STATISTICAL ISSUES IN COMPARATIVE EFFECTIVENESS RESEARCH

by

Yi-Fan Chen

B.S. National Chiao Tung University, Taiwan (ROC), 2004

M.S. National Yang-Ming University, Taiwan (ROC), 2006

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Yi-Fan Chen

It was defended on

June 19, 2013

and approved by

Dissertation Advisor:
Lisa A. Weissfeld, PhD, Professor
Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Committee Members:
Chung-Chou H. Chang, PhD, Professor
Department of Medicine, School of Medicine
University of Pittsburgh

Jong-Hyeon Jeong, PhD, Associate Professor
Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Sachin Yende, MD, MS, Associate Professor
Department of Critical Care Medicine, School of Medicine
University of Pittsburgh

Lisa A. Weissfeld, PhD

STATISTICAL ISSUES IN COMPARATIVE EFFECTIVENESS RESEARCH

Yi-Fan Chen, PhD

University of Pittsburgh, 2013

ABSTRACT

The goal of this research is to provide empirical results that can be used to guide decisions regarding treatments/interventions. This work focuses on two different problems of interest in comparative effectiveness research. The first problem is to understand if the proportion of an event changes over time, when the populations are nested within each other. This often happens in the health care system and is illustrated here by a study of hospital readmissions within 48 hours of a first visit to an emergency department (ED). The nested structure of the data must be taken into account at the analysis stage and there are no standard statistical methods for doing this. We propose a likelihood ratio test based on the product of conditional probabilities in the form of generalized mixed model. This test accommodates conditionality, within subject dependence and between hospital cluster effects. Simulations show that it preserves the type-I error level given no difference, and provides estimates that are less biased in the presence of a large cluster effect. This approach can be implemented using SAS PROC NLMIXED making it easy to apply in this setting.

The second problem focuses on the identification of subgroups within a clinical trial, with the goal being the identification of subjects who benefit from the treatment of interest. The focus is on the use of interaction trees which are an extension of the classification and regression trees (CART). The use of interaction trees overcomes both the subjectivity and multiple comparisons that plague a conventional subgroup analysis. However, the method is greedy in finding each local node by exhausting every predictor and its available values. We propose a greediness reduction interaction tree (GRIT) algorithm that integrates random forests and the evolutionary algorithm into the interaction trees. Simulations show that this proposed method outperforms the interaction trees without accessing every predictor given the interaction. The strengths of the proposed method are demonstrated through a real data example from the Biological Markers for Recovery of Kidney (BioMaRK) study. **Public Health Significance:** Two methodologies proposed provide less bias and more accurate information under certain circumstances. One is for medical and public policy decisions based on administrative datasets and the other is for finding subgroups and generating hypotheses for future clinical trials.

Keywords: evolutionary/genetic algorithm; interaction trees; likelihood ratio test; nested/conditional proportions; non-linear mixed model; random forests; subgroup analysis

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

The area of comparative effectiveness research has become of great importance in recent years. The goal of this research is to provide empirical results that can be used to guide decisions regarding new and/or currently existing treatments. The gold standard for decision making is the clinical trial, where subjects are randomly assigned to receive either a new or currently existing treatment. The advantage of this methodology is that the randomization insures that the subjects are comparable across the groups and that the treatment was administered in a standardized fashion. However, clinical trials are often not feasible for a variety of reasons such as cost, feasibility, ethics, time, and the lack of sufficient information to guide selection of best treatments. Thus researchers have often turned to large administrative data bases to glean information about current and potentially new treatments for diseases, to better understand the patient population, and to gain information on individuals who fail treatment.

This work focuses on two different problems that are of interest in comparative effectiveness research. The first of these problems is to understand the "failures" in a given population. This group is defined as subjects who "fail" within the health care system and is illustrated by a study of hospital readmissions within 48 hours of a first visit to an emergency department (ED). The goal is to understand the readmissions within this population. The second problem focuses on the identification of subgroups within a clinical trial, with the goal being the identification of subjects who benefit from the treatment of interest.

Many different statistical methods are available for analyzing datasets in the comparative effectiveness area. Two widely used parametric methods, which we focus on in the first part of this work, is the general mixed model and the likelihood ratio test. The general mixed model [Brown et al., 2006] is useful due to its ability to analyze correlated data, including repeated measures, and clustered data. This model is composed of fixed and random effects leading to the "mixed" model. Here, a fixed effect is the average effect and a random effect allows different intercepts or slopes across units that is use to account for dependence within a single unit. A normal distribution is used to account for the correlation structure induced by the multiple observations within a single unit. Based on the normal assumption, a likelihood function is formed and estimates are obtained using either maximum likelihood estimation (MLE) or restricted maximum likelihood estimation (REML) [Brown et al., 2006]. This approach has been extended to include other distributions, such as the binominal distribution, following the same idea used in the normal case, except that the link function between the outcome and the linear combination of effects is no longer the identity link. In this case, the model is referred to as a nonlinear/generalized mixed model.

The likelihood ratio test [Pawitan, 2001], however, is used for hypothesis testing, within the likelihood framework. It can be applied as long as two likelihood functions exist and one is nested within the other. In other words, the distribution of a variable related to the hypothesis of interest should be identified beforehand. Then the difference between two likelihood functions, where one satisfies the alternative hypothesis and the other does not, follows a chi-squared distribution. Note that the common basis for both the mixed model and the likelihood ratio test is the existence of a likelihood function. This makes it is easy to apply the likelihood ratio test to a mixed model as long as a likelihood function related to the hypothesis of interest can be computed. Moreover, both can be implemented in standard statistical software. In SAS (SAS Institute Inc., Cary, NC, USA), PROC NLMIXED allows for a user-defined likelihood function and hence is more flexible for a less commonly used model and a complicated data structure.

When using parametric methods, however, since all of these advantages heavily rely on the distributional assumption, they may be invalid if these assumptions are violated. This leads to the usage of non-parametric methods which do not require the assumption of a parametric distribution and are data-driven analyses. Yet, unlike parametric methods, it is tough to justify a non-parametric method through a theoretical derivation, and the implementation may require expertise in programming. One popular example is tree analysis which is often used for the classification of subjects into meaningful groups.

Tree modeling [Breiman et al., 1984], the focus of the second part of this work, is a method for clustering data into similar groups based on a set of covariates for predicting a given outcome. It is very powerful in finding complex interactions using automated techniques. Without specifying any form or predictors beforehand, it detects the best predictor to split one group of subjects into two, such that within each group subjects are as homogenous as possible with respect to the outcome of interest. These groups then form the next branches in the tree and the method finds the next predictor that best separates each subgroup, until certain stopping criteria are met. The most famous splitting criterion is the Gini index, which is similar in concept to the likelihood function but does not rely on any distributions. The stopping criteria could be the restriction of the minimum number of subjects in one node or the maximum depth of a tree etc. After pruning and validating the tree, the result of the analysis is a tree structure showing different paths with nodes and their splitting criteria and at the end of each path is the terminal node indicating the number of subjects in the node and the outcome. Many of the tree-related methods that have been developed over the years are an extension of the simple tree analysis, such as random forests [Breiman, 2001]. However, because the software is only available in R (R Development Core Team, 2012) or other commercial software, it is not easily applied by a non-biostatistician. In spite of these limitations the methods are now getting more attention by clinicians.

In this dissertation, we propose one parametric method and one non-parametric method to solve different clinical issues due to different data structures and study designs. For the first part, i.e., Chapter 2, the clinical issue is to test if the admitted rates in the emergency department (ED) in 30 hospitals are different between the first and the return visits. It is an observational and sequentially collected dataset with a binary outcome and cluster effects. We propose a likelihood ratio test for comparing two nested proportions based on the product of conditional probabilities in the form of a generalized mixed model. The SAS procedure PROC NLMIXED is used to demonstrate the method. In the second part, i.e., Chapter 3, a greediness reduction interaction tree (GRIT) algorithm is proposed to lessen the greediness in interaction trees for subgroup analysis in clinical trials. Here, the algorithm of interaction trees is an extension of the conventional tree analysis, except that it uses the test statistic of an interaction as the splitting criterion. The idea is to incorporate randomness in the interaction trees by using the genetic evolutionary algorithm and random forests. An R program is written for implementing the method. At the end of this dissertation, an overall summary and future work for these two parts is given in Chapter 4.

2.0 A LIKELIHOOD RATIO TEST FOR NESTED PROPORTIONS

2.1 INTRODUCTION

The idea of comparing two nested proportions is not uncommon and can be useful for making effective medical and health and public policy decisions. For example, it is important to compare the divorce rates between first marriages and subsequent marriages [Clarke et al., 1994; Chadwick et al., 1999] just as it is critical to understand the rates of successful remission for patients who undergo initial chemotherapy and the same chemotherapy undertake as a second line option, or initial revascularization for significant coronary lesions and repeat revascularization for the same lesion. With greater numbers of registry and large administrative databases available for analysis, similar questions involving an outcome comparison of a group that experiences a defined situation and the subset that again faces that same situation will arise. However, based on our knowledge, a valid statistical testing method is not available for this circumstance.

When two proportions are collected successively from the same group of subjects, the statistical method should incorporate the dependence within a given subject. A simple two-sample test for proportions does not incorporate this dependence, and a paired test for proportions cannot be utilized if only grouped data are available for the analysis. If we frame this as an analysis from a cross-over design, a longitudinal repeated measures analysis, or a recurrent event analysis, we encounter a large number of missing values for the second time point, because many or even most subjects have only one event. Moreover, it is unnecessary to utilize an incomplete data analysis approach, because we are not interested in estimating the outcomes for the second time point for those individuals without any subsequent events (e.g. marriages, diagnoses, recurrences, diagnoses). Thus, repeated measures analyses are unable to appropriately capture the data structure of the conditionality or nesting in a single model. Although study designs, such as the outcome-dependent sampling [Zhou et al., 2007] or the case-cohort study design [Prentice, 1986], collect only a subset of data from a whole cohort, the conditionality is based on a manipulated sampling scheme. This predefined sampling probability is then used in the statistical analysis to draw inferences for the entire population. The subset data in our study, however, is observed, based on the occurrence of the event of interest, and recovering the information for the whole population is not of interest. Without an appropriate and direct statistical method, most researchers explore these issues by presenting descriptive statistics and a simple *t*-test or stratification and lack the necessary methodology for statistical modeling or testing [Alessandrini et al., 2004; Clarke et al., 1994; Chadwick et al., 1999].

This study is motivated by a clinical and hospital policy question about hospital admission procedures for emergency department (ED) visits. In particular, the investigators wanted to compare the hospital admission rates between the first/index emergency department visit and any ED visits that occurred within 48 hours of the first visit in 30 hospitals. Three hypotheses were of interest (1) the hospital admission rate at the index ED visit is different from the hospital admission rate at the 48-hour return ED visit for patients who were admitted at the first visit; (2) the hospital admission rate at the index ED visit is different from the hospital admission rate at the 48-hour return ED visit for patients who were discharged at the first visit,

(3) the hospital admission rate at the index ED visit is different from the hospital admission rate among all return visits within 48 hours. Figure 1 depicts this process of coming to the ED for care and admission to the hospital. Considering these hypotheses, three statistical issues should be considered. First, these two rates are nested, because only patients who returned to the ED contribute to the second rate. Second, for those patients who came to the ED twice, the dependence within a subject needs to be taken into account in any analyses. Finally, the dataset consists of 30 hospitals, and so the method should also account for the cluster effect within the hospitals.

The goal of this study is to propose a method for performing the comparison of nested proportions, while simultaneously accounting for the conditional structure, the within-subject dependence, and the hospital cluster effect. This method is based on the conditional probability and the likelihood ratio test. It can be easily extended to more than two layers, to individual patient data, and to alternative outcome variables through the use of other distribution functions. Standard statistical software is available for this type of modeling and we will demonstrate how it can be used to implement the proposed method. The organization of this article is as follow. We describe the notation and our proposed method in Section 2.2-2.3 and evaluate its performance via simulation in Section 2.4. The ED data described previously is used as an example to demonstrate the method in Section 2.5. Finally, the discussion is presented in Section 2.6.

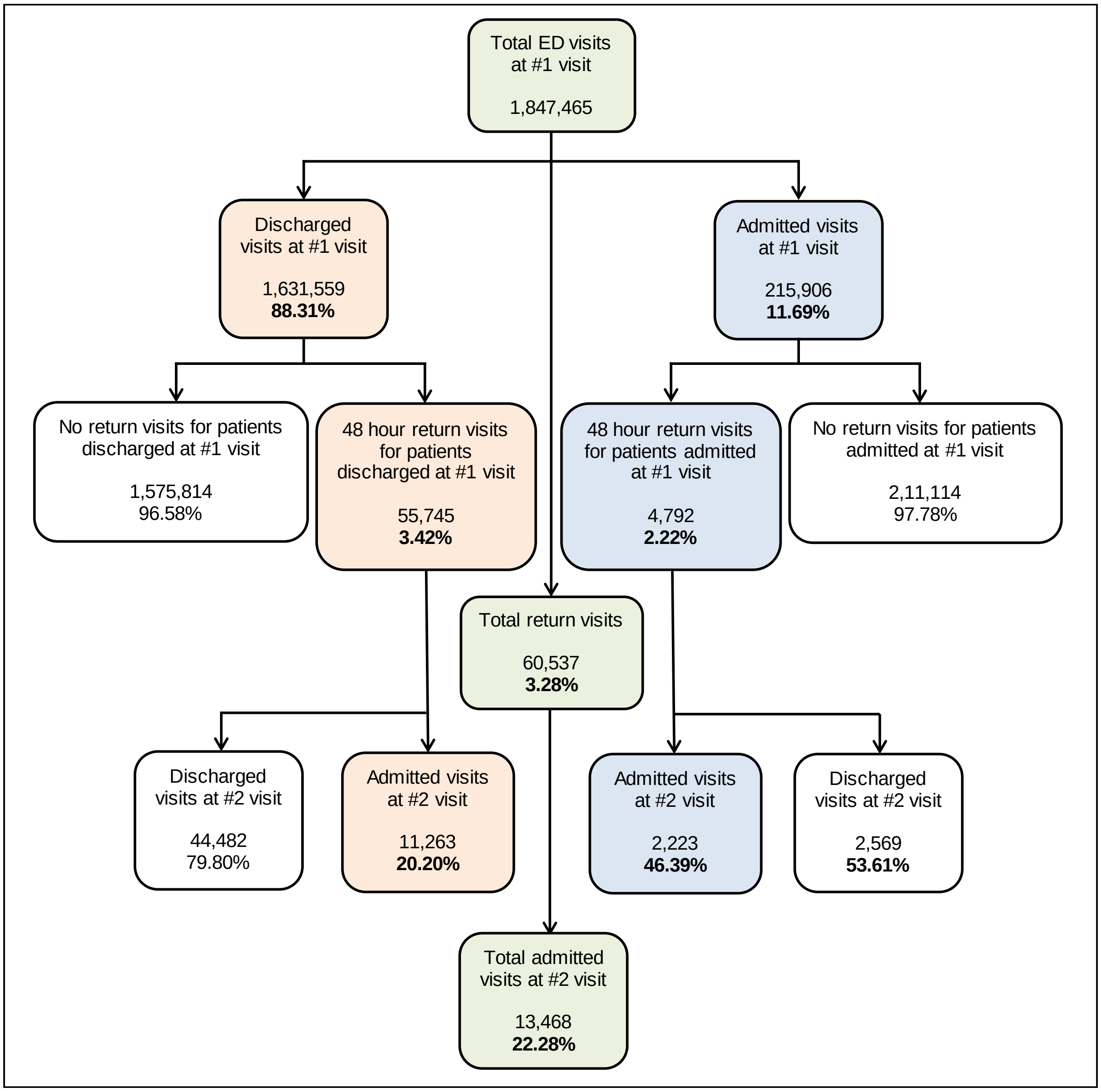

Figure 1. Schematic diagram of the ED data

2.2 NOTATION

Let $j = 1,...,M$ denote the hospital with the total number of hospitals being M , and $i = 1,...,n_j$ be the subject with n_j as the number of subjects within j^{th} hospital. The total number of subjects is $N = \sum_{j=1}^{M} n_j$. We assume a Bernoulli distribution for each admission and return with corresponding proportion parameters described below, and that the hospital cluster effect follows a normal distribution, $b_j \sim N(0,\sigma^2)$. For simplicity, we only illustrate the proposed method in the framework of the first hypothesis, where we compare the admission rates between the index visit and the return visit for patients admitted at the index visit. The same idea is applied to the other two hypotheses and possibly to even more general scenarios. To generalize the notation, we number the visits, instead of using the index visit and the return visit, to accommodate cases with more than two visits. We denote these visits as follows:

$$
\begin{aligned}
A_{1ij} &= I\{\text{Admission at visit 1 for } ith \text{ patient in hospital } j\} \sim Bernoulli(p_1) \\
R_{ij} &= I\{\text{Return visit for } ith \text{ patient in hospital } j\} \sim Bernoulli(p_r) \\
A_{2ij} &= I\{\text{Admission at visit 2 for } ith \text{ patient in hospital } j\} \sim Bernoulli(p_2) \\
R_{ij}^{(1)} &= I\left\{\text{Return visit for } ith \text{ patient in hospital } j \middle| A_{1ij} = 1\right\} \sim Bernoulli(p_r^{(1)}) \\
A_{2ij}^{(1)} &= I\left\{\text{Admission at visit 2 for } ith \text{ patient in hospital } j \middle| R_{ij}^{(1)} = 1\right\} \sim Bernoulli(p_2^{(1)}) \\
R_{ij}^{(0)} &= I\left\{\text{Return visit for } ith \text{ patient in hospital } j \middle| A_{1ij} = 0\right\} \sim Bernoulli(p_r^{(0)}) \\
A_{2ij}^{(0)} &= I\left\{\text{Admission at visit 2 for } ith \text{ patient in hospital } j \middle| R_{ij}^{(0)} = 1\right\} \sim Bernoulli(p_2^{(0)})
\end{aligned}
$$

The corresponding schematic plot depicting the quantities listed above is shown below.

