox's proportional hazards model to identify treatment-sensitive subsets for each treatment using multiple biomarkers based on a linear combination of the multiple biomarkers. Let $\mathbf{X_i} = (x_{i1}, x_{i2}, \cdots, x_{id})'$ be a d-dimensional vector of exposure variables, such as treatment group indicators, for a patient i and $\mathbf{Z_i} = (z_{i1}, z_{i2}, \cdots, z_{ip})'$ be a p-dimensional vector which are the observed values of p biomarkers from the i-th patient ($i = 1, 2, \cdots, n$). Define an indicator function $I_{(\mathbf{Z_i'}\boldsymbol{\gamma_j} > c_j)}$ to be used to define the treatment-sensitive subset of patients for the j-th treatment, where $\boldsymbol{\gamma_j}$ is a p-dimensional vector used to combine biomarkers linearly and c_j is the threshold parameter. Denote $\mathbf{W_i} = (\mathbf{X_i'}, \mathbf{Z_i'})$. The proposed model can be written as

$$h(t|\mathbf{W_i}) = h_0(t) \exp\left\{\boldsymbol{\beta' X_i} + \sum_{j=1}^{d} \eta_j I_{(\mathbf{Z_i'}\boldsymbol{\gamma_j} > c_j)} + \sum_{j=1}^{d} \alpha_j x_j I_{(\mathbf{Z_i'}\boldsymbol{\gamma_j} > c_j)}\right\}, \tag{4}$$

where $h(t)$, $h_0(t)$, and $\boldsymbol{\beta}$ are the same defined in last section. The parameters $\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_d)'$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_d)'$ model the main effect of biomarker and the treatment-biomarker interaction, respectively. A significant treatment-biomarker interaction implies the treatment effect varies across subsets defined by $I_{(\mathbf{Z_i'}\boldsymbol{\gamma_j} > c_j)}$ and, consequently, the treatment-sensitive subsets for each treatment can be determined.

To obtain estimators of the parameters in the model, a maximum penalized smoothed partial likelihood method has been proposed. First, assume that data are available from n independent patients, where $i = 1, 2, \cdots, n$. Denote $\boldsymbol{\Gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_d)'$, $\boldsymbol{c} = (c_1, c_2, \cdots, c_d)'$, and $\boldsymbol{\theta} = (\boldsymbol{\beta', \eta', \alpha', c', \Gamma'})'$. Then the partial likelihood of the parameters in model (4) can be written as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\frac{\exp\left\{\boldsymbol{\beta' X_i} + \sum_{j=1}^{d} \eta_j I_{(\mathbf{Z_i'}\boldsymbol{\gamma_j} > c_j)} + \sum_{j=1}^{d} \alpha_j x_{ij} I_{(\mathbf{Z_i'}\boldsymbol{\gamma_j} > c_j)}\right\}}{\sum_{k \in R(T_i)} \exp\left\{\boldsymbol{\beta' X_k} + \sum_{j=1}^{d} \eta_j I_{(\mathbf{Z_k'}\boldsymbol{\gamma_j} > c_j)} + \sum_{j=1}^{d} \alpha_j x_{kj} I_{(\mathbf{Z_k'}\boldsymbol{\gamma_j} > c_j)}\right\}} \right]^{\delta_i}. \tag{5}$$

Since the partial likelihood function is not continuous at some parameters, the estimator of $\boldsymbol{\theta}$ cannot be obtained by maximizing the partial likelihood function (5). He et al. (2018) proposed a local distribution function $\Phi((\mathbf{Z_i'}\boldsymbol{\gamma}_j - c_j)/h)$ as a smooth approximation to the indicator function $I(\mathbf{Z_i'}\boldsymbol{\gamma_j} > c_j)$, where Φ is the distribution function of the standard normal variable and the bandwidth h converges

to zero as the sample size increases. With this approximation, the smoothed partial likelihood (SPL) function can be defined as

$$S(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\frac{\exp\{\boldsymbol{\beta}'\boldsymbol{X_i} + \sum_{j=1}^{d} \eta_j \Phi((\boldsymbol{Z_i'\gamma_j} - c_j)/h) + \sum_{j=1}^{d} \alpha_j x_{ij} \Phi((\boldsymbol{Z_i'\gamma_j} - c_j)/h)\}}{\sum_{k \in R(T_i)} \exp\{\boldsymbol{\beta}'\boldsymbol{X_k} + \sum_{j=1}^{d} \eta_j \Phi((\boldsymbol{Z_k'\gamma_j} - c_j)/h) + \sum_{j=1}^{d} \alpha_j x_{kj} \Phi((\boldsymbol{Z_k'\gamma_j} - c_j)/h)\}} \right]^{\delta_i}. \tag{6}$$