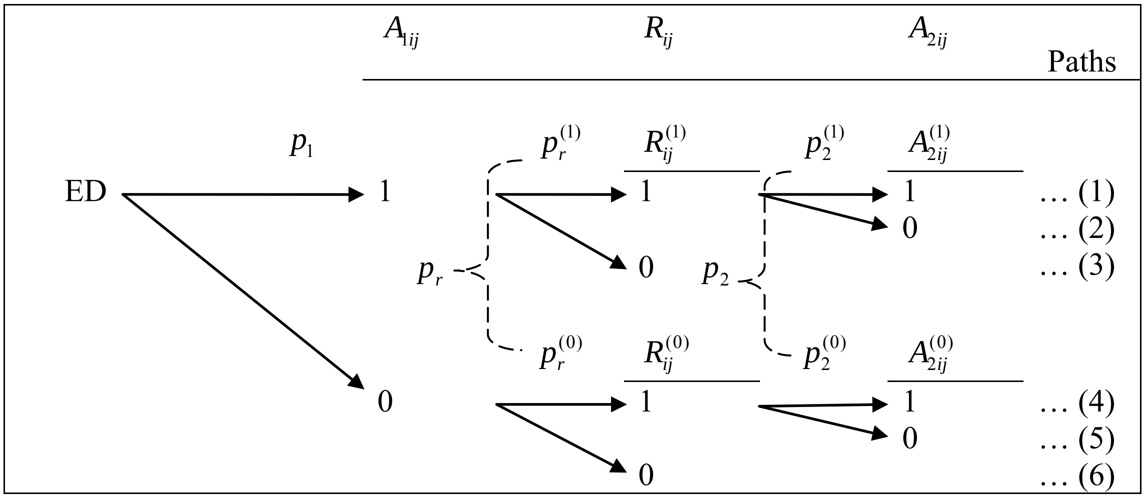

Figure 2. Sample paths associated with the ED Data

2.3 LIKELIHOOD RATIO TEST

2.3.1 Likelihood function

To incorporate the nested structure, we construct the likelihood function for each subject as the product of conditional probabilities, which correspond to the process of coming to the emergency department for care and admission to the hospital. The likelihood function for each patient is

$$f\left(A_{1ij}, R_{ij}, A_{2ij}, b_j; \sigma^2\right) = f\left(b_j \middle| \sigma^2\right) f\left(A_{1ij} \middle| b_j\right) f\left(R_{ij} \middle| b_j, A_{1ij}\right) f\left(A_{2ij} \middle| b_j, A_{1ij}, R_{ij}\right).$$

At the right hand side of this function, the first term represents the likelihood of the hospital cluster effect, and second one is the likelihood of the first visit admission rate for a subject given the cluster effect. Likewise, the third term describes the likelihood of return after knowing a patient's admission status at the first visit, and the last term is the likelihood of admission at the second visit for a subject given the admission status at the first visit and the return visit. This function can take account for both the within-patient fixed-effect dependence between the ED visits and the cluster random-effect within the hospital. For the hypothesis $H_0 : p_1 = p_2^{(1)}\ vs. H_A : p_1 \neq p_2^{(1)}$, we assume the following:

(1) $A_1, R\big|\left(A_1 = 1\right), A_2\big|\left(A_1 = 1\,\&\, R = 1\right) \perp R\big|\left(A_1 = 0\right)$ and

(2) $A_1, R\big|\left(A_1 = 1\right), A_2\big|\left(A_1 = 1\,\&\, R = 1\right) \perp A_2\big|\left(A_1 = 0\,\&\, R = 1\right)$

to avoid the correlation between the two paths of being admitted or not at the first visit and to simplify the likelihood function. Therefore, the likelihood function for the target sample related to p_1 and $p_2^{(1)}$, i.e. paths (1)-(3), can be simply written as

$$
\begin{aligned}
L\left(p_1, p_r^{(1)}, p_2^{(1)}, b_j, \sigma^2\right) &= \prod_{j=1}^{M}\prod_{i=1}^{n_j} f\left(A_{1ij}, R_{ij}, A_{2ij}, b_j; \sigma^2\right) \\
&= \prod_{j=1}^{M}\prod_{i=1}^{n_j} f\left(b_j\middle|\sigma^2\right) f\left(A_{1ij} = a_{1ij}\middle|b_j\right) f\left(R_{ij} = r_{ij}\middle|b_j, A_{1ij}\right) f\left(A_{2ij} = a_{2ij}\middle|b_j, A_{1ij}, R_{ij}\right) \\
&\propto \prod_{j=1}^{M}\prod_{i=1}^{n_j} \left\{ \begin{array}{l} \left\{ \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{b_j^2}{2\sigma^2}\right) \right\} \left\{ p_1^{a_{1ij}} \left(1 - p_1\right)^{\left(1 - a_{1ij}\right)} \right\} \\ \times \left\{ p_r^{(1) r_{ij}} \left(1 - p_r^{(1)}\right)^{\left(1 - r_{ij}\right)} \right\}^{a_{1ij}} \left\{ p_2^{(1) a_{2ij}} \left(1 - p_2^{(1)}\right)^{\left(1 - a_{2ij}\right)} \right\}^{a_{1ij} r_{ij}} \end{array} \right\},
\end{aligned}
$$

which only uses the information from the target sample by the restriction of indicators. Also notice that the hospital random effect is incorporated in the probabilities as

$$
p_1 = 1 \big/ \left\{1 + \exp\left\{-\left(\beta_{0,1} + b_j\right)\right\}\right\},\ p_r^{(1)} = 1 \big/ \left\{1 + \exp\left\{-\left(\beta_{0,r}^{(1)} + b_j\right)\right\}\right\} \text{ and}
$$

$$
p_2^{(1)} = 1 \big/ \left\{1 + \exp\left\{-\left(\beta_{0,2}^{(1)} + b_j\right)\right\}\right\},
$$

based on a generalized mixed model with a random intercept for a binary outcome when no covariate is included, but only the intercept, $\beta_{0,1}, \beta_{0,r}^{(1)}$, and $\beta_{0,2}^{(1)}$. This model can be extended to include covariates in order to make an adjusted analysis possible and will be discussed later.

Based on our experience, however, using the full data should be more efficient than using a restricted sample when estimating the cluster effect. Therefore, the following likelihood function for the entire sample comprising paths (1)-(6) is preferred.

$$L\left(p_1, p_r^{(1)}, p_2^{(1)}, p_r^{(0)}, p_2^{(0)}, b_j, \sigma^2\right) = \prod_{j=1}^{M}\prod_{i=1}^{n_j}\left\{\begin{array}{l} \left\{\dfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\dfrac{b_j^2}{2\sigma^2}\right)\right\}\left\{p_1^{a_{1ij}}\left(1-p_1\right)^{\left(1-a_{1ij}\right)}\right\} \\ \times\left\{p_r^{(1)r_{ij}}\left(1-p_r^{(1)}\right)^{\left(1-r_{ij}\right)}\right\}^{a_{1ij}}\left\{p_2^{(1)a_{2ij}}\left(1-p_2^{(1)}\right)^{\left(1-a_{2ij}\right)}\right\}^{a_{1ij}r_{ij}} \\ \times\left\{p_r^{(0)r_{ij}}\left(1-p_r^{(0)}\right)^{\left(1-r_{ij}\right)}\right\}^{\left(1-a_{1ij}\right)}\times\left\{p_2^{(0)a_{2ij}}\left(1-p_2^{(0)}\right)^{\left(1-a_{2ij}\right)}\right\}^{\left(1-a_{1ij}\right)r_{ij}} \\ \end{array}\right\}$$

2.3.2 Likelihood ratio test

To test the hypothesis, $H_0 : p_1 = p_2^{(1)}\ vs. H_A : p_1 \neq p_2^{(1)}$, the likelihood ratio statistic is

$$\lambda = \frac{L\left(p_1 = p_2^{(1)} = \hat{p}, \hat{p}_r^{(1)}, \hat{p}_r^{(0)}, \hat{p}_2^{(0)}, \hat{b}_j, \hat{\sigma}^2\right)}{L\left(\hat{p}_1, \hat{p}_2^{(1)}, \hat{p}_r^{(1)}, \hat{p}_r^{(0)}, \hat{p}_2^{(0)}, \hat{b}_j, \hat{\sigma}^2\right)}$$

$$= \frac{\prod_{j=1}^{M}\prod_{i=1}^{n_j}\left\{\hat{p}^{\left\{a_{1ij}+a_{2ij}a_{1ij}r_{ij}\right\}}\left(1-\hat{p}\right)^{\left\{1-a_{1ij}+\left(1-a_{2ij}\right)a_{1ij}r_{ij}\right\}}\right\}}{\prod_{j=1}^{M}\prod_{i=1}^{n_j}\left\{\hat{p}_1^{a_{1ij}}\left(1-\hat{p}_1\right)^{\left(1-a_{1ij}\right)}\right\}\left\{\hat{p}_2^{(1)a_{2ij}}\left(1-\hat{p}_2^{(1)}\right)^{\left(1-a_{2ij}\right)}\right\}^{a_{1ij}r_{ij}}},$$

after simplifying the following equation

$$\lambda = \frac{L\left(p_1 = p_2^{(1)} = \hat{p}, \hat{p}_r^{(1)}, \hat{p}_r^{(0)}, \hat{p}_2^{(0)}, \hat{b}_j, \hat{\sigma}^2\right)}{L\left(\hat{p}_1, \hat{p}_2^{(1)}, \hat{p}_r^{(1)}, \hat{p}_r^{(0)}, \hat{p}_2^{(0)}, \hat{b}_j, \hat{\sigma}^2\right)}$$

$$= \frac{\prod_{j=1}^{M}\prod_{i=1}^{n_j}\left\{\begin{array}{l} \left\{\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\exp\left(-\frac{b_j^2}{2\hat{\sigma}^2}\right)\right\}\left\{\hat{p}^{a_{1ij}}\left(1-\hat{p}\right)^{\left(1-a_{1ij}\right)}\right\} \\ \times\left\{\hat{p}_r^{(1)r_{ij}}\left(1-\hat{p}_r^{(1)}\right)^{\left(1-r_{ij}\right)}\right\}^{a_{1ij}}\left\{\hat{p}^{a_{2ij}}\left(1-\hat{p}^{a_{2ij}}\right)^{\left(1-a_{2ij}\right)}\right\}^{a_{1ij}r_{ij}} \\ \times\left\{\hat{p}_r^{(0)r_{ij}}\left(1-\hat{p}_r^{(0)}\right)^{\left(1-r_{ij}\right)}\right\}^{\left(1-a_{1ij}\right)}\times\left\{\hat{p}_2^{(0)a_{2ij}}\left(1-\hat{p}_2^{(0)}\right)^{\left(1-a_{2ij}\right)}\right\}^{\left(1-a_{1ij}\right)r_{ij}} \end{array}\right\}}{\prod_{j=1}^{M}\prod_{i=1}^{n_j}\left\{\begin{array}{l} \left\{\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\exp\left(-\frac{b_j^2}{2\hat{\sigma}^2}\right)\right\}\left\{\hat{p}_1^{a_{1ij}}\left(1-\hat{p}_1\right)^{\left(1-a_{1ij}\right)}\right\} \\ \times\left\{\hat{p}_r^{(1)r_{ij}}\left(1-\hat{p}_r^{(1)}\right)^{\left(1-r_{ij}\right)}\right\}^{a_{1ij}}\left\{\hat{p}_2^{(1)a_{2ij}}\left(1-\hat{p}_2^{(1)}\right)^{\left(1-a_{2ij}\right)}\right\}^{a_{1ij}r_{ij}} \\ \times\left\{\hat{p}_r^{(0)r_{ij}}\left(1-\hat{p}_r^{(0)}\right)^{\left(1-r_{ij}\right)}\right\}^{\left(1-a_{1ij}\right)}\times\left\{\hat{p}_2^{(0)a_{2ij}}\left(1-\hat{p}_2^{(0)}\right)^{\left(1-a_{2ij}\right)}\right\}^{\left(1-a_{1ij}\right)r_{ij}} \end{array}\right\}}$$

and then computing $-2\log\lambda \sim \chi_1^2$.

2.3.3 Likelihood function and likelihood ratio test in terms of relative risk

From the prospective of researchers, a relative risk is more interpretable than two separate probabilities. Let $RR^{(1)} = p_2^{(1)}/p_1$. The likelihood function is

$$L\left(p_1, RR^{(1)}, p_r^{(1)}, p_r^{(0)}, p_2^{(0)}, b_j, \sigma^2\right)$$

$$= \prod_{j=1}^{M}\prod_{i=1}^{n_j}\left\{\begin{array}{l} \left\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{b_j^2}{2\sigma^2}\right)\right\}\left\{p_1^{a_{1ij}}\left(1-p_1\right)^{\left(1-a_{1ij}\right)}\right\} \\ \times\left\{p_r^{(1)r_{ij}}\left(1-p_r^{(1)}\right)^{\left(1-r_{ij}\right)}\right\}^{a_{1ij}}\left\{\left(p_1RR^{(1)}\right)^{a_{2ij}}\left(1-p_1RR^{(1)}\right)^{\left(1-a_{2ij}\right)}\right\}^{a_{1ij}r_{ij}} \\ \times\left\{p_r^{(0)r_{ij}}\left(1-p_r^{(0)}\right)^{\left(1-r_{ij}\right)}\right\}^{\left(1-a_{1ij}\right)}\times\left\{p_2^{(0)a_{2ij}}\left(1-p_2^{(0)}\right)^{\left(1-a_{2ij}\right)}\right\}^{\left(1-a_{1ij}\right)r_{ij}} \end{array}\right\}.$$

The likelihood ratio test for $H_0 : RR^{(1)} = 1 \, vs. \, H_A : RR^{(1)} \neq 1$ is

$$\lambda = \frac{L\left(p_1 = \hat{p}, R\hat{R}^{(1)} = 1, p_r^{(1)}, \hat{p}_r^{(0)}, \hat{p}_2^{(0)}, \hat{b}_j, \hat{\sigma}^2\right)}{L\left(\hat{p}_1, R\hat{R}^{(1)}, \hat{p}_r^{(1)}, \hat{p}_r^{(0)}, \hat{p}_2^{(0)}, \hat{b}_j, \hat{\sigma}^2\right)}$$

$$= \frac{\prod_{j=1}^{M}\prod_{i=1}^{n_j}\left\{\hat{p}^{\left\{a_{1ij}+a_{2ij}a_{1ij}r_{ij}\right\}}\left(1-\hat{p}\right)^{\left\{1-a_{1ij}+\left(1-a_{2ij}\right)a_{1ij}r_{ij}\right\}}\right\}}{\prod_{j=1}^{M}\prod_{i=1}^{n_j}\left\{\hat{p}_1^{a_{1ij}}\left(1-\hat{p}_1\right)^{\left(1-a_{1ij}\right)}\right\}\left\{\left(\hat{p}_1 R\hat{R}^{(1)}\right)^{a_{2ij}}\left(1-\left(\hat{p}_1 R\hat{R}^{(1)}\right)\right)^{\left(1-a_{2ij}\right)}\right\}^{a_{1ij}r_{ij}}}$$

and $-2\log\lambda \sim \chi_1^2$, which gives exactly the same result as using the probabilities.

2.3.4 Property and estimation of the proposed method

The essence of the proposed method is based on the conventional likelihood ratio test, and therefore this method inherits all of the properties of the likelihood ratio test. When no random effects are identified, the maximum likelihood estimator (MLE) is valid for estimation. When a random effect is included, the MLE with adaptive Gauss-Hermite quadrature is used. These estimators have all of the nice statistical properties of the MLE, and can be easily calculated in SAS (SAS Institute Inc., Cary, NC, USA) by using PROC NLMIXED. Sample code is provided in the Appendix A.

2.4 SIMULATION STUDY

We evaluate the performance of the proposed method via simulation with respect to two properties in comparison to the ordinary two-sample test for proportion -- a naïve method that ignores the complexity and dependence inherent in the data structure. First, we investigate the bias of the estimated parameters of interest and the corresponding standard error estimators. Second, we examine power and type I error. We study the properties of the proposed method under varying sample sizes, effect sizes, and clustering effects. Sample sizes were chosen to be 45,000 and 500 to represent, respectively, large and moderate datasets. Effect sizes were varied by specifying the true difference between p_1 and $p_2^{(1)}$ at 3 levels: severe, mild, and none. We studied 6 levels cluster effects: 0, 0.001, 0.01, 0.1, 0.5 and 1. These values were chosen as they resemble the data features in the example that will be described in detail later. The total number of hospitals, M , was set to 30 and 1000 datasets were simulated for each scenario. Details are found in Table 1.

Table 1. Simulation setting for testing nested proportions

Items	Values	
Hospital size: M	30	
Sample size: N	45,000	500
True difference: $\left(p_1, p_2^{(1)}\right)$	(0.46, 0.46) (0.30, 0.46) (0.12, 0.46)	(0.46, 0.46) (0.38, 0.46) (0.30, 0.46)
Cluster effect: σ^2	0 / 0.001 / 0.01 / 0.1 / 0.5 / 1	
$\left(p_r^{(1)}, p_r^{(0)}, p_2^{(0)}\right)$	(0.02, 0.03, 0.20)	(0.50, 0.55, 0.20)

Tables 2 and 3 show the simulation results for $N = 45{,}000$ and $N = 500$ respectively. With respect to consistent estimation, the estimators are unbiased in most of the probability parameters from the proposed method. The estimated standard errors are close to the standard deviations, and both increase from the first visit to the return and the second visit due to the declining sample size. Non-convergence, a common problem occurring when using PROC NLMIXED to fit models with random effects, is also encountered in some of the datasets. Notice that when the cluster correlation is very small, under 0.001, with a large sample size or 0.1 with a small sample size, the proposed method tends to be instable and provides a biased estimation for the relative risk (results are shown in the Appendix B). One solution for this is to remove the cluster effect from the likelihood function and results are presented in the tables for these cases. In contrast to the proposed method, the two-sample proportion test gives a biased relative risk when a cluster effect exists and the severity of the bias expands as the effect increases from 0.1. The estimated standard deviations and standard errors under the two-sample proportion test are for the relative risks in the natural log scale which approximately follow the normal distribution. However, the standard deviations and standard errors estimated from the proposed method are directly for the relative risks following the normal assumption under the MLE. Hence, they are not directly comparable between two methods.

With respect to the power, the two-sample proportion test exhibits a serious type I error when no difference exists. The type I error is 32% to 100% when the cluster effect is more than 0.1 and the sample size is large, and is 44% to 84% when the cluster effect is bigger than 0.5 and the sample size is small. In the same scenarios, the proposed method provides reasonable error rates, around 4-6%, regardless of the sample size. When the difference increases to a mild or large level in a large dataset, the proposed method and the proportion test perform equally well

Table 2. Simulation results comparing the proposed method to the ordinary test for proportions under different scenarios with N=45,000.[a,b]

	Proposed method														Two-sample proportion test				
		p_1			$p_r^{(1)}$			$p_2^{(1)}$			$RR^{(1)}$				$RR^{(1)}$				
Difference	Simulation Est/Power	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Power	Simu-lation	Bias	Est. SD[c]	Est. SE[d]	Power[e]
None: $(p_1, p_2^{(1)})=$ (0.46,0.46)																			
Cluster effect: **0**	**977**	**-0.01**	**0.24**	**0.23**	**<0.01**	**<0.01**	**0.10**	**0.04**	**2.44**	**2.45**	**0.09**	**5.33**	**5.35**	**0.044**	**1000**	**0.09**	**5.37**	**5.36**	**0.044**
0.001	**983**	**-0.01**	**0.27**	**0.23**	**<0.01**	**0.10**	**0.10**	**0.09**	**2.47**	**2.45**	**0.22**	**5.38**	**5.35**	**0.047**	**1000**	**0.22**	**5.41**	**5.35**	**0.046**
0.01	1000	-0.01	0.51	0.50	<0.01	0.11	0.10	-0.01	2.51	2.48	-0.01	5.40	5.33	0.053	1000	0.77	5.38	5.30	0.054
0.1	1000	-0.01	1.44	1.41	<0.01	0.15	0.14	<0.01	2.71	2.74	0.03	5.05	5.17	0.040	1000	7.53	5.05	4.78	0.324
0.5	949	-0.09	3.15	3.10	0.01	0.27	0.26	-0.07	3.69	3.76	0.06	4.59	4.71	0.043	1000	30.71	6.76	3.30	0.995
1	893	-0.13	4.45	4.35	0.02	0.36	0.36	-0.07	4.79	4.77	0.15	4.32	4.40	0.043	1000	49.75	8.47	2.34	1.000
Mild: $(p_1, p_2^{(1)})=$ (0.30,0.46)																			
Cluster effect: **0**	**999**	**<0.01**	**0.22**	**0.22**	**<0.01**	**0.13**	**0.12**	**0.01**	**3.01**	**3.03**	**0.04**	**10.11**	**10.17**	**1.000**	**1000**	**0.02**	**6.63**	**6.66**	**1.000**
0.001	**998**	**<0.01**	**0.25**	**0.22**	**<0.01**	**0.12**	**0.12**	**0.05**	**3.03**	**3.03**	**0.16**	**10.15**	**10.16**	**0.999**	**1000**	**0.13**	**6.65**	**6.65**	**0.999**
0.01	1000	<0.01	0.44	0.43	<0.01	0.13	0.12	-0.02	3.05	3.05	-0.06	10.20	10.12	1.000	1000	1.08	6.62	6.56	1.000
0.1	1000	<0.01	1.22	1.20	<0.01	0.16	0.16	-0.03	3.16	3.20	-0.04	9.67	9.80	1.000	1000	10.80	6.12	5.82	1.000
0.5	936	-0.02	2.69	2.62	<0.01	0.27	0.26	-0.10	3.93	3.99	0.07	8.91	9.13	1.000	1000	41.86	7.59	3.80	1.000
1	888	-0.04	3.74	3.72	0.01	0.36	0.36	-0.08	4.88	4.94	0.52	9.28	9.02	1.000	1000	63.61	9.48	2.60	1.000
Severe: $(p_1, p_2^{(1)})=$ (0.12,0.46)																			
Cluster effect: **0**	**980**	**<0.01**	**0.16**	**0.15**	**<0.01**	**0.19**	**0.19**	**-0.19**	**4.85**	**4.79**	**-1.57**	**40.79**	**40.23**	**1.000**	**1000**	**-1.17**	**10.84**	**10.64**	**1.000**
0.001	**985**	**0.01**	**0.17**	**0.15**	**<0.01**	**0.19**	**0.19**	**-0.12**	**4.82**	**4.79**	**-1.10**	**40.47**	**40.19**	**1.000**	**1000**	**-0.76**	**10.69**	**10.61**	**1.000**
0.01	997	0.01	0.25	0.24	<0.01	0.19	0.19	-0.07	4.68	4.78	-0.74	39.19	40.00	1.000	1000	1.50	10.31	10.44	1.000
0.1	1000	0.02	0.63	0.61	<0.01	0.21	0.21	-0.03	4.68	4.69	-0.56	38.23	38.49	1.000	1000	22.52	9.44	9.04	1.000
0.5	984	0.06	1.37	1.34	0.01	0.29	0.28	0.04	4.86	4.82	0.21	34.34	35.76	1.000	1000	76.56	9.40	5.39	1.000
1	915	0.03	1.90	1.88	0.01	0.37	0.37	-0.05	5.45	5.42	2.25	37.21	36.65	1.000	1000	95.45	11.72	3.44	1.000

[a] Bias, SD and SE are in the scale of 100 times.

[b] Rows in bold are from likelihood functions without random effects.

[c] Estimated standard deviation of ln(RR)

[d] Estimated standard error of ln(RR)

[e] This is the type I error under the scenario of no difference.

Table 3. Simulation results comparing the proposed method to the ordinary test for proportions under different scenarios with N=500.[a,b]

	Proposed method														Two-sample proportion test				
		p_1			$p_r^{(1)}$			$p_2^{(1)}$			$RR^{(1)}$				$RR^{(1)}$				
Difference	Simulation Est/Power	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Power	Simulation	Bias	Est. SD[c]	Est. SE[d]	Power[e]
None: $(p_1, p_2^{(1)})=$ (0.46,0.46)																			
Cluster effect: **0**	**1000**	**0.12**	**2.24**	**2.23**	**0.01**	**3.19**	**3.29**	**0.16**	**4.53**	**4.63**	**0.33**	**11.10**	**11.21**	**0.056**	**1000**	**0.33**	**11.07**	**11.23**	**0.055**
0.001	**997**	**0.13**	**2.25**	**2.23**	**0.01**	**3.17**	**3.29**	**0.12**	**4.47**	**4.63**	**0.22**	**11.00**	**11.20**	**0.050**	**1000**	**0.25**	**10.99**	**11.24**	**0.050**
0.01	**999**	**0.13**	**2.25**	**2.23**	**0.19**	**3.20**	**3.29**	**0.35**	**4.49**	**4.63**	**0.71**	**10.94**	**11.20**	**0.042**	**1000**	**0.71**	**10.86**	**11.18**	**0.042**
0.1	939	0.14	2.65	2.66	0.10	3.63	3.65	0.14	4.69	4.90	0.27	11.12	11.33	0.042	1000	5.25	10.40	10.67	0.065
0.5	962	-0.06	3.78	3.89	-0.05	4.39	4.71	-0.07	5.50	5.67	0.21	11.17	11.58	0.043	1000	19.84	9.90	9.32	0.439
1	912/913	-0.15	4.82	4.99	-0.04	5.51	5.73	0.06	6.31	6.49	0.72	11.68	11.91	0.048	1000	32.06	9.97	8.35	0.838
Mild: $(p_1, p_2^{(1)})=$ (0.38,0.46)																			
Cluster effect: **0**	**1000**	**0.10**	**2.15**	**2.17**	**0.08**	**3.50**	**3.62**	**0.27**	**4.99**	**5.09**	**0.80**	**15.07**	**15.15**	**0.286**	**1000**	**0.80**	**12.34**	**12.52**	**0.283**
0.001	**998/997**	**0.10**	**2.15**	**2.17**	**0.09**	**3.50**	**3.62**	**0.24**	**4.91**	**5.09**	**0.72**	**14.90**	**15.15**	**0.303**	**1000**	**0.74**	**12.20**	**12.52**	**0.289**
0.01	**997**	**0.17**	**2.21**	**2.17**	**0.24**	**3.56**	**3.61**	**0.46**	**4.95**	**5.08**	**1.08**	**14.97**	**15.11**	**0.318**	**1000**	**1.09**	**12.20**	**12.45**	**0.307**
0.1	958/957	0.14	2.55	2.57	0.05	3.94	3.95	0.23	5.10	5.33	0.60	15.03	15.25	0.299	1000	6.43	11.73	11.84	0.451
0.5	938/936	0.05	3.70	3.72	0.05	4.64	4.96	0.06	5.75	6.00	0.54	15.09	15.43	0.301	1000	22.91	10.87	10.19	0.883
1	828/826	0.38	4.53	4.73	0.44	5.67	5.88	0.58	6.36	6.72	0.92	15.16	15.58	0.312	1000	35.91	10.87	9.05	0.984
Severe: $(p_1, p_2^{(1)})=$ (0.30,0.46)																			
Cluster effect: **0**	**1000**	**0.09**	**2.07**	**2.05**	**0.07**	**3.98**	**4.07**	**0.24**	**5.66**	**5.73**	**1.04**	**21.71**	**21.93**	**0.780**	**1000**	**1.04**	**14.06**	**14.32**	**0.752**
0.001	**1000**	**0.09**	**2.05**	**2.05**	**0.07**	**3.99**	**4.07**	**0.27**	**5.55**	**5.73**	**1.14**	**21.45**	**21.94**	**0.777**	**1000**	**1.14**	**13.85**	**14.31**	**0.758**
0.01	**1000**	**0.14**	**2.09**	**2.05**	**0.27**	**4.07**	**4.07**	**0.51**	**5.65**	**5.72**	**1.72**	**21.88**	**21.89**	**0.789**	**1000**	**1.72**	**14.07**	**14.23**	**0.773**
0.1	951/950	0.08	2.46	2.39	-0.02	4.35	4.38	0.16	5.85	5.93	0.97	22.46	22.04	0.787	1000	8.14	13.55	13.47	0.878
0.5	885	0.14	3.40	3.37	0.22	5.13	5.28	0.35	6.23	6.45	1.51	21.96	22.06	0.806	1000	26.99	12.19	11.38	0.996
1	713	0.46	4.43	4.24	0.44	6.30	6.12	0.57	6.99	7.05	1.02	22.41	22.21	0.788	1000	40.04	12.31	9.99	1.000

[a] Bias, SD and SE are in the scale of 100 times.

[b] Rows in bold are from likelihood functions without random effects.

[c] Estimated standard deviation of ln(RR)

[d] Estimated standard error of ln(RR)

[e] This is the type I error under the scenario of no difference.

with almost 100% power. However, as the sample size decreases, the proposed method only has a 29% to 32% and 78% to 81% power to discover a mild and severe difference respectively, while the t-test has a stronger power when the large random effect is large. Nonetheless, the proportion test pays off by the highly biased estimated relative risk.