Because a large number of covariates may be available but only a few of them may be relevant in the definition of treatment-sensitive subsets, He et al. (2018) added a penalty function to the SPL function for efficiently selecting relevant biomarkers from large amount of biomarkers in practice. In their procedure, the smoothly clipped absolute deviation (SCAD) penalty function was used and the penalized smoothed partial likelihood (PSPL) function was defined as

$$L_n(\boldsymbol{\theta}) = \log\{S(\boldsymbol{\theta})\} - n \sum_{j=1}^{d} \sum_{k=1}^{p} P_\lambda(|\lambda_{jk}|), \tag{7}$$

where λ_{jk} is the component k of $\boldsymbol{\gamma_j}$ and $P_\lambda(\cdot)$ is the SCAD penalty function with a regularization parameter λ. By maximizing PSPL function (7), the estimations of $\boldsymbol{\theta}$ can be obtained. Therefore, when at least one of the α_j is significantly different from 0, corresponding treatment-sensitive subset of patients for the treatment j can be determined by the estimate $\hat{c}_j$ of c_j as $\{i : I_{(\boldsymbol{Z_i'\gamma_j} > \hat{c}_j)}\}$.

2.4 An Interaction Tree Approach

Su et al. (2008) proposed a procedure to construct an interaction tree $\mathscr{T}$ based on survival outcomes which can be used to identify treatment-sensitive subsets of patients. There are three steps in the construction of an interaction tree which are introduced in details below.

The first step is to grow a large initial tree. Let s be a single binary split of patients in the tree construction based on a biomarker z measured on patients. If z is continuous, then the split s is induced by whether or not $z \leq c$, where the threshold c can be any constant. However, in practice the threshold c is chosen as one of the observed values of z. If z is ordinal, the split s can be induced by the similar procedure. If z is a categorical variable with categories $C = \{c_1, \cdots, c_r\}$, then the split can be induced by the form of $z \in A$ with $A \subset C$. In order to reduce the computational burden, the treatment effect within each category is often estimated first and then the categories of z are reordered according to the treatment effect.

Splitting on z can then be induced by treating z as an ordinal variable. Next we need to select the best split from all possible splits, which has the greatest difference in the treatment effect between its two child nodes. The splitting selection approach in Su et al. (2008) is to choose the split to maximize a statistic for test $H_0 : \beta_3 = 0$ in the following Cox model:

$$h(t|\mathbf{W_i}) = h_0(t) \exp\{\beta_1 X_i + \beta_2 I^{(s)} + \beta_3 X_i I^{(s)}\}, \tag{8}$$

where X_i is a treatment indicator, $I^{(s)} = I_{(z \in A)}$ or $I^{(s)} = I_{(z \leq c)}$, and $W_i = (X_i, I^{(s)})$. In their method, they chose to use the following partial likelihood ratio test (PLRT) statistic as the test statistic for $H_0 : \beta_3 = 0$:

$$G(s) = -2(l_2 - l_1), \tag{9}$$

where l_2 is the maximized partial likelihood (Cox 1975) of model (8) and l_1 is the maximized partial likelihood of the reduced model under H_0:

$$h(t|\mathbf{W_i}) = h_0(t) \exp\{\beta_1 X_i + \beta_2 I^{(s)}\}. \tag{10}$$

The best split s^* can be determined by $G(s^*) = \max_s G(s)$. After choosing the best split, the patients can be divided into two subsets and therefore the tree grows two child nodes. The same procedure is then implemented to split both child nodes based on different variables such as the values of other biomarkers. A large initial tree $\mathcal{T}_0$ can be obtained by repeating the above process recursively.

Since the initial tree is large, it needs to be pruned until it has an appropriate size. Su et al. (2008) introduced the following penalty function for a node h of the initial tree:

$$g(h) = \frac{G(\mathcal{T}_h)}{|\mathcal{T}_h - \tilde{\mathcal{T}}_h|},$$

where $\mathcal{T}_h$ is the branch of tree with h as its root, $\tilde{\mathcal{T}}_h$ represents the set of all terminal nodes of $\mathcal{T}_h$, and $|\mathcal{T}_h - \tilde{\mathcal{T}}_h|$ denotes the number of all internal nodes of $\mathcal{T}_h$. By minimizing $g(h)$ over all the internal nodes of $\mathcal{T}_0$, the weakest link (or the most ineffective split) h^* can be determined. Denote $\mathcal{T}_1$ as the subtree after pruning off the branch $\mathcal{T}_{h^*}$ from $\mathcal{T}_0$ and apply the same pruning procedure to the subtree $\mathcal{T}_1$. After the above process is repeated recursively, a nested sequence of subtrees can be defined as $\mathcal{T}_M \prec \cdots \prec \mathcal{T}_m \prec \mathcal{T}_{m-1} \cdots \prec \mathcal{T}_1 \prec \mathcal{T}_0$, where $\mathcal{T}_M$ is a tree only having the root node and $\prec$ means "a subtree of."