2.5 EXAMPLE

Our example is a retrospective cohort study of 30 children's hospital emergency departments using the Pediatric Health Information System (PHIS) 2009 database. The data contains inpatient and ED data for patients < 18 years of age. We study the number of index visits, return visits, and admitted and discharged rates for each visit. To be specific, the index visit is defined as a visit without a prior ED visit within 48-hrs, while the return visit is a visit within 48 hours after the index visit. The admission rate is the proportion of subjects admitted to the hospital among the total ED visits for the index visit, and is the proportion of the subjects admitted to the hospital among those who returned, either for patients who were discharged or admitted at the index ED visit. The clinical question of interest was to investigate the quality of ED and inpatient care by comparing the proportions of admission rates between the index and the return visits. Considering the relative risk by using the index visit as the reference, i.e., $RR^{(1)} = p_2^{(1)} / p_1$, the following three hypotheses were of interest: (1) $H_0 : RR^{(1)} = 1 \, vs. \, H_A : RR^{(1)} \neq 1$, (2) $H_0 : RR^{(0)} = 1 \, vs. \, H_A : RR^{(0)} \neq 1$, and (3) $H_0 : RR = 1 \, vs. \, H_A : RR \neq 1$. Two rationales are behind these comparisons to answer the clinical questions. Firstly, the admission rate to a hospital at the index visit, i.e., p_1, is treated as the general admitted probability for any person in the population.

Secondly, it is assumed that patients are fully recovered once they leave the hospital from their index visit. Therefore, the admission rate at the second visit should be similar to the first one as these two visits were assumed to be independent. Any significant differences between these two sequential admission rates might indicate certain quality issues regarding to the caring in the hospital.

In the data, there are 1,847,465 total index ED visits, with an 11.69% (n=215,906) admission rate. For patients admitted and discharged at the index visit, the return rates are 2.22% (n=4,792), and 3.42% (n=55,745) respectively, and the overall return rate is 3.28% (n=60,537). Among those who returned, the admission rates are 46.39% (n=2,223), 20.20% (n=11,263), and 22.28% (n=13,486) for patients admitted, discharged at the index visit, and pooled patients respectively. The admission at the 48-hr return visit is more likely for both patients discharged and admitted at the index visit than the admission at the index visit. To test if this difference is statistically significant, we use both of the proposed method and the two-sample proportion test. Because the cluster effect (<0.0001) in the data set is extremely small, we do not include any random effects in the likelihood function of the proposed method, as what the simulation study suggested. In Table 4, both methods provide same estimations. This echoes the simulation study and results in same conclusions that the differences in rates are statistically significant. The significant differences observed deserve further evaluation to identify a multitude of underlying quality of care issues including physician behavior, incomplete medical treatment, missed diagnoses, or failure of adequate discharge planning.

Table 4. Relative risk for admission at a 48-hr return visit compared to an index visit

		Two-proportion t-test	Proposed method[a]		
	Crude rate % (SE)	Relative risk (95% CI)	Estimated rate % (SE)	Relative risk (95% CI)	p-value
Admission rate at the index visit	11.69 (0.02)	Reference	11.69 (0.02)	reference	reference
Admission rate at the return visit (overall)	22.28 (0.17)	1.91 (1.88, 1.94)	22.28 (0.17)	1.91 (1.88, 1.94)	< 0.0001
Admission rate at the return visit for patients discharged at the index visit	20.20 (0.17)	1.73 (1.70, 1.76)	20.20 (0.17)	1.73 (1.70, 1.76)	< 0.0001
Admission rate at the return visit for patients admitted at the index visit	46.39 (0.72)	3.97 (3.85, 4.09)	46.39 (0.72)	3.97 (3.85, 4.09)	< 0.0001

[a] Approach without random effects was used because of the extremely small cluster effect in the data

2.6 DISCUSSION

This study proposes a likelihood ratio test for comparing nested proportions. Statistically, the method can accommodate a data structure with conditionality, within-subject dependence and between-cluster heterogeneity. It can be easily extended to individual patient data with covariate adjustment, other distributions of outcomes, and more than two time points of visits. When focusing on one likelihood function with covariate adjustment, the model works just like a conventional model; however, with a more generalized and flexible format to accommodate more than one distribution and a more complicated data structure. This model can be implemented in SAS PROC NLMIXED without an extra programming requirement. Compared to the naïve two-sample proportion test, it preserves the type-I error level when no difference exists, and provides less bias estimates given a large cluster effect. Generally, it performs well when the sample size is large and does require a large sample size to detect a mild to a severe difference. Non-convergence of estimates can be an issue and a model without the random effect is preferred when the cluster effect is close to 0. As to covariate adjustment, it is not clear if the covariate' effect should be fixed across all distributions in the likelihood function in terms of interpretation. Also, further study is needed to relax the assumptions of independent paths, and to include the random effect at the subject level as well as the hospital level. Clinically, the proposed method could provide more precise results and avoid false-positive findings, when the naïve method tends to claim difference easily. Moreover, it could be applied in many areas, when one distribution cannot capture the conditional structure in the dataset.

3.0 IMPROVEMENTS TO THE INTERACTION TREES ALGORITHM FOR SUBGROUP ANALYSIS IN CLINICAL TRIALS

3.1 INTRODUCTION

A core question of comparative effectiveness research (CER) is the determination of the best treatment, for whom, and under what circumstances [AHRQ, 2009]. This is not only a critical issue for a clinician, but also a methodological challenge for a biostatistician. While several methods have been proposed to provide individualized treatment rules (ITR), mostly based on the causal effect [Cai et al. 2011; Deng et al., 2012; Huang et al., 2012; Imai et al.; Rubin et al., 2012; Song et al., 2004; Zhao et al., 2011; Zhao et al., 2012], improved methods of subgroup analysis were also developed for clinical trials [Foster et al., 2011; Lipkovich et al. 2011]. The idea of a subgroup analysis is to explore heterogeneous effects with respect to a treatment. Such differential effects are common in a clinical trial, and of great interest when evaluating a treatment. For example, compared to men, women have a higher platelet aggregability after taking aspirin [Becker et al., 2006; Shen et al., 2009] and lower coronary events after cholesterol-lowering therapy [Sacks et al., 1996]. With a subgroup analysis it is possible to identify subgroups that benefit or are harmed most from the treatment and to generate hypotheses for future trials. In general, a conventional subgroup analysis requires a predetermined number of subgroups before any statistical analysis. When the underlying mechanism of the treatment and/or condition is not well understood, the selection of the subgroups can be very subjective and important subgroups may not have been prespecified. Moreover, the process introduces the problem of multiple comparisons and other issues [Cook et al., 2004; Pocock et al., 2002; Sleight, 2000; Wang et al., 2012].

To overcome this subjectivity, data-driven/machine learning methods can be very useful. One such approach is the method of interaction trees [Su et al., 2008, 2009, 2011] which is an extension of classification and regression trees (CART) [Breiman et al., 1984]. Instead of using purity measures, such as the Gini and information indexes, interaction trees split a node according to the test statistics of a treatment interaction in a model. Similar to CART, the interaction trees exhaust all variables of interest and their possible values, and then the variable with the largest test statistic is chosen as the splitting criterion with its best value. By repeating this process sequentially and performing pruning and validation at the end, it forms a tree structure that detects complex interaction with respect to the treatment automatically, and each path with splitting factors in the tree is used to distinguish subgroups. The entire procedure is applicable to any data structure when an appropriate model is performed [Su et al., 2008, 2009, 2011]. Moreover, no multiple comparisons are concerned, because the p-value is not of interest.

However, this kind of tree algorithm is always subject to instability and greediness, which are the weakness of the method. The greediness refers to the searching of a local optimal node by evaluating each predictor and its available values. This does not guarantee a global optimal tree as a better result and could be time consuming especially when the number of predictors is huge. One solution for the greediness is the evolutionary/genetic algorithm [Chatterjee et al., 1996; Goldberg, 1988; Holland, 1975]. The evolutionary algorithm obtains a globally optimal tree by evolving a forest of multiple trees through genetic mechanisms. The idea behind these mechanisms is to generate random variations which expand the searching domain for building a less greedy tree. This algorithm guarantees an improved generation of the trees at each iteration. The improvement can be quantified by various fitness evaluations, such as the misclassification rate.

In general, a genetic algorithm randomly assigns each splitting criterion. However, if a local criterion is a must, such as the test statistic in interaction trees, a connection between the randomness for the global optimum and the regional criterion for the local optimum is necessary. The method of random forests [Breiman, 2001], which comprises multiple trees, is able to link these two conflicting goals. It finds a local optimum for each node, while it allows randomness by resampling. It not only bootstraps a sample for each tree, but also randomly samples covariates with a predetermined size for each node. So far, only a few evolutionary algorithms [Gray et al., 2008; Zorman et al., 2000], have been proposed for tree analysis without a random forests algorithm, and they are not available for the interaction trees at this point in time.

The goal of this study is to build a greediness reduction interaction tree (GRIT) by integrating random forests and the evolutionary algorithm into the interaction trees algorithm. The proposed method reduces the greediness in the conventional interaction trees and preserves the tree structure for interpretation after incorporating random forests. Moreover, the analysis result helps to identify the heterogeneous treatment effects, and so that subgroups within treatment arms might be better identified in a clinical trial for a future study. The organization of this article is as follow. We describe the notation and our proposed method in Section 3.2-3.3. Then we evaluate the properties of the GRIT algorithm and compare it with the original interaction trees algorithm via simulations in Section 3.4. The strengths of the proposed method are demonstrated through a real data example from a Biological Markers for Recovery of Kidney (BioMaRK) study in Section 3.5. Finally, the discussion is presented in Section 3.6.

3.2 NOTATION

Let $O_i = (y_i, trt_i, x_i), i = 1,...,N$ be the observed data for the i^{th} subject with the outcome of interest denoted as y_i, the treatment assignment, trt_i, and $x_i = (x_{i1},...,x_{ip})$ be a $p \times 1$ vector of covariates. There are N subjects in total and we denote a tree by T. In a given tree T, $\widetilde{T}$ represents all terminal nodes with $|\widetilde{T}|$ as the number of terminal nodes, and $T - \widetilde{T}$ denotes all internal nodes with the number of internal nodes computed as $|T - \widetilde{T}|$. If h is an internal node, T_h indicates a branch of T that roots from h and includes all descendants of h in T. To describe the splitting criteria, $g_i^{(s)}$ is an indicator if a subject meets the criteria for s split. When c is a cut point for a continuous covariate $X_j, j = 1,..., p, g_i^{(s)} = I(X_j \le c)$. Otherwise, $g_i^{(s)} = I(X_j \in A)$ for a categorical variable X_j, as A is a subset of all possible categories, $C = \{c_1,...,c_k\}$.

3.3 GREEDINESS REDUCTION INTERACTION TREE (GRIT) ALGORITHM

3.3.1 Overview

The proposed method consists of interaction trees, random forests, and the evolutionary algorithm as its three main components. Each component has a sub-algorithm within itself. The outline of each part is described below and a corresponding flow chart that depicts the relationship between components is presented in Figure 3. For simplicity, we describe every detail starting from the interaction trees algorithm, and only focus on independent data with a continuous outcome.

Step 1. Random forests algorithm

Step a. Bootstrap data of sample size *N* for each tree in Step 2

Step b. Randomly select *m* variables as splitting candidates for each node in Step2

Step 2. Interaction trees algorithm

Step a. Grow a large initial tree by using the *t* test statistic of the treatment and a covariate interaction in a multiple regression

Step b. Prune each node locally by using the chi-squared distribution with df = the depth of the current tree at 5% alpha level

Step c. Use interaction complexity measure for pruning to get a subtree

Step d. Determine the best tree size via validation

Step 3. Evolutionary algorithm

Step a. Evaluate each tree by *G* statistic

Step b. Select trees based on their probabilities proportional to *G* statistics, and perform genetic mechanisms with local pruning

Step c. Repeat 3a-3b until the best tree converges.

Notice that the local pruning here is not part of the general procedure for both the interaction trees and the evolutionary algorithm. The justification and details will be given later.

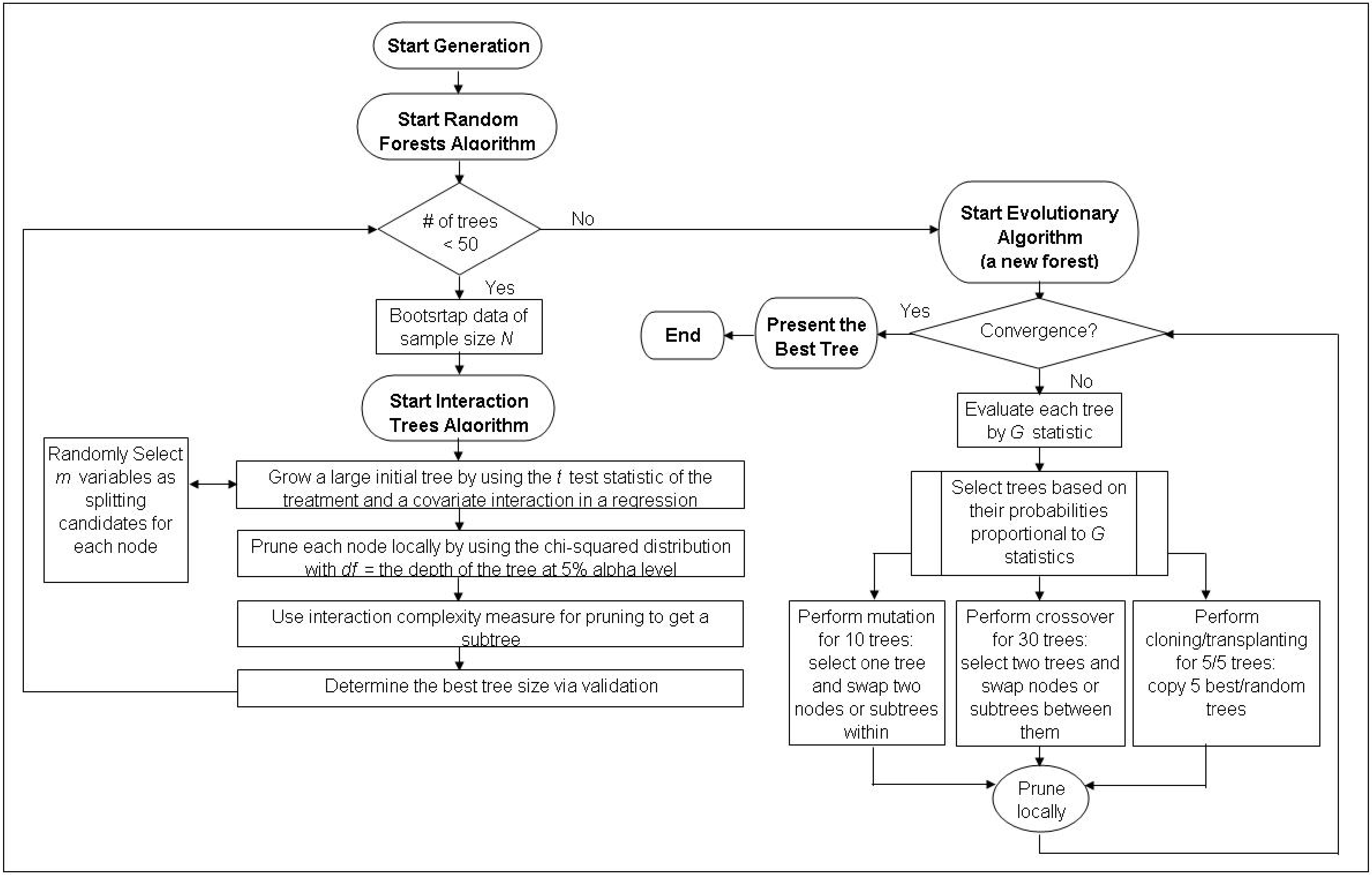
Start Generation
Start Random Forests Algorithm
# of trees < 50
No
Yes
Bootsrtap data of sample size *N*
Start Interaction Trees Algorithm
Randomly Select *m* variables as splitting candidates for each node
Grow a large initial tree by using the *t* test statistic of the treatment and a covariate interaction in a regression
Prune each node locally by using the chi-squared distribution with *df* = the depth of the tree at 5% alpha level
Use interaction complexity measure for pruning to get a subtree
Determine the best tree size via validation
Start Evolutionary Algorithm (a new forest)
Convergence?
Yes
Present the Best Tree
End
No
Evaluate each tree by *G* statistic
Select trees based on their probabilities proportional to *G* statistics
Perform mutation for 10 trees: select one tree and swap two nodes or subtrees within
Perform crossover for 30 trees: select two trees and swap nodes or subtrees between them
Perform cloning/transplanting for 5/5 trees: copy 5 best/random trees
Prune locally

Figure 3. GRIT algorithm flow chart

3.3.2 Interaction trees algorithm

Step a. Grow a large initial tree by using the t test statistic of the treatment and a covariate interaction in a multiple regression

For a single binary split, a multiple regression with the treatment effect, one covariate of interest and their interaction is fitted by using least squares.

$$\mu_i = \beta_0 + \beta_1 trt_i + \beta_2 g_i^{(s)} + \beta_3 trt_i g_i^{(s)}$$

The splitting statistic is the t^2 test statistic for $H_0 : \beta_3 = 0\, vs.\, H_A : \beta_3 \neq 0$ and is given by $G(s) = \left\{\hat{\beta}_3 / se\left(\hat{\beta}_3\right)\right\}^2$. After trying all possible candidate covariates with their available values, the best split, s^*, is the one with the highest test statistic as represented by $G(s^*) = \max_s \{G(s)\}$. The chosen predictor and its value form a splitting criterion as a parent node. Subjects who meet the criterion usually go to the left branch/child in a tree structure; otherwise, they go to the right branch/child. By repeating this process for each node until a stopping rule is met, an initial tree is built. Stopping rules could be restrictions for the maximum number of nodes, the minimum number of subjects in a node, and the maximum depth of a tree. We set the default for the minimum number of subjects in a node as 15 for a small data or 40 for a large data and the maximum depth of a tree as 10 in our method.

Step b. Prune each node locally by using the chi-squared distribution with df = the depth of the current tree at 5% alpha level

A noisy node could be formed if non-informative variables as the splitting candidates are selected from predictors in a random process, such as random forests. To remove such a noisy node, a local pruning threshold is applied after a node is found. This threshold is the critical value of a chi-squared distribution at an alpha level of 0.05 with the degrees of freedom set to be the current depth of the tree. If the splitting statistic t^2 of a node is less than the threshold, the node is removed from the tree. Otherwise, it stays in the tree. The justification of this threshold is the assumption that the splitting statistic t^2 follows a chi-squared distribution. Through the augmentation of the degrees of freedom in this chi-squared distribution, the threshold increases as the tree grows. This enforces the requirement of the need for stronger evidence for a node to stay at the bottom than at the top of the tree. It not only characterizes the need of a more informative node as the sample size decreases along the path of the tree, but also provides a simple way to possibly avoid over fitting a tree.

Step c. Use interaction complexity measure for pruning to get a subtree

To simplify a built initial tree and to avoid over fitting, we use the interaction complexity measure given by $G_\lambda(T) = G(T) - \lambda\left|T - \widetilde{T}\right|$ to prune the tree. Here, $G(T) = \sum_{h \in (T - \widetilde{T})} G(h)$, the sum of all of the test statistics in a tree, evaluates the performance of the tree. A larger value of $G(T)$ indicates a better tree. A penalty is then applied based on the tree size given a parameter $\lambda \geq 0$. Starting from the smallest to the largest value of λ, one node, h, and its subtree are identified for removal so that the tree has the largest $G_\lambda(T)$ for each λ. By doing this, nodes are withdrawn one by one until the root node is the only one left. Then the relationship between the subtrees is found as $T_M \prec ... \prec T_1 \prec T_0$, where $\prec$ denotes 'the subtree of' with the tree association as $T_{j+1} = T_j - T_h, j = 0,..,M-1$ and T_M is the root node.

Step d. Determine the best tree size via validation

The best tree, T^*, among all subtrees is determined by the tree performance in either via cross-validation or resampling samples when the sample size is small; otherwise it is determined via an independent test sample. The tree that maximizes $G_\lambda(T)$ in the former case or $\hat{G}_\lambda(T)$in the later case is the best tree. Thus $\hat{G}_\lambda(T)$ is a bias-corrected interaction complexity measure. The bias refers to the optimism of using the same data to build and validate a tree. In this case $\hat{G}_\lambda(T)$ reduces the bias through a resampling scheme. LeBlanc and Crowley [LeBlanc et al., 1993] suggested $2 \le \lambda \le 4$ and Su et al. [Su et al., 2009, 2011] found that $\lambda = \log(N)$ outperformed 2, 3, and 4 in their simulations. The recommendation for the number of bootstrap samples is $B \ge 25$ in LeBlanc and Crowley [LeBlanc et al., 1993]. We adopt $\lambda = \log(N)$ and $B = 25$ as our default. The properties of λ, and a brief description of the resampling approach are discussed in Remarks 1 and 2, respectively. More details can also be found in LeBlanc and Crowley [LeBlanc et al., 1993] and Su et al. [Su et al., 2009, 2011].

Remark 1. $\lambda = 2,4,$ and $\log(N)$ roughly correspond to AIC, $\chi_1^2(0.95)$, and BIC.

Remark 2. The description of the resampling method for the bias correction of $G_\lambda(T)$

Step i. For each λ, draw B bootstrap samples and build a tree for each sample.

Step ii. Calculate $\hat{G}_\lambda(T) = G_\lambda(T;O) - (1/B)\sum_{b=1}^{B}\{G_\lambda(T_{(b)};O) - G_\lambda(T_{(b)};O_b)\}$, where $G(tree_{(training\ data)};testing\ data)$ means applying a tree built upon a training data on a testing data, and then calculating G_λ. Notice that $T = T_{(O)}$ and O_b is the b^{th} bootstrap sample.

Step iii. Choose the tree that maximizes $\hat{G}_\lambda(T)$.

3.3.3 Random forests algorithm

To connect the interaction trees and the evolutionary algorithm, a forest of 50 interaction trees is formed. In Step a, a bootstrap sample of size *N* is drawn and used for each tree, and then in Step b, a set of *m* covariates is randomly selected as splitting candidates for each node. Usually, $m = \sqrt{\# of \;\; predictors}$ in the random forests.

3.3.4 Evolutionary algorithm

To obtain the best tree that is less greedy, we implement the following steps through iteratively evolving trees in a forest by genetic mechanisms.

Step a. Evaluate each tree by G statistic

Each tree in the forest is evaluated by the *G* statistic, $G_\lambda(T)$, for which a larger value indicates a better tree. Then the probability of a tree being selected for evolving is proportional to the $G_\lambda(T)$ statistic. This implies that a tree that describes the data better will have a larger$G_\lambda(T)$ and a higher chance for evolving and being kept in the next generation of the forest.

Step b. Select trees based on their probabilities proportional to G statistics, and perform genetic mechanisms with local pruning

The evolving process includes crossover, mutation, cloning and transplanting. For crossover, two trees are selected as parent trees based on their probabilities of being selected. Then either one node or one subtree from each of them is randomly chosen and both are swapped to form two child trees. Among each pair of parent-child, the better one is kept. By doing this, the variation between trees expands the domain of searching. In contrast, mutation tries to extend the variation within one tree by switching two random nodes or subtrees in the tree itself or by randomly assigning another splitting criterion for one node. Again, the better one among the parent-child trees stays in the new generation. Finally, cloning and transplanting are similar in keeping trees from the previous generation. The only difference between these concepts is that cloning preserves good trees and transplanting holds random trees. The former one implements the idea of elitism, and the later one increases the variation in the forest. Along these four mechanisms, the local pruning described in the interaction trees algorithm is also implemented. In general, the evolutionary algorithm appoints the proportion of trees generating from each evolving mechanism in the new generation. Here we assign 30, 10, 5, and 5 out of 50 trees in a forest built through crossover, mutation, cloning, and transplanting respectively. A visual presentation of these mechanisms is in Figure 4.

Step c. Repeat the former two steps until the best tree converges

Since a better tree has a greater chance to stay in a new generation, as we continue updating a forest by iterating the previous two steps, we obtain an improved new generation each time. When the best tree does not change a lot as comparing $G_{\lambda}(T)$ across sequential best trees, we claim that the algorithm converges, and the best tree is the final result of the analysis.

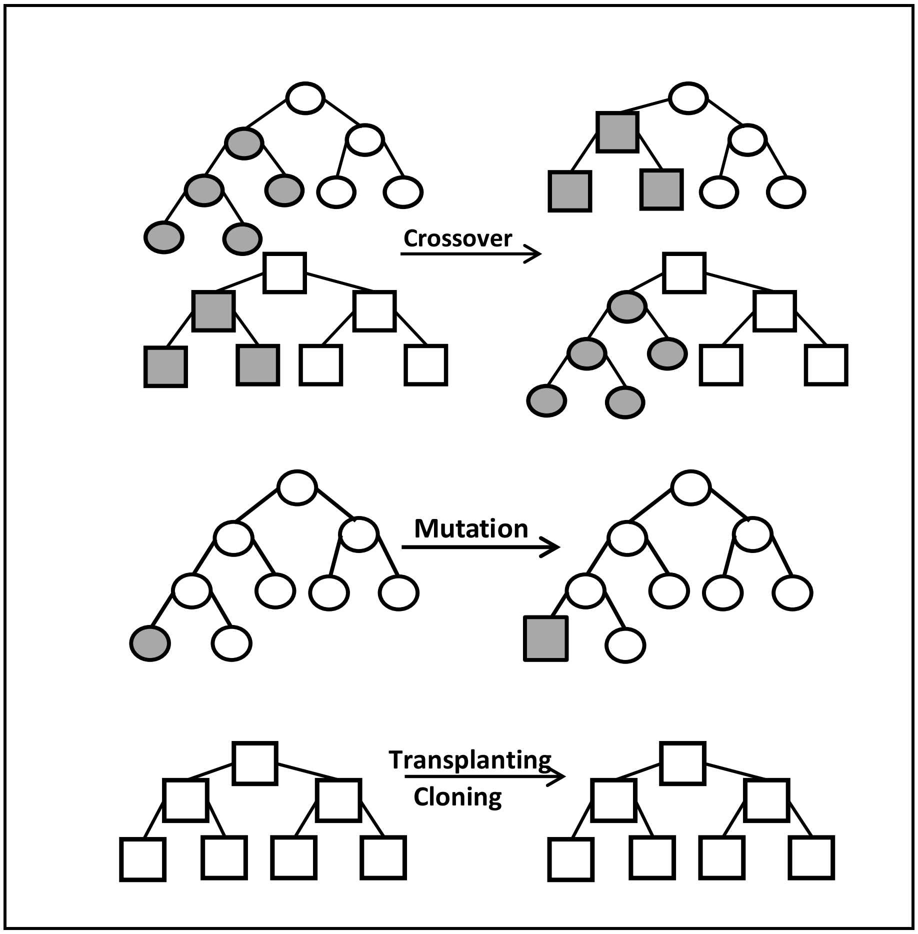

Figure 4. Genetic mechanisms

3.3.5 Miscellaneous

Test for overall interaction

To exam if a subgroup analysis is necessary, we could test $H_0 : T_M = T^* vs. H_A : T_M \neq T^*$, or equivalently we can test $H_0 : \beta_3 = 0 \; vs. H_A : \beta_3 \neq 0$, in the model $\mu_i = \beta_0 + \beta_1 trt_i + \beta_2^{'} g_i^{(T^*)} + \beta_3^{'} trt_i g_i^{(T^*)}$ by a Wald test statistic, $W = \beta_3^{'} \overset{\wedge}{\text{cov}}(\hat{\beta}_3)^{-1} \hat{\beta}_3$. Here, $g_i^{(T^*)}$ represents a vector of dummy variables for terminal nodes with dimension, $|\widetilde{T}^*| - 1$. With a large sample, $W \sim \chi^2(|\widetilde{T}^*| - 1)$. When the sample size is small, a permutation test could be used. A brief description of this permutation test is in Remark 3 and we refer it to Su et al. [Su et al., 2011] for more details.

Remark 3. The permutation test:

Step i. Grow the best tree, T^* ,based on data O with the tree size $r = |\widetilde{T}^*|$, and calculate W

Step ii. Randomize data (y_i, trt_i) and x_i to obtain permuted data O_q.

Step iii. Build a tree by using O_q with tree size r, and calculate W_q.

Step iv. Repeat Step ii-iii for Q times, i.e. $q = 1,...,Q$.

Step v. Calculate the empirical *p*-value as $\left\{1 + \sum_{q=1}^{Q} I(W_q \geq W)\right\} / (Q+1)$.

Amalgamation

When several terminal nodes share a similar treatment effect, it is reasonable to merge them as one group, although they might follow different causal paths. This amalgamation can be accomplished within two steps. First, iteratively merge a pair of terminal nodes that have the smallest $t = \hat{\beta}_3 / se(\hat{\beta}_3)$, until all subgroups have quite different treatment effects from each other. Second, for a clear display, final subgroups should be sorted by the magnitude of the treatment effect.

Software

We develop the whole algorithm in R based on an existing program for interaction trees [Su et al., 2011] and plan to improve its efficiency by cooperating it with the C++ codes in a R package 'etree' [Grubinger et al., 2011]. The integrated codes will be efficient with respect to the computational time and could be easily implemented in the interface of R.

3.4 SIMULATION STUDY

We evaluate the performance of the proposed method in comparison with the interaction trees in two aspects. The first one is the tree size and the second one is hits. They are the percentages of correctly identifying the number of terminal nodes and predictors respectively. Moreover, the average and the standard deviation of $G_\lambda(T)$ are presented as the reference. We generated four covariates, $x_1, x_2, x_3,$ and x_4, from a discrete uniform distribution containing 4 possible values: 0, 0.25, 0.5, and 1. Then seven different models in Table 5 were studied. Among them, Model A

Table 5. Simulation setting for the subgroup analysis by interaction trees

Items	Values	
	GRIT	ITs
Sample size: N	150 / 500	
Number of sampled covariates in random forests: m	1 / 2 / 4	
	2	4
Model[a,b]…{Tree size}	$A: Y = 2 + 2 \cdot trt + 2x_1 + 2x_2 + \varepsilon, \varepsilon \sim N(0,1)$ .... {1} $B: Y = 2 + 2 \cdot trt + 2x_1 + 2x_2 + 2z_1 z_2 \cdot trt + \varepsilon, \varepsilon \sim N(0,1)$ .... {3} $C: Y = 2 + 2 \cdot trt + 2x_1 + 2x_2 + 2z_1 \cdot trt + 2z_2 \cdot trt + \varepsilon, \varepsilon \sim N(0,1)$ .... {4} $D: Y = 2 + 2 \cdot trt + 2x_1 + 2x_2 + 4z_1 \cdot trt + 4z_2 \cdot trt + \varepsilon, \varepsilon \sim N(0,1)$ .... {4} $E: Y = 10 + 10 \cdot trt + \exp\left\{(x_1 - 0.5)^2 + (x_2 - 0.5)^2\right\} + \varepsilon, \varepsilon \sim N(0,1)$ $F: Y = 2 + 2 \cdot trt + 2x_1 + 2x_2 + 2z_1 \cdot trt + 2z_2 \cdot trt + \varepsilon, \varepsilon \sim Uniform\left(-\sqrt{3}, \sqrt{3}\right)$ ...... {4} $G: Y = 2 + 2 \cdot trt + 2x_1 + 2x_2 + 2z_1 \cdot trt + 2z_2 \cdot trt + \varepsilon, \varepsilon \sim Exp(1)$ .... {4}	

[a] $x_1, x_2, x_3, x_4 = \{0, 0.25, 0.5, 1\}$, following a discrete uniform distribution

[b] $z_1 = I(x_1 \le 0.5)$ and $z_2 = I(x_2 \le 0.5)$

had no treatment interactions, and thus is used to evaluate the type I error of the methods. Model B included only one three-way interaction, while Model C and D contained two two-way interactions which had stronger signals in Model D than C. Model E provided a non-linear interaction form of one exponential and two power functions. Model F and G were generated to test the robustness of the methods by generating the error term from a uniform distribution with the range from $-\sqrt{3}$ to $\sqrt{3}$ and from an exponential distribution with rate 1 respectively. In the models, $z_1 = I(x_1 \le 0.5)$ and $z_2 = I(x_2 \le 0.5)$, and only x_1 and x_2 were the true predictors interacting with the treatment. The true tree sizes for these models, with the exception of Model E, in alphabetical order were 1, 3, 4, 4, 4, and 4.