After the pruning procedure is finished, the last step of the proposed procedure is to select the best size of the tree. For this purpose, following the split-complexity pruning algorithm for survival tree (LeBlanc & Crowley 1993), the following interaction-complexity measure is introduced to evaluate the overall goodness-of-

interaction of a given tree $\mathscr{T}$:

$$G_\lambda(\mathscr{T}) = G(\mathscr{T}) - \lambda \cdot |\mathscr{T} - \tilde{\mathscr{T}}|, \tag{11}$$

where $\tilde{\mathscr{T}}$ denotes a set of all terminal nodes of $\mathscr{T}$ and $|\mathscr{T} - \tilde{\mathscr{T}}|$ the number of all internal nodes of $\mathscr{T}$, $G(\mathscr{T}) = \sum_{h \in \mathscr{T} - \tilde{\mathscr{T}}} G(h)$, which is the sum of $G(h)$, the splitting statistic defined in (9), over node h (including its split to its child nodes), and $\lambda (\geqslant 0)$ is a penalty parameter for each added node. With this measure, an optimally sized tree $\mathscr{T}^*$ can be determined by maximizing $G_\lambda(\mathscr{T})$ as following:

$$G_\lambda(\mathscr{T}^*) = \max_{m=0,\cdots,M} \{G(\mathscr{T}_m) - \lambda \cdot |\mathscr{T}_m - \tilde{\mathscr{T}}_m|\},$$

where the penalty parameter λ can be pre-specified within the range $2 \leqslant \lambda \leqslant 4$ (LeBlanc & Crowley 1993). After the optimally sized tree is determined, the treatment-sensitive subsets of patients can be defined based on the terminal nodes of the tree $\mathscr{T}^*$.

3 Statistical Methods for Treatment-Sensitive Subset Identification Based on Longitudinal Measurements

Longitudinal measurements, which are repeated observations measured on the same patients at different points in time, are often collected in clinical trials or other medical studies. For example, although the treatment effect in cancer clinical trials are traditionally evaluated by relatively objective endpoints such as tumor response, relapse-free survival, or overall survival, it is argued that these endpoints may not provide adequate information in understanding of the treatment effect. Recently, evaluations of more subjective endpoints, such as patient reported quality of life (QoL), have become increasingly recognized in cancer clinical trials, since these endpoints can help patients to make the treatment decisions by providing detailed information on side effects of the treatment (Blazeby et al. 2001). Also these endpoints can help future patients understand the consequences of their illness and treatment (Bezjak et al. 2006). These patient reported outcomes are usually assessed at several timepoints before, during, and after patients have received the treatment.

Multilevel or hierarchical models are often used for the analysis of longitudinal data, as these models incorporate the variation at different levels of the hierarchy into analysis. This class of models includes multilevel models, linear mixed models, random effects ANOVA models, generalized estimating equations (GEE), etc. In this section, some statistical methods proposed for identifying treatment-sensitive subsets of patients based on these models when the outcomes of clinical trials are longitudinal or repeated measures are reviewed.

3.1 A Procedure Based on Multilevel Models

To establish notations, let y_{ij} be the longitudinal measurement at j-th observation time t_{ij} ($j = 1, 2, \cdots, n_i$) from patient i ($i = 1, 2, \cdots, N$). The observation times are usually called as level-1 units in a multilevel model, while patients are called as the level-2 units. Also denote X_i as the treatment indicator with $X_i = 1$ if the patient is assigned into the treatment group and $X_i = 0$ if the patient is assigned into the control group. Consider the following two-level linear regression model proposed in Moineddin et al. (2008) for these longitudinal measurements: the first level of the model assumes that the measurement y_{ij} is a linear function of observation time t_{ij}, which can be written as

$$y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}, \tag{12}$$

where e_{ij} is the random error term assumed to follow a normal distribution with mean zero and a constant variance σ_e^2 and β_{0i} and β_{1i} are, respectively, a random intercept and slope associated with the ith patient. It is assumed further that β_{0i} and β_{1i} can be explained by a linear function of X_i in the following second level of the model:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} X_i + u_{0i},$$
$$\beta_{1i} = \gamma_{10} + \gamma_{11} X_i + u_{1i},$$