Other than the scenarios presented here, various scenarios were also considered. The sample size was either 150 or 500 to represent a small or a large data set. The number of sampled covariates for splitting in the random forests had 4 levels: 1, 2, 4 for both methods, and 2 for the proposed method and 4 for the interaction trees. The last level implies the case where each method uses its default setting for individual's best performance. Each scenario had 200 simulated datasets. All of these settings were similar to those presented in Sue et al, 2009 and 2011 [Su et al., 2009, 2011]. Details are found in Table 5.

Table 6 and Table 7 present the simulation results for *N*=500 and *N*=150 respectively. In terms of the tree size and the hits, the interaction tree outperforms the proposed method when it is allowed to check every covariate for splitting, i.e., *m*=4, with a large sample size. In this case, the interaction tree correctly identifies the tree size in 73% to 100% of the simulated datasets and detects the true predictors in more than 96% of the simulated datasets. Whereas the proposed method has only 74% to 86% probabilities to claim the true size and 81.5% to 94% chances to find the correct predictors. Moreover, when no interaction exits, the interaction tree is superior

Table 6. Simulation results comparing the proposed method to the interaction trees under different scenarios with N=500. (True tree size in bold.)

	GRIT									ITs								
	Tree Size (%)							Hits	$G_\lambda(T)$	Tree Size (%)							Hits	$G_\lambda(T)$
	1	2	3	4	5	6	≥7	(%)	Mean±SD	1	2	3	4	5	6	≥7	(%)	Mean±SD
Model A: $Y = 2 + 2 \cdot trt + 2x_1 + 2x_2 + \varepsilon, \varepsilon \sim N(0,1)$																		
m=1	**81.5**	15.0	3.0	0.5	0.0	0.0	0.0	81.5	0.31±1.20	**98.0**	1.5	0.5	0.0	0.0	0.0	0.0	98.0	0.09±0.63
m=2	**80.5**	12.5	5.5	1.0	0.5	0.0	0.0	80.5	0.36±1.21	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.00±0.00
m=4	**83.0**	11.5	4.5	1.0	0.0	0.0	0.0	83.0	0.33±1.22	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.00±0.00
m=2/4	**79.5**	17.0	3.0	0.5	0.0	0.0	0.0	79.5	0.22±0.74	**100.0**	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.00±0.00
Model B: $Y = 2 + 2 \cdot trt + 2x_1 + 2x_2 + 2z_1 z_2 \cdot trt + \varepsilon, \varepsilon \sim N(0,1)$																		
m=1	0.0	0.0	**75.0**	19.5	5.0	0.5	0.0	85.0	70.51±18.90	29.0	16.5	**15.5**	17.0	10.5	5.5	6.0	18.0	27.59±26.26
m=2	0.0	0.0	**78.0**	15.5	6.0	0.5	0.0	86.5	73.00±18.24	5.5	3.0	**44.0**	15.0	18.5	7.5	6.5	49.5	54.19±25.06
m=4	0.0	0.0	**74.0**	21.5	4.5	0.0	0.0	86.5	72.03±19.52	0.0	0.0	**96.0**	2.5	1.0	0.0	0.5	98.0	68.28±19.18
m=2/4	0.0	0.0	**80.0**	17.5	2.5	0.0	0.0	89.0	73.15±17.71	0.0	0.0	**98.0**	2.0	0.0	0.0	0.0	99.0	68.68±17.08
Model C: $Y = 2 + 2 \cdot trt + 2x_1 + 2x_2 + 2z_1 \cdot trt + 2z_2 \cdot trt + \varepsilon, \varepsilon \sim N(0,1)$																		
m=1	0.0	0.0	0.0	83.0	14.5	2.0	0.5	89.5	185.61±28.59	4.0	12.5	10.5	12.5	13.0	14.5	33.0	12.5	96.33±45.94
m=2	0.0	0.0	0.0	77.0	18.5	4.5	0.0	86.5	188.14±30.04	0.0	2.0	5.0	25.5	16.5	14.5	36.5	36.0	155.64±39.96
m=4	0.0	0.0	0.0	83.0	13.0	4.0	0.0	88.5	187.72±27.75	0.0	0.0	0.0	98.0	1.0	1.0	0.0	98.5	185.79±27.44
m=2/4	0.0	0.0	0.0	75.0	20.5	4.5	0.0	85.0	188.57±30.57	0.0	0.0	0.5	97.0	2.0	0.5	0.0	97.5	186.21±30.07
Model D: $Y = 2 + 2 \cdot trt + 2x_1 + 2x_2 + 4z_1 \cdot trt + 4z_2 \cdot trt + \varepsilon, \varepsilon \sim N(0,1)$																		
m=1	0.0	0.0	0.0	**81.5**	12.0	4.5	2.0	88.0	644.98±59.54	2.5	3.5	5.0	**11.0**	15.0	18.0	45.0	12.0	349.24±156.15
m=2	0.0	0.0	0.0	**82.5**	14.0	3.5	0.0	85.0	644.34±55.68	0.0	0.0	0.0	**19.0**	13.0	17.5	50.5	23.5	547.53±99.98
m=4	0.0	0.0	0.0	**86.0**	13.0	1.0	0.0	91.0	648.13±56.19	0.0	0.0	0.0	**98.5**	1.5	0.0	0.0	99.0	632.70±54.98
m=2/4	0.0	0.0	0.0	**81.5**	15.5	3.0	0.0	88.0	653.26±60.73	0.0	0.0	0.0	**98.5**	0.5	1.0	0.0	98.5	635.51±57.11
Model E: $Y = 10 + 10 \cdot trt + \exp\{(x_1 - 0.5)^2 + (x_2 - 0.5)^2\} + \varepsilon, \varepsilon \sim N(0,1)$																		
m=1	0.0	0.0	1.0	**65.0**	21.0	8.0	5.0	91.5	236.14±39.84	4.0	10.5	14.5	**16.5**	19.5	13.5	21.5	22.5	116.94±60.43
m=2	0.0	0.0	2.5	**59.5**	23.0	13.0	2.0	91.0	237.84±38.86	0.0	0.5	17.0	**26.5**	19.0	13.5	23.5	49.5	187.86±48.66
m=4	0.0	0.0	1.0	**54.5**	30.5	12.0	2.0	94.0	235.01±37.85	0.0	0.0	9.0	**73.0**	5.0	9.0	4.0	96.0	227.79±38.03
m=2/4	0.0	0.0	0.0	**65.0**	24.0	7.5	3.5	93.5	238.18±36.78	0.0	0.0	4.0	**81.0**	6.5	4.5	4.0	96.5	229.86±35.69
Model F: $Y = 2 + 2 \cdot trt + 2x_1 + 2x_2 + 2z_1 \cdot trt + 2z_2 \cdot trt + \varepsilon, \varepsilon \sim Uniform\left(-\sqrt{3}, \sqrt{3}\right)$																		
m=1	0.0	0.0	0.0	**81.0**	15.0	2.5	1.5	86.0	189.87±30.50	4.0	8.5	11.5	**12.5**	13.5	13.0	37.0	17.0	103.39±48.85
m=2	0.0	0.0	0.0	**83.0**	14.0	2.5	0.5	92.0	190.76±30.16	0.5	1.0	3.0	**21.0**	16.5	16.0	42.0	27.0	157.80±40.69
m=4	0.0	0.0	0.0	**80.5**	17.5	2.0	0.0	88.5	184.76±27.59	0.0	0.0	0.0	**98.0**	1.5	0.5	0.0	98.0	183.21±27.51
m=2/4	0.0	0.0	0.0	**74.0**	21.5	3.5	1.0	81.5	190.72±28.47	0.0	0.0	0.5	**97.5**	1.0	0.5	0.5	98.0	188.49±28.48
Model G: $Y = 2 + 2 \cdot trt + 2x_1 + 2x_2 + 2z_1 \cdot trt + 2z_2 \cdot trt + \varepsilon, \varepsilon \sim Exp(1)$																		
m=1	0.0	0.0	0.0	**82.0**	15.0	2.5	0.5	90.0	191.93±33.45	3.0	8.0	8.5	**8.0**	16.5	15.5	40.5	13.0	103.19±45.94
m=2	0.0	0.0	0.0	**82.5**	13.5	4.0	0.0	86.0	190.00±34.15	0.0	0.5	8.0	**17.5**	16.0	16.5	41.5	28.5	157.36±41.65
m=4	0.0	0.0	0.0	**81.5**	16.5	1.5	0.5	88.5	193.37±35.57	0.0	0.0	0.0	**100.0**	0.0	0.0	0.0	100.0	190.57±35.32
m=2/4	0.0	0.0	0.0	**82.5**	14.5	1.5	1.5	88.0	189.26±34.83	0.0	0.0	0.5	**98.5**	0.5	0.5	0.0	99.0	186.49±34.48

Table 7. Simulation results comparing the proposed method to the interaction trees under different scenarios with N=150. (True tree size in bold.)

	GRIT									ITs								
	Tree Size (%)							Hits	$G_\lambda(T)$	Tree Size (%)							Hits	$G_\lambda(T)$
	1	2	3	4	5	6	≥7	(%)	Mean±SD	1	2	3	4	5	6	≥7	(%)	Mean±SD
Model A: $Y = 2 + 2\cdot trt + 2x_1 + 2x_2 + \varepsilon, \varepsilon \sim N(0,1)$																		
m=1	**72.5**	22.0	5.5	0.0	0.0	0.0	0.0	72.5	0.67±1.74	**94.0**	5.5	0.5	0.0	0.0	0.0	0.0	94.0	0.18±1.02
m=2	**72.0**	24.5	3.5	0.0	0.0	0.0	0.0	72.0	0.48±1.30	**90.0**	7.0	2.5	0.5	0.0	0.0	0.0	90.0	0.26±0.95
m=4	**68.5**	25.0	6.5	0.0	0.0	0.0	0.0	68.5	0.60±1.52	**84.0**	7.0	7.0	2.0	0.0	0.0	0.0	84.0	0.58±1.45
m=2/4	**68.0**	28.0	4.0	0.0	0.0	0.0	0.0	68.0	0.53±1.38	**81.0**	13.0	5.5	0.5	0.0	0.0	0.0	81.0	0.53±1.41
Model B: $Y = 2 + 2\cdot trt + 2x_1 + 2x_2 + 2z_1z_2\cdot trt + \varepsilon, \varepsilon \sim N(0,1)$																		
m=1	6.0	8.5	**78.5**	7.0	0.0	0.0	0.0	81.5	16.97±11.77	55.5	22.0	**17.5**	4.5	0.5	0.0	0.0	13.0	4.83±8.45
m=2	4.0	11.5	**74.5**	10.0	0.0	0.0	0.0	74.5	16.05±10.05	21.5	33.5	**40.0**	4.0	1.0	0.0	0.0	34.0	8.28±8.90
m=4	5.5	8.5	**79.0**	7.0	0.0	0.0	0.0	80.0	16.85±10.79	4.5	9.5	**71.5**	14.0	0.5	0.0	0.0	77.5	14.94±9.95
m=2/4	4.5	10.0	**77.5**	8.0	0.0	0.0	0.0	80.5	16.85±10.94	9.5	4.5	**72.5**	13.0	0.5	0.0	0.0	74.0	15.47±10.54
Model C: $Y = 2 + 2\cdot trt + 2x_1 + 2x_2 + 2z_1\cdot trt + 2z_2\cdot trt + \varepsilon, \varepsilon \sim N(0,1)$																		
m=1	0.0	1.0	18.5	**80.0**	0.5	0.0	0.0	96.5	49.98±16.89	23.0	28.5	27.0	**19.5**	2.0	0.0	0.0	17.5	19.05±17.05
m=2	0.0	1.0	14.5	**84.5**	0.0	0.0	0.0	98.5	48.49±16.17	1.5	18.0	44.0	**33.0**	3.5	0.0	0.0	56.5	33.26±16.57
m=4	0.0	0.5	18.5	**79.5**	1.5	0.0	0.0	94.5	50.74±14.76	0.0	0.5	16.5	**80.5**	2.5	0.0	0.0	95.0	49.62±14.50
m=2/4	0.0	0.0	12.5	**85.5**	2.0	0.0	0.0	97.0	49.76±15.91	0.0	0.0	13.5	**84.5**	2.0	0.0	0.0	94.5	48.51±15.60
Model D: $Y = 2 + 2\cdot trt + 2x_1 + 2x_2 + 4z_1\cdot trt + 4z_2\cdot trt + \varepsilon, \varepsilon \sim N(0,1)$																		
m=1	0.0	0.0	3.5	**96.5**	0.0	0.0	0.0	99.5	189.10±36.45	13.0	26.0	34.0	**24.5**	2.0	0.5	0.0	14.5	57.75±51.62
m=2	0.0	0.0	2.0	**97.5**	0.5	0.0	0.0	99.5	185.26±33.33	0.0	18.0	42.5	**31.0**	8.5	0.0	0.0	54.0	112.75±49.90
m=4	0.0	0.0	2.0	**98.0**	0.0	0.0	0.0	100.0	187.63±32.71	0.0	0.0	2.5	**96.0**	1.5	0.0	0.0	99.5	179.26±31.07
m=2/4	0.0	0.0	4.0	**96.0**	0.0	0.0	0.0	100.0	188.05±35.09	0.0	0.0	3.5	**96.0**	0.5	0.0	0.0	100.0	178.00±32.71
Model E: $Y = 10 + 10\cdot trt + \exp\left\{(x_1 - 0.5)^2 + (x_2 - 0.5)^2\right\} + \varepsilon, \varepsilon \sim N(0,1)$																		
m=1	0.0	1.0	88.0	10.0	1.0	0.0	0.0	95.5	52.41±16.52	21.0	30.0	30.0	17.5	1.5	0.0	0.0	16.0	20.74±18.97
m=2	0.0	2.0	88.0	10.0	0.0	0.0	0.0	95.5	51.56±14.83	3.5	25.5	43.0	24.5	3.5	0.0	0.0	44.0	35.20±20.24
m=4	0.0	2.0	84.5	13.0	0.5	0.0	0.0	95.0	54.60±17.27	0.0	3.5	66.5	22.5	7.0	0.5	0.0	84.5	51.89±17.03
m=2/4	0.0	1.5	86.0	12.5	0.0	0.0	0.0	96.0	52.17±16.07	0.0	4.0	67.5	25.5	3.0	0.0	0.0	86.5	48.34±16.30
Model F: $Y = 2 + 2\cdot trt + 2x_1 + 2x_2 + 2z_1\cdot trt + 2z_2\cdot trt + \varepsilon, \varepsilon \sim Uniform\left(-\sqrt{3}, \sqrt{3}\right)$																		
m=1	0.0	1.0	18.5	**80.0**	0.5	0.0	0.0	96.0	47.90±15.59	22.0	25.0	32.5	**19.0**	1.5	0.0	0.0	17.0	17.69±16.31
m=2	0.0	1.0	14.5	**84.5**	0.0	0.0	0.0	97.0	50.79±15.53	0.5	22.5	36.0	**38.5**	2.5	0.0	0.0	55.5	36.56±16.38
m=4	0.0	0.0	20.0	**79.0**	1.0	0.0	0.0	96.0	50.18±15.44	0.0	0.5	16.5	**82.0**	1.0	0.0	0.0	96.5	49.34±15.20
m=2/4	0.0	0.0	17.5	**81.0**	1.5	0.0	0.0	96.5	48.22±16.05	0.0	1.0	16.0	**80.0**	3.0	0.0	0.0	94.5	46.8±15.6
Model G: $Y = 2 + 2\cdot trt + 2x_1 + 2x_2 + 2z_1\cdot trt + 2z_2\cdot trt + \varepsilon, \varepsilon \sim Exp(1)$																		
m=1	0.0	1.0	27.5	**70.5**	1.0	0.0	0.0	96.0	52.26±20.27	18.0	22.5	41.0	**17.5**	1.0	0.0	0.0	18.0	19.82±18.57
m=2	0.0	1.5	26.0	**72.0**	0.5	0.0	0.0	92.0	48.62±18.04	1.0	19.0	44.5	**31.0**	4.5	0.0	0.0	56.5	33.22±17.68
m=4	0.0	0.5	19.5	**79.0**	1.0	0.0	0.0	95.0	52.41±19.92	0.0	1.5	16.0	**80.0**	2.5	0.0	0.0	92.0	50.47±19.40
m=2/4	0.0	0.0	15.5	**83.0**	1.5	0.0	0.0	97.5	56.30±19.74	0.0	0.0	15.5	**83.0**	1.5	0.0	0.0	97.0	54.16±19.08

with a more than 98% and 81% probability of uncovering the tree without splitting respectively under the large and small sample sizes. In contrast, the proposed method tends to declare at least one false-positive interaction with a more than 20.5% probability for a large data and a more than 32% chance for a small data.