where γ_{rs} ($r = 0, 1$ and $s = 0, 1$) are population average fixed effect parameters and u_{0i} and u_{1i} are random errors which follow a bivariate normal distribution with mean zero and variance-covariance $var(u_{0i}) = \sigma_0^2$, $var(u_{1i}) = \sigma_1^2$ and $cov(u_{0i}, u_{1i}) = \sigma_{01}^2$. From the definition of X_i as a treatment indicator, it can be seen that the fixed effects γ_{00} and γ_{10} are, respectively, the population average of the measurement y_{ij} at baseline (intercept) and the population average of change over time (slope) for patients in the control group, while the parameters γ_{01} and γ_{11} can be interpreted as the differences in, respectively, the population averages of the measurement y_{ij} at baseline (intercepts) and the population average of changes over time (slopes) between the treatment and the control groups. Parameter σ_0^2 is the residual variance of the measurement y_{ij} at baseline (intercept) , σ_1^2 is the residual variance of the change rate (slope), and σ_{01}^2 is the residual covariance between the baseline the measurement and rate of change.

It is known that u_{1i} represents the residuals of the regression slopes across the patients. When the variance of u_{1i} is significant at a two-sided 0.05 level, Moineddin et al. (2008) suggested that treatment-sensitive subsets of patients can be identified based on a baseline factor (age, gender, biomarker, etc.) of patients by correlating u_{1i} with this factor using a t-test or analysis of variance if the factor is categorical and the Pearson or Spearman correlations if the factor is continuous. When the association is significant at two-sided 0.05 level, treatment-sensitive subsets of

patients can be defined by the natural grouping generated by the categories of the baseline factor when it is categorical (for example, female and male subsets if the gender is the baseline factor). When the factor is continuous such as the age or value of a biomarker, however, a cutpoint is required. Only an ad hoc approach using the median of the factor as a cutpoint was suggested and there was no formal procedure proposed to estimate the cutpoint.

3.2 A Prediction Model Approach

Andrews et al. (2017) proposed a complete procedure which can be used for both identification of the treatment-sensitive subsets of patients and validation of the subsets identified based on longitudinal measurements. First step in the proposed procedure is to use a linear mixed model which includes a random effect term to evaluate the individual treatment effect and a fixed effect term to evaluate the population average treatment effect. Based on the estimates of individual treatment effect, various classifying methods can then be used to build prediction models which can be used to identify treatment-sensitive subsets of patients based on the characteristics of patients. A validation step is then followed to select the best prediction model under a marginal regression framework.

Specifically, consider the following random intercept-slope linear mixed model:

$$y_{ij} = \beta_0 + \alpha_{0i} + (\beta_1 + \alpha_{1i})X_i t_{ij} + \beta_2 t_{ij} + e_{ij}, \tag{13}$$

where X_i, t_{ij}, y_{ij} and random error term e_{ij} are the same as defined in the last subsection, β_0 and β_1 represent, respectively, the population average of the initial status and the treatment effect over time, α_{0i} and α_{1i} are, respectively, the random intercept and slope for patient i, and β_2 is the fixed effect of time. The interaction effect $\beta_1 + \alpha_{1i}$ between the treatment and time in this model describes the trend of individual treatment effect over time.

To simplify the presentation of the procedure, model (13) can be rewritten in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{D\alpha} + \boldsymbol{e}, \tag{14}$$

where $\boldsymbol{Y}$ is a n-dimensional vector of the responses with $n = \sum_{i=1}^{N} n_i$, $\boldsymbol{X}$ and $\boldsymbol{D}$ are an $n \times 3$ and $n \times 2N$ matrices of covariates corresponding to the fixed effects $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ and random effects $\boldsymbol{\alpha} = (\alpha_{01}, \cdots, \alpha_{0N}, \alpha_{11}, \cdots, \alpha_{1N})$, respectively, and $\boldsymbol{e}$ is a m-dimensional vector of the random errors. It is assumed that $E(\boldsymbol{\alpha}) = \boldsymbol{0}$ and $E(\boldsymbol{e}) = \boldsymbol{0}$. In addition, it is assumed that $\boldsymbol{\alpha}$ and $\boldsymbol{e}$ are independent and distributed as multivariate normal as