However, when the interaction exits and only partial covariates are sampled as splitting candidates, i.e., $m < 4$, the proposed method shows 1.77 (78% vs. 44%) to 10.25 (82% vs. 8%) times the power of the interaction trees to discover the tree size with a large sample. In addition, the correct covariates are identified in 85 to 92% of the simulated datasets by the proposed method and only 12.5% to 49.5% by the interaction trees. When the sample size is small and the interaction exists, the proposed method even competes or is better with the interaction trees as m=4 and m=2/4. There, the chances of finding the true tree size and the predictors are 71.5% to 98.0% and 74% to 100% respectively for both methods and the proposed method is better for both Model B and E. As m declines with a small sample size, the proposed method shows its strength when compared to the interaction trees as is the case in a large sample setting. In this case, the proposed method has a 1.86 (74.5% vs. 40%) to 4.49 (78.5% vs. 17.5%) fold chance to identify the correct tree size when compared to the interaction tree approach. Also, the true predictors are identified in 74.5% to 99.5% of the simulated datasets by the proposed method but only in 13% to 56.5% of the datasets by the interaction trees. When considering different models, the proposed method performs especially well for Model E which includes non-linear interactions. Also, the pattern of the results is similar in Model B, C and D, and is alike in Model F and G.

In addition, the proposed method is more robust than the interaction trees by providing tighter distributions of the claimed tree sizes and smaller variations of the $G_\lambda(T)$ statistics. As to the tree size, when the sample size is small the proposed method finds 3 to 4 different sizes, while the interaction trees declare 4 to 6 distinct ones. The phenomenon gets worse for the interaction trees when the sample size is large, where the tree sizes spread from 3 to ≥7. For the $G_\lambda(T)$ statistics, trees built by the proposed method have a higher average of $G_\lambda(T)$ statistics than the interaction trees as expected, since the proposed method determines the best tree by a large $G_\lambda(T)$ statistic. Also, the variation of the $G_\lambda(T)$ statistics from the proposed method is smaller than the one from the interaction trees, when only partial covariates are selected as splitting candidates, i.e., m=1 or 2.

3.5 EXAMPLE

The illustrative example is the Biological Markers for Recovery of Kidney (BioMaRK) Study. It is a nested prospective observational cohort study as an ancillary study to the Veterans Affairs (VA)/NIH Acute Renal Failure Trial Network (ATN) study [Ronco et al., 2008], which is a multicenter prospective trial. The goal of the VA/NIH study (n=1124) was to investigate if a high intensive renal-replacement therapy (RRT) (i.e., dialysis at a higher dose) decreases mortality among critically ill patients with acute kidney injury (AKI) more than a less-intensive RRT, and no overall significant differences were found. Following the same idea, the aim of the BioMaRK study was to understand the role of 11 plasma biomarkers concentrations in determining the relationship between the treatment and the survival outcome of interest at day 60. Therefore, the BioMaRK study only included participants who were in the VA/NIH ATN study and gave additional blood sample for banking. There, the biomarkers of interest are plasma interleukin-6, IL-8, IL-10, IL-18, IL-1β, macrophage migration inhibitory factor (MIF) and tumor necrosis factor (TNF) for inflammation, and tumor necrosis factor (TNFRI)-I, TNFR-II, and death receptor (DR)-5 for apoptosis, and granulocyte macrophage colony stimulating factor (GM-CSF) for as a growth factor.

One specific question in the BioMaRK study was whether any particular groups benefit or are harmed most from the intensive RRT. This is exactly what the proposed GRIT algorithm was developed for in a randomized trial. To reframe the question as the preferred setting of the proposed method, we only include subjects who had no missing values and died in 60 days. For the former issue, it is because the proposed method must be applied to complete data. For the later issue, it is because the proposed method works best for outcomes that are normally distributed. Therefore, only the 156 participants who died in 60 days with complete data in the BioMaRK study are used in this analysis. The outcome of interest is the survival days in 60 days in a log scale as a normal transformation and the primary covariates are the 11 plasma biomarkers concentrations at day 1 before RRT. We use both of the interaction trees and the proposed method to analyze this data and m is 4 for the proposed method and is 11 for the interaction trees. Notice that this example is only used as an illustration of the proposed method and the data used here is the version before the missing issue of the biomarker data is fixed.

Table 8 and Figure 5 are respectively the subgroup summary and the tree built based on the interaction trees. Four splitting are found including log(TNFRI) with the cut point 9.35, log(IL18) with 4.19, log(TNFRII) with 8.93, and log(MIF) with 6.23. Originally, five terminal nodes are identified but nodes 111, 1211, and 122 are merged as one subgroup because the intervention effect is not effective to these groups. The final results show that 11 patients receiving the intensive RRT in subgroup I, i.e., who had log(TNFRI) less or equal to 9.35 and log(IL18) larger than 4.19, have the average survival day as 13.82, whereas 17 subjects receiving

Table 8. Summary of the subgroup analysis by using ITs for BioMaRK study (n=156)

	Intensive RRT		Less- Intensive RRT		Intervention effect	
	Size	Mean survival day	Size	Mean survival day	*t*-test	p-value
Terminal nodes						
111	14	23.07	15	22.33	0.16	0.8751
112	11	13.82	17	28.76	-3.91	0.0006
1211	17	21.29	23	19.52	0.49	0.6266
1212	14	29.14	10	12.60	3.85	0.0017
122	14	25.00	21	34.95	-1.96	0.0590
Subgroups						
I (112)	11	13.82	17	28.76	-3.91	0.0006
II (111+1211+122)	45	23.00	59	25.73	-1.00	0.3205
III (1212)	14	29.14	10	12.60	3.85	0.0017

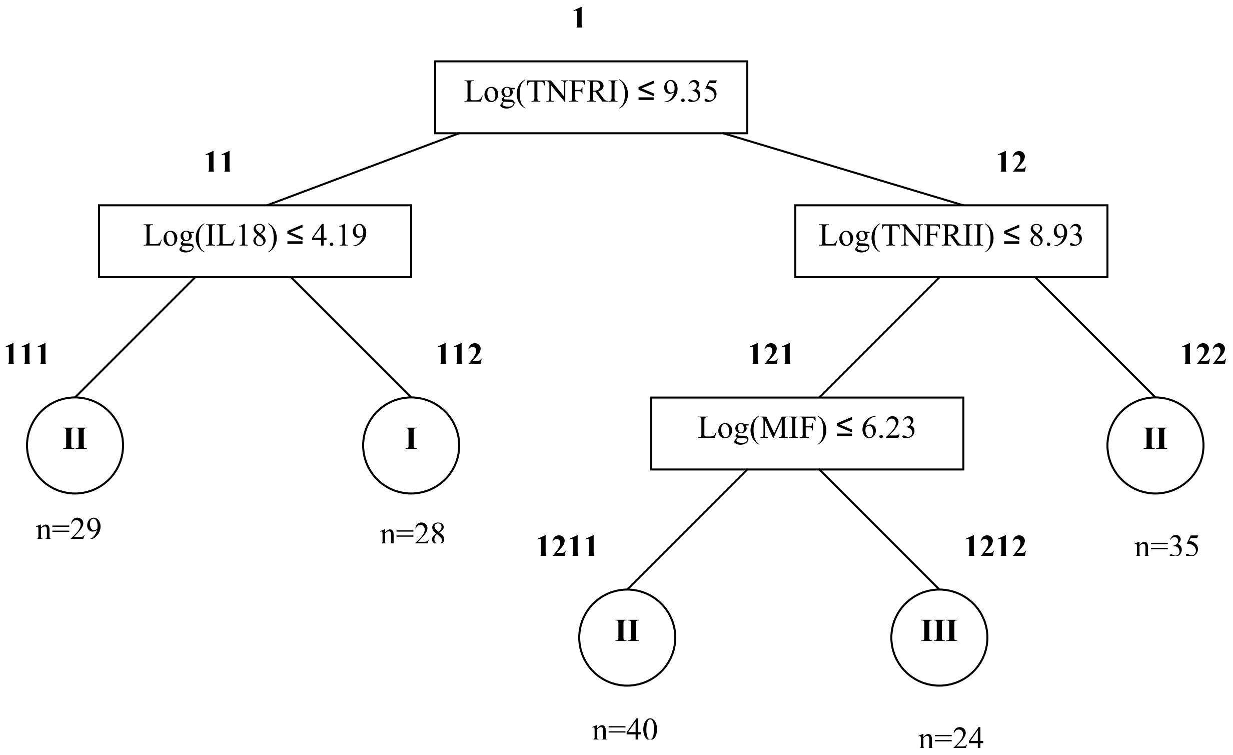

Figure 5. Tree structure of the subgroup analysis by using ITs for BioMaRK study

the less-intensive RRT have the average survival day as 28.76. The significant p-value (p-value = 0.006) and negative *t* statistic (*t* statistic-3.91) indicates a possible harmful effect of the intensive RRT on this subgroup. In contrast, subgroup III experienced a beneficial treatment effect, where the average survival day of 14 patients in the intensive RRT group is 29.14 and the average survival day of 10 patients in the less-intensive RRT group is 12.60. For this significant intervention effect, the *t* statistic is 3.85 and the p-value is 0.0017. The overall interaction test has the Wald statistic around 6.51 and the p-value<0.0001.

However, the proposed GRIT algorithm merely recognizes two subgroups by one splitting which is log(TNFRI) with the cut point 9.35. Only patients who had log(TNFRI) less or equal to 9.35 experienced the harmful effect from the intensive RRT when compared to the less-intensive treatment. There, the average survival day of 25 participants in the intensive RRT group is 19 and for 32 participants in the less-intensive RRT group the average survival day is 25.75. Moreover, the *t* test is not highly statistical significant (*t* statistic = -2.06; p-value = 0.0454). The overall interaction test has the Wald statistic around 6.41 and the p-value about 0.0124. These results are presented in Table 9 and Figure 6. Although the IT and GRIT give different results, the results obtained from the GRIT analysis might be preferred. This is the case due to the simulation study results showing that the interaction trees approach has the power to identify no interaction, thus when IT claims one in this analysis, it indicates that the interaction is likely to be present. Moreover, the proposed method outperforms or competes with the ITs when the sample size is small as 150 and an interaction is present. Therefore, the additional splitting from the ITs might be due to its instability and the more conservative result from the proposed method might be more reasonable.

Table 9. Summary of the subgroup analysis by using GRIT for BioMaRK study (n=156)

	Intensive RRT		Less- Intensive RRT		Intervention effect	
	Size	Mean survival day	Size	Mean survival day	*t*-test	p-value
Terminal nodes						
11	25	19.00	32	25.75	-2.06	0.0454
12	45	24.89	54	24.24	0.22	0.8265

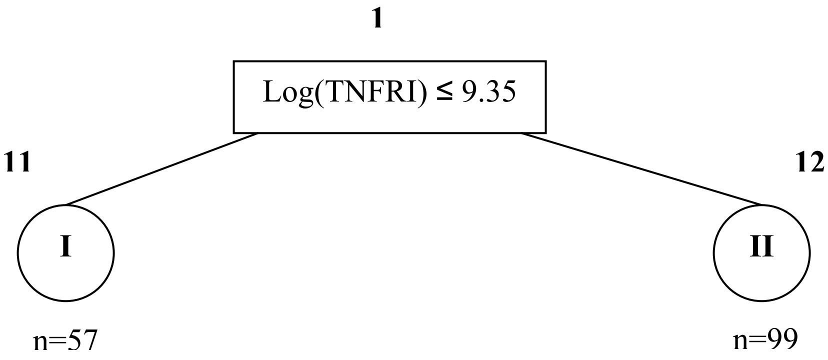

Figure 6. Tree structure of the subgroup analysis by using GRIT for BioMaRK study

3.6 DISCUSSION

In this study, we propose a greediness reduction interaction tree (GRIT) algorithm to identify subgroups with respect to the treatment in clinical trials. Statistically, the proposed method reduces the greediness in the interaction trees by increasing the variation when searching for a node. To be specific, it incorporates the evolutionary algorithm into the interaction trees through random forests. Similar to the interaction trees approach, the proposed method does not suffer from the issues of the subjectivity and multiple comparisons as in a conventional subgroup analysis, because it is a data-driven process and the p-value is not of interest. It is also able to reduce the greediness and to preserve the interpretation of a single tree even after cooperating with the interaction trees and random forests. It can be applied to other types of outcomes by using appropriate models and test statistics. Moreover, outliers in the data do not affect the analysis directly, because only the maximum, i.e., the rank, of the *t* statistic matters. Compared to the interaction trees, the proposed method performs well when the interaction exits and only partial covariates are sampled as splitting candidates. When the sample size is small, it can compete with the interaction trees even when using each one's preferred setting when the interaction exits. It is especially able to detect non-linear interaction regardless of the sample size. Also, it is more robust than the interaction trees in terms of the possible tree sizes and the variation of the $G_{\lambda}(T)$ statistic as partial covariates are sampled. However, it has a high false-positive rate when no interaction exits and it is computationally intensive, especially for a large dataset. As to the local pruning in the proposed method, alternatives can be considered in addition to the threshold of a chi-squared distribution. One empirical option is to choose the threshold as a certain percentage of the previous threshold. This might reduce the false-positive declaration in the proposed method, since the first cut point is always $\chi_1^2(0.05) \approx 3.84$ and it may not be appropriate for every data set. Also, further study is needed to add a validation process for the final tree after considering the computational burden, and to provide a measure of variable importance after weighting the contribution of each variable in the whole algorithm. Clinically, a more objective and precise tool for the subgroup analysis could advance the progress of the personalized medicine to target possible beneficial or harmful subgroups to an intervention.