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}\right).$$

By using the conventional maximum likelihood method for the linear mixed model, the parameter estimates for the fixed and random effects can be obtained as following:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y},$$
$$\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{G}}\boldsymbol{D}'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

where $\boldsymbol{\Sigma} = \boldsymbol{DGD}' + \boldsymbol{R}$ and $\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{R}}$ are obtained by maximizing the following likelihood function:

$$l(\boldsymbol{R}, \boldsymbol{G}|\boldsymbol{Y}, \boldsymbol{X}) = -\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y})'\boldsymbol{\Sigma}^{-1}$$
$$(\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\log(2\pi),$$

where $|\boldsymbol{\Sigma}|$ is the determinant of the variance-covariance matrix $\boldsymbol{\Sigma}$. The asymptotic consistency and efficiency of these estimates were proved by Hartley and Rao (1967). Furthermore, if the variance estimation is biased, the restricted maximum likelihood would be a viable alternative method (Verbeke & Molenberghs 2009).

Since the random slope $\beta_1 + \alpha_{1i}$ describes the treatment effect over time, patients can be divided into two subsets based on whether its estimate $\hat{\beta}_1 + \hat{\alpha}_{1i}$ is positive. Define C_i as the subset indicator based on this definition. That is,

$$C_i = \begin{cases} 1 & \hat{\beta}_1 + \hat{\alpha}_{1i} > 0 \\ -1 & \hat{\beta}_1 + \hat{\alpha}_{1i} \leq 0. \end{cases}$$

Since some baseline characteristics or covariates $\mathbf{W}_i$ of patients, such as age, gender, blood pressure, and gene expression, might influence the treatment effect, a prediction model

$$f(\mathbf{W}_i) = P(C_i = 1|\mathbf{W}_i)$$

based on the subset indicator C_i and these baseline characteristics or covariates $\boldsymbol{W}_i$ may be used to classify patients into two subsets which have differential treatment effects. In general, the relationship between C_i and $\mathbf{W}_i$ is unknown, which could be linear or nonlinear, so the predictive function $f(\cdot)$ in the above prediction model needs to be estimated. Andrews et al. (2017) suggested various linear or nonlinear supervised learning algorithms, such as logistic regression, support vector machine (SVM) with linear kernel, linear discriminant analysis (LDA), decision tree, random forest, etc., may be used to estimate $f(\cdot)$. Once the estimated prediction function

$\hat{f}(\mathbf{W}_i)$ is obtained from the data, it was proposed that patient i can be classified in the subset of patients who may benefit from the treatment if $\hat{f}(\mathbf{W}_i) > 0.5$.

Andrews et al. (2017) also developed a validation procedure to assess the effectiveness of the method proposed above for the treatment-sensitive subset identification but the choice of 0.5 as the cutpoint for estimated predictive function to define the subsets is ad hoc, which may have large impact on the performance of the proposed method.

3.3 A Procedure Based on a Threshold Linear Mixed Model

Ge et al. (2020) introduced a threshold linear mixed model which can be used simultaneously to determine the cutpoint of a continuous covariate, such as age or the expression level of a biomarker, in the definition of treatment-sensitive subsets of patients and to assess the interaction effect between the treatment and subset indicator based on longitudinal measurements. The standard likelihood method is difficult to apply to the inference of the parameters in the model because the likelihood function is not continuous for some parameters. They therefore proposed a smoothing likelihood function to approximate the original likelihood function and developed an inference procedure for the parameters in the model based on this new likelihood function. Finally, they used the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970), which belongs to quasi-Newton methods and is included in R package "maxLik" (Henningsen & Toomet 2011), to implement the proposed procedure.

Specifically, denote a column vector $\mathbf{Y_i} = (y_{i1}, y_{i2}, \cdots, y_{in_i})$ for the longitudinal measurements observed from the i-th patient. For each patient, denote also $\mathbf{X_i} = (\mathbf{x_{i1}}, \mathbf{x_{i2}}, \cdots, \mathbf{x_{in_i}})'$ as an $(n_i \times p)$ designed matrix of covariates for fixed effect $\boldsymbol{\beta}$ and $\mathbf{Z_i} = (\mathbf{z_{i1}}, \mathbf{z_{i2}}, \cdots, \mathbf{z_{in_i}})'$ as an $(n_i \times q)$ designed matrix of covariates for random effect $\boldsymbol{\alpha_i}$. Assume b_i is an indicator of the treatment received by patient i with either $b_i = 1$ if the patient receiving a new therapy or $b_i = 0$ if not. Denote w_i as a continuous covariate at baseline for patient i and assume two subsets of patients can be defined based on whether w_i exceeds an unknown cutpoint c. The following threshold linear mixed model was proposed to assess the potential differential treatment effects between these two subsets:

$$\mathbf{Y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\boldsymbol{\alpha_i} + \eta_1 I(w_i > c)\mathbf{1} + \eta_2 b_i I(w_i > c)\mathbf{1} + \boldsymbol{\varepsilon_i}, \tag{15}$$

where $\boldsymbol{\varepsilon_i} = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{in_i})'$ is a vector of random errors and $\mathbf{1}$ is a n_i-dimensional vector with its all elements as 1. In model (15), the response y_{ij} of patient i measured at the time t_{ij} is modeled by three components: the fixed effects of all covariates $\boldsymbol{x'_{ij}}\boldsymbol{\beta} + \eta_1 I(w_i > c) + \eta_2 b_i I(w_i > c)$, the patient effect $\boldsymbol{z'_{ij}}\boldsymbol{\alpha_i}$, and the random error ε_{ij}. The columns of $\mathbf{X_i}$ may include intercept, time or its function, treatment, and other confounding variables, and the columns of $\mathbf{Z_i}$ are assumed to

be a subset of the columns of $\mathbf{X_i}$. In order to simplify the presentation, model (15) can be rewritten in the matrix form as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\eta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{16}$$

where $\mathbf{Y} = [\mathbf{Y_1'}, \mathbf{Y_2'}, \cdots, \mathbf{Y_N'}]'$, $\mathbf{X} = [\mathbf{X_1'}, \mathbf{X_2'}, \cdots, \mathbf{X_N'}]'$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha_1'}, \boldsymbol{\alpha_2'}, \cdots, \boldsymbol{\alpha_N'})'$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon_1'}, \boldsymbol{\varepsilon_2'}, \cdots, \boldsymbol{\varepsilon_N'})'$ and $\mathbf{W} = [\mathbf{W_1'}, \mathbf{W_2'}, \cdots, \mathbf{W_N'}]'$, and

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z_1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_2} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & & \mathbf{Z_N} \end{pmatrix} \quad , \quad \boldsymbol{W_i} = \begin{pmatrix} I(w_i > c)\ b_i \times I(w_i > c) \\ I(w_i > c)\ b_i \times I(w_i > c) \\ \vdots \qquad\qquad \vdots \\ I(w_i > c)\ b_i \times I(w_i > c) \end{pmatrix}_{n_i \times 2} .$$

For the vector of random effects $\boldsymbol{\alpha}$ and vector of random errors $\boldsymbol{\varepsilon}$ in the model, it is assumed that $E(\boldsymbol{\alpha}) = \mathbf{0}$ and $E(\boldsymbol{\varepsilon}) = \mathbf{0}$. In addition, it is assumed that $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ are independent and distributed as multivariate normal, that is,

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right).$$

In the proposed model, they assumed that $\mathbf{R} = \sigma^2\mathbf{I}$ (σ is an unknown parameter) and $\mathbf{G} = \sigma^2\rho^2\mathbf{I}$ (ρ is also an unknown parameter). Following Patterson and Thompson (1971), the covariance-variance matrix of the observation $\mathbf{Y}$ can be written as

$$Var(\mathbf{Y}) = \sigma^2(\rho^2\mathbf{Z}\mathbf{Z}' + \mathbf{I}) = \sigma^2\mathbf{H},$$

where $\mathbf{H} = \rho^2\mathbf{Z}\mathbf{Z}' + \mathbf{I}$.

Under the assumptions and notations mentioned above, $\mathbf{Y}$ follows a multivariate normal distribution as $N(\mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\eta}, \sigma^2\mathbf{H})$. Denote $n = \sum_{i=1}^{N} n_i$ as the total number of observations, The log-likelihood for the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}, c, \rho^2, \sigma^2)$ in model (16) based on longitudinal outcomes $\mathbf{Y}$ can be written as

$$l(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}) = -\frac{1}{2}\Bigg\{ \log(2\pi) + n\log\sigma^2 + \log|\mathbf{H}| + \frac{(\mathbf{Y} - \boldsymbol{X\beta} - \boldsymbol{W\eta})'\mathbf{H}^{-1}(\mathbf{Y} - \boldsymbol{X\beta} - \boldsymbol{W\eta})}{\sigma^2} \Bigg\}. \tag{17}$$

However, due to the presence of the indicator functions $I(w_i > c)$, the log-likelihood function is not continuous with respect to c, which makes the conventional maximum likelihood theory and algorithm difficult to apply. Following a

smoothing procedure used by Brown and Wang (2007), they proposed to use a kernel smooth function

$$\Phi\left(\frac{w_i - c}{h}\right) \tag{18}$$

as a smooth approximation to the indicator function $I(w_i > c)$, where Φ is the distribution function of the standard normal distribution and h is a bandwidth which converges to zero as the sample size increases. Using this approximation, a smoothed log-likelihood function can be defined by replacing $\mathbf{W_i}$ in the definition of $\mathbf{W}$ in (17) with

$$\widetilde{\mathbf{W}}_\mathbf{i} = \begin{bmatrix} \Phi(\frac{w_i-c}{h})\ b_i & \times\ \Phi(\frac{w_i-c}{h}) \\ \Phi(\frac{w_i-c}{h})\ b_i & \times\ \Phi(\frac{w_i-c}{h}) \\ \vdots & \vdots \\ \Phi(\frac{w_i-c}{h})\ b_i & \times\ \Phi(\frac{w_i-c}{h}) \end{bmatrix}_{n_i \times 2},$$

therefore the smoothed log-likelihood function of $\boldsymbol{\theta}$ is given by

$$sl(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}) = -\frac{1}{2}\Bigg\{ \log(2\pi) + n\log\sigma^2 + \log|\mathbf{H}| + \frac{(\mathbf{Y} - \boldsymbol{X\beta} - \widetilde{\boldsymbol{W}}\boldsymbol{\eta})'\mathbf{H}^{-1}(\mathbf{Y} - \boldsymbol{X\beta} - \widetilde{\boldsymbol{W}}\boldsymbol{\eta})}{\sigma^2} \Bigg\} \tag{19}$$

where $\widetilde{\mathbf{W}} = [\widetilde{\mathbf{W}}_\mathbf{1}', \widetilde{\mathbf{W}}_\mathbf{2}', \cdots, \widetilde{\mathbf{W}}_\mathbf{n}']'$. The maximum smoothed likelihood estimates (MSLE) of $\boldsymbol{\theta}$ can be obtained by maximizing the smoothed log-likelihood function (19). Based on this estimate, a treatment-sensitive subset of patients can be defined as $\{i : I(w_i > \hat{c})\}$, where $\hat{c}$ is an estimate of c, if η_2 is found significantly different from 0 based on its estimate and associated variance estimate.

4 Discussions and Future Work

Most of the methods reviewed in this article assume a specific statistical model for the clinical outcomes of the study. For example, the Cox proportional hazards models were assumed when the clinical outcomes are survival times and longitudinal outcomes are required to be normally distributed because of assumptions underlying the linear mixed models. The proportional hazards assumption behind the Cox model and the normality assumption required for linear mixed models may not be satisfied by the data. Some more robust methods with more realistic assumptions may be preferred. For example, since quality of life scores are restricted

to an interval, a linear mixed model with beta (Hunger et al. 2012) or simplex (Qiu et al. 2008) distributions may be more appropriate. For patients with early stage of cancer, some of them may be cured by the treatment they received and, therefore, cure models may be more useful for the observed survival times (Othus et al. 2012). Extensions of the methods reviewed in this article to these models may be of interest. When the cutpoint of a single biomarker is known and pre-specified and survival times are the clinical outcomes of a study, a nonparametric measure of interaction was proposed recently by Jiang et al. (2016). Development of statistical methods which use this measure of interaction to identify treatment-sensitive subsets of patients may also be of interest but can be difficult when there are multiple biomarkers.

In many clinical studies, both survival times and longitudinal measurements are collected but they are usually analyzed separately. Joint analysis of longitudinal outcomes and survival times may identify treatment-sensitive subsets of patients for both of these outcomes. But technically this may be more difficult because additional random effects are required to connect the Cox proportional hazards with linear mixed models, which will require novel computation methods to make inferences on the parameters in both of these models.

When the clinical outcomes are longitudinal, only the case where a single covariate is available to define the subsets of patients has been considered. Similar procedures as that presented in Sect. 2.3 would be generalized from the case where the clinical outcomes are survival times to the case where longitudinal outcomes are outcomes of interest to combine multiple covariates or biomarkers when they are available.

There is so far no systematic comparison between the treatment-sensitive subsets of patients identified from different approaches. As noted by Janes et al. (2015), accuracy measures such as sensitivity, specificity, and positive and negative predictions employed for the comparison of statistical procedures for the identification of prognostic groups are difficult to define for the comparisons of statistical procedures for the identification of treatment-sensitive subsets. A consensus is required among medical researchers and statisticians on the measures which could be used for the comparisons.

References

Andrews, N., & Cho, H. (2017). Validating effectiveness of subgroup identification for longitudinal data. *Statistics in Medicine, 37*, 98–106.

Bezjak, A., Tu, D., Seymour, L., Clark, G., Trajkovic, A., Zukin, M., Ayoub, J., Lago, S., de Albuquerque Ribeiro, R., Gerogianni, A., Cyjon, A., Noble, J., Laberge, F., Chan, R. T. T., Fenton, D., Pawel, J., Reck, M., & Shepherd, F. (2006). Symptom improvement in lung cancer patients treated with erlotinib: quality of life analysis of the National Cancer Institute of Canada Clinical Trials Group study BR.21. *Journal of Clinical Oncology, 24*, 3831–3837.

Blazeby, J. M., Brookes, S. T., & Alderson, D. (2001). The prognostic value of quality of life scores during treatment for oesophageal cancer. *Gut, 49*, 227–230.

Brown, B. M., & Wang, Y.-G. (2007). Induced smoothing for rank regression with censored survival times. *Statistics in Medicine, 26*, 828–836.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics, 6*, 76–90.

Chen, B. E., Jiang, W., & Tu, D. (2014). A hierarchical Bayes model for biomarker subset effects in clinical trials. *Computational Statistics & Data Analysis, 71*, 324–334.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*, 187–202.

Cox, D. R. (1975). Partial likelihood. *Biometrika, 62*, 269–276.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal, 13*, 317–322.

Ge, X., Peng, Y., & Tu, D. (2020). A threshold linear mixed model for identification of treatment-sensitive subsets in a clinical trial based on longitudinal outcomes and a continuous covariate. *Statistical Methods in Medical Research, 10*, 2919–2931.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation, 24*(109), 23–26.

Hartley, H. O., & Rao, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika, 54*, 93–108.

He, Y., Lin, H., & Tu, D. (2018). A single-index threshold *c*ox proportional hazard model for identifying a treatment-sensitive subset based on multiple biomarkers. *Statistics in Medicine, 37*, 3267–3279.

Henningsen, A., & Ott Toomet, O. (2011). MaxLik: A package for maximum likelihood estimation in R. *Computational Statistics, 26*, 443–458.

Hunger, M., Döring, A., & Holle, R. (2012). Longitudinal beta regression models for analyzing health-related quality of life scores over time. *BMC Med Res Methodol*, 144: https://doi.org/10.1186/1471-2288-12-144.

Janes, H., Pepe, M. S., McShane, L. M., Sargent, D. J., & Heagerty, H. J. (2015). The fundamental difficulty with evaluating the accuracy of biomarkers for guiding treatment. *Journal of National Cancer Institute, 107*, djv157.

Jiang, W., Freidlin, B., & Simon, R. (2007). Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute, 99*, 1036–1043.

Jiang, S., Chen, B., & Tu, D. (2016). Inference on treatment-covariate interaction based on a nonparametric measure of treatment effects and censored survival data. *Statistics in Medicine, 35*, 2715–2725.

LeBlanc, M., & Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association, 88*, 457–467.

Moineddin, R., Butt, D. A., Tomlinson, G., & Beyene, J. (2008). Identifying subpopulations for subgroup analysis in a longitudinal clinical trial. *Contemporary Clinical Trials, 29*, 817–822.

Othus, M., Barlogie, B., LeBlanc, M. L., & Crowley, J. J. (2012). Cure models as a useful statistical tool for analyzing survival. *Clin Cancer Res, 18*, 3731–3736.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika, 58*, 545–554.

Qiu, Z., Song, P., & Tan, M. (2008). Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics, 35*, 577–596.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation, 24*, 647–656.

Shultz, D. B., Pai, J., Chiu, W., Ng, K., Hellendag, M. G., Heestand, G., Chang, D. T., Tu, D., Moore, M. J., Parulekar, W. R., & Koong, A. (2016). A novel biomarker panel examining response to gemcitabine with or without erlotinib for pancreatic cancer therapy in NCIC Clinical Trials Group PA.3. *PloS One, 11*, e0147995.

Su, X., Zhou, T., Yan, X., Fan, J., & Yang, S. (2008). Interaction trees with censored survival data. *The International Journal of Biostatistics, 4*, 2.

Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Science & Business Media.

Wells, C., O'Callaghan, C., Karapetis, C. S., Jonker, D., Tu, D., Liu, G., Shapiro, J., Simes, J., Siu, L., Tebbutt, N., & Price, T. (2008). Outcomes of older patients ($\geq$ 70 years) treated with targeted therapy in metastatic chemorefractory colorectal cancer: Retrospective analysis of NCIC CTG CO.17 and CO.20. *Clinical Colorectal Cancer, 18*, e140–e149.