4.0 CONCLUSION

4.1 SUMMARY

In this dissertation, two statistical issues of importance in the area of comparative effectiveness research were considered. They are essentially different issues within this framework. While the first method presented focuses on hypothesis testing in observational studies by using parametric methods, the second method focuses on subgroup analysis in clinical trials by using machine learning techniques.

In the first part of this dissertation, i.e., Chapter 2, we proposed a likelihood ratio test for comparing two nested proportions based on the product of conditional probabilities. To cover the statistical issues in the example of the ED data used, our test not only considers the dependence within subjects and the cluster effect between hospitals, but also the conditional structure. Simulations showed that the method provides unbiased estimates and a reasonable power especially when the sample size is large and no difference exists. The proposed method can be performed directly by using SAS PROC NLMIXED.

In the second part of this dissertation, i.e., Chapter 3, the greediness reduction interaction tree (GRIT) algorithm was proposed to reduce the greediness in the interaction trees for subgroup analysis in clinical trials. The proposed method integrates the evolutionary algorithm into the interaction trees through random forests. Simulations showed that this proposed method outperforms the interaction trees without accessing every predictor when the interaction exits and it is especially useful when the sample size is small. Also, the GRIT is relatively robust to the interaction trees. We demonstrated it in a real data example from the Biological Markers for Recovery of Kidney (BioMaRK) study. A set of R code was developed to implement this method.

4.2 FUTURE WORK

For the first part of this dissertation, although the framework seems to be restricted to the example provided, the methodology may be extended to various fields faced with similar analysis issues. In the long term, due to the conditional structure, future work may relate the proposed likelihood function to a more efficient study design such as adaptive treatment strategies [Murphy, 2003]. In the short term, the proposed method may be extended to individual patient data with covariate adjustment after clarifying the interpretation across different distributions, and to including the random effect at the subject level.

For the second part of this dissertation, although conceptually the proposed method can be extended to other types of outcomes, the literature [Su et al., 2008, 2009, 2011] showed that the performance differs across distinct outcomes. In the long term, additional work is necessary to tune the methodology accordingly for each type of outcome. In the short term, more efforts are needed to understand the effect of the local pruning on the performance of the proposed method to reduce the false-positive rate. Alternatives of local pruning such as empirical thresholds, the validation of GRIT, or the mixture of the interaction trees and the GRIT especially for choosing the root node may be considered as well. Moreover, variable importance and weighting in GRIT algorithm may be emphasized in the future work due to their clinical usefulness.

APPENDIX A

SAMPLE CODE OF LIKELIHOOD RATIO TEST FOR NESTED PROPORTIONS

```
**************************** SAMPLE CODE ****************************;
*     Definition:
*           1. t.visitsWeight: the name of your dataset of group form
*           2. beta0_p1, beta0_pr_0, beta0_pr_1, beta0_p2_0, beta0_p2_1:
*                       random intercepts for probabilities in all paths
*           3. var: variance of random effect
*           4. Weight: number of subjects for each path
*********************************************************************;

*-- calculate log likelihood;
proc nlmixed data=t.VisitsWeight itdetails tech=trureg maxiter=1000
maxfunc=2000;
      parms beta0_p1=0.1 beta0_pr_0=0.1 beta0_pr_1=0.1
        beta0_p2_0=0.1 beta0_p2_1=0.1 var=1;

      p1=min(max(1/(1+exp(-(beta0_p1+bj))),1E-10),1-1E-10);
    l_1=p1**A1 * (1-p1)**(1-A1);

      if A1=1 then do;
            pr_1=min(max(1/(1+exp(-(beta0_pr_1+bj))),1E-10),1-1E-10);
            l_2=pr_1**R * (1-pr_1)**(1-R);

            if R=1 then do;
                  p2_1=min(max(1/(1+exp(-(beta0_p2_1+bj))),1E-10),1-1E-10);
                  l_3=p2_1**A2 * (1-p2_1)**(1-A2); end; end;

      else if A1=0 then do;
            pr_0=min(max(1/(1+exp(-(beta0_pr_0+bj))),1E-10),1-1E-10);
            l_2=pr_0**R * (1-pr_0)**(1-R);

            if R=1 then do;
                  p2_0=min(max(1/(1+exp(-(beta0_p2_0+bj))),1E-10),1-1E-10);
                  l_3=p2_0**A2 * (1-p2_0)**(1-A2); end; end;

      if R=0 then ll=Weight*log(l_1*l_2);
      else if R=1 then ll=Weight*log(l_1*l_2*l_3);

      model A2 ~ general(ll);
      random bj ~ normal(0,log(exp(var))) subject=hospID out=re;
      estimate "p1" 1/(1+exp(-(beta0_p1)));
      estimate "pr_1" 1/(1+exp(-(beta0_pr_1)));
      estimate "p2_1" 1/(1+exp(-(beta0_p2_1)));
      estimate "RR_11" (1/(1+exp(-(beta0_p2_1))))/(1/(1+exp(-(beta0_p1))));
      predict p1 out=p1;
      predict pr_1 out=pr_1;
      predict p2_1 out=p2_1;
      ods output FitStatistics=lld;
      ods output AdditionalEstimates=Est;
run;

proc nlmixed data=t.VisitsWeight itdetails tech=trureg maxiter= 1000
maxfunc=2000;
      parms beta0_p=0.1 beta0_pr_0=0.1 beta0_pr_1=0.1 beta0_p2_0=0.1 var=1;

      p=min(max(1/(1+exp(-(beta0_p+bj))),1E-10),1-1E-10);
    l_1=p**A1 * (1-p)**(1-A1);

      if A1=1 then do;
            pr_1=min(max(1/(1+exp(-(beta0_pr_1+bj))),1E-10),1-1E-10);
            l_2=pr_1**R * (1-pr_1)**(1-R);

            if R=1 then do;
                  p=min(max(1/(1+exp(-(beta0_p+bj))),1E-10),1-1E-10);
                  l_3=p**A2 * (1-p)**(1-A2); end; end;

      else if A1=0 then do;
            pr_0=min(max(1/(1+exp(-(beta0_pr_0+bj))),1E-10),1-1E-10);
            l_2=pr_0**R * (1-pr_0)**(1-R);

            if R=1 then do;
                  p2_0=min(max(1/(1+exp(-(beta0_p2_0+bj))),1E-10),1-1E-10);
                  l_3=p2_0**A2 * (1-p2_0)**(1-A2); end; end;

      if R=0 then ll=Weight*log(l_1*l_2);
      else if R=1 then ll=Weight*log(l_1*l_2*l_3);

      model A2 ~ general(ll);
      random bj ~ normal(0,log(exp(var))) subject=hospID;
      ods output FitStatistics=lln;
run;

*-- likelihood ratio test;
data LRT;
      retain descr lln lld lrt pvalue;
      merge lln (rename=(value=lln)) lld(rename=(value=lld)); by descr;
      lrt=lln-lld;
      pvalue=1 - probchi(lrt, 1);
      if descr~='-2 Log Likelihood' then delete;
run;

*-- Output;
proc print data=Est; proc print data=LRT; run;
```

APPENDIX B

SUPPLEMENTARY TABLES

Table 10. Simulation results comparing the proposed method without random effects to the ordinary test for proportions under different scenarios with N=45,000.[a]

	Proposed method														Two-sample proportion test				
		p_1			$p_r^{(1)}$			$p_2^{(1)}$			$RR^{(1)}$				$RR^{(1)}$				
Difference	Simulation Est/Power	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Power	Simu-lation	Bias	Est. SD[b]	Est. SE[c]	Power[d]
None: $(p_1, p_2^{(1)})=$ (0.46,0.46)																			
Cluster effect:																			
0.0001	981	-0.01	0.24	0.23	<0.01	0.10	0.10	0.06	2.45	2.45	0.16	5.33	5.35	0.044	1000	0.13	5.36	5.36	0.045
0.003	977	<0.01	0.34	0.23	0.01	0.10	0.10	0.16	2.50	2.45	0.35	5.43	5.34	0.057	1000	0.32	5.44	5.34	0.056
0.005	982	<0.01	0.39	0.23	0.01	0.10	0.10	0.19	2.49	2.44	0.41	5.41	5.34	0.053	1000	0.42	5.43	5.33	0.052
0.007	984	0.01	0.44	0.23	0.01	0.11	0.10	0.29	2.52	2.44	0.62	5.45	5.33	0.057	1000	0.59	5.45	5.32	0.057
0.009	983	<0.01	0.48	0.23	0.02	0.11	0.10	0.34	2.51	2.44	0.74	5.40	5.33	0.051	1000	0.70	5.41	5.31	0.050
0.01	980	<0.01	0.51	0.23	0.02	0.11	0.10	0.38	2.51	2.44	0.83	5.38	5.33	0.056	1000	0.77	5.38	5.30	0.054
Mild: $(p_1, p_2^{(1)})=$ (0.30,0.46)																			
Cluster effect:																			
0.0001	1000	<0.01	0.22	0.22	<0.01	0.12	0.12	0.01	3.03	3.03	0.05	10.14	10.17	0.999	1000	0.05	6.65	6.66	0.999
0.003	999	0.01	0.30	0.22	0.01	0.13	0.12	0.12	3.03	3.03	0.35	10.18	10.15	1.000	1000	0.34	6.65	6.63	1.000
0.005	999	0.02	0.35	0.22	0.01	0.13	0.12	0.21	3.04	3.02	0.59	10.21	10.14	0.999	1000	0.57	6.67	6.61	0.999
0.007	999	0.03	0.39	0.22	0.02	0.13	0.12	0.29	3.05	3.02	0.82	10.20	10.12	0.999	1000	0.80	6.65	6.59	0.999
0.009	998	0.04	0.43	0.22	0.02	0.13	0.12	0.35	3.05	3.02	0.99	10.18	10.10	1.000	1000	0.96	6.63	6.57	1.000
0.01	1000	0.04	0.44	0.22	0.02	0.13	0.12	0.39	3.05	3.01	1.08	10.19	10.10	1.000	1000	1.08	6.62	6.56	1.000
Severe: $(p_1, p_2^{(1)})=$ (0.12,0.46)																			
Cluster effect:																			
0.0001	984	<0.01	0.16	0.15	<0.01	0.18	0.19	-0.16	4.85	4.79	-1.33	40.73	40.23	1.000	1000	-0.97	10.77	10.63	1.000
0.003	976	0.01	0.19	0.15	0.01	0.19	0.19	-0.03	4.82	4.78	-0.61	40.44	40.10	1.000	1000	-0.20	10.58	10.58	1.000
0.005	982	0.02	0.21	0.15	0.01	0.19	0.19	0.08	4.77	4.77	0.05	40.02	40.02	1.000	1000	0.25	10.49	10.54	1.000
0.007	973	0.03	0.23	0.15	0.02	0.19	0.19	0.12	4.77	4.77	0.10	40.00	39.94	1.000	1000	0.71	10.46	10.51	1.000
0.009	973	0.04	0.24	0.15	0.03	0.20	0.19	0.19	4.73	4.75	0.35	39.49	39.82	1.000	1000	0.98	10.37	10.47	1.000
0.01	975	0.05	0.25	0.15	0.03	0.19	0.19	0.28	4.73	4.75	0.95	39.39	39.79	1.000	1000	1.50	10.31	10.44	1.000

[a] Bias, SD and SE are in the scale of 100 times.

[b] Estimated standard deviation of ln(RR)

[c] Estimated standard error of ln(RR)

[d] This is the type I error under the scenario of no difference.

Table 11. Simulation results comparing the proposed method with random effects to the ordinary test for proportions under different scenarios with N=45,000.[a]

	Proposed method														Two-sample proportion test				
		p_1			$p_r^{(1)}$			$p_2^{(1)}$			$RR^{(1)}$				$RR^{(1)}$				
Difference	Simulation Est/Power	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Power	Simu-lation	Bias	Est. SD[b]	Est. SE[c]	Power [d]
None: $(p_1, p_2^{(1)})=$ (0.46,0.46)																			
Cluster effect: 0	987	-0.09	0.26	0.29	-0.15	0.16	0.09	-4.31	4.40	2.43	-9.20	9.45	5.29	0.023	1000	0.09	5.37	5.36	0.044
0.0001	991	-0.07	0.26	0.29	-0.13	0.16	0.10	-3.78	4.41	2.43	-8.08	9.48	5.29	0.027	1000	0.13	5.36	5.36	0.045
0.001	1000	-0.02	0.28	0.28	-0.04	0.13	0.10	-1.05	3.71	2.44	-2.24	8.02	5.33	0.042	1000	0.22	5.41	5.35	0.046
0.003	1000	<0.01	0.34	0.33	<0.01	0.10	0.10	<0.01	2.53	2.46	0.01	5.48	5.34	0.053	1000	0.32	5.44	5.34	0.056
0.005	1000	-0.01	0.39	0.39	<0.01	0.10	0.10	0.01	2.50	2.46	0.04	5.43	5.34	0.053	1000	0.42	5.43	5.33	0.052
0.007	1000	-0.01	0.44	0.44	<0.01	0.10	0.10	0.01	2.53	2.47	0.04	5.46	5.34	0.052	1000	0.59	5.45	5.32	0.057
0.009	1000	-0.01	0.49	0.48	<0.01	0.11	0.10	-0.01	2.52	2.48	-0.01	5.42	5.33	0.048	1000	0.70	5.41	5.31	0.050
Mild: $(p_1, p_2^{(1)})=$ (0.30,0.46)																			
Cluster effect: 0	984/740	-0.08	0.25	0.27	-0.18	0.20	0.12	-5.91	5.89	2.98	-19.40	19.40	9.96	0.922	1000	0.02	6.63	6.66	1.000
0.0001	990/776	-0.07	0.25	0.27	-0.16	0.20	0.12	-5.35	5.96	2.98	-17.57	19.62	9.98	0.929	1000	0.05	6.65	6.66	0.999
0.001	997/939	-0.01	0.25	0.25	-0.05	0.16	0.12	-1.76	4.83	3.02	-5.81	16.04	10.10	0.983	1000	0.13	6.65	6.65	0.999
0.003	1000/997	<0.01	0.30	0.30	<0.01	0.13	0.12	-0.11	3.21	3.04	-0.36	10.78	10.15	0.998	1000	0.34	6.65	6.63	1.000
0.005	1000/999	<0.01	0.35	0.34	<0.01	0.13	0.12	-0.04	3.09	3.04	-0.12	10.38	10.15	0.999	1000	0.57	6.67	6.61	0.999
0.007	1000	<0.01	0.39	0.38	<0.01	0.13	0.12	<0.01	3.05	3.04	0.01	10.22	10.14	0.999	1000	0.80	6.65	6.59	0.999
0.009	1000	<0.01	0.43	0.41	<0.01	0.13	0.12	-0.02	3.06	3.05	-0.07	10.22	10.13	1.000	1000	0.96	6.63	6.57	1.000
Severe: $(p_1, p_2^{(1)})=$ (0.12,0.46)																			
Cluster effect: 0	971	-0.06	0.18	0.18	-0.13	0.19	0.19	-13.21	11.36	4.39	-109.04	93.82	36.90	0.999	1000	-1.17	10.84	10.64	1.000
0.0001	978	-0.06	0.18	0.18	-0.12	0.19	0.19	-12.14	11.49	4.43	-100.16	94.90	37.21	0.999	1000	-0.97	10.77	10.63	1.000
0.001	974	-0.03	0.18	0.18	-0.07	0.20	0.19	-7.38	11.07	4.57	-61.00	91.81	38.34	1.000	1000	-0.76	10.69	10.61	1.000
0.003	991	<0.01	0.19	0.19	-0.02	0.19	0.19	-1.75	7.31	4.74	-14.51	61.02	39.76	1.000	1000	-0.20	10.58	10.58	1.000
0.005	991	0.01	0.21	0.20	-0.01	0.19	0.19	-0.45	5.47	4.77	-3.85	45.80	40.03	1.000	1000	0.25	10.49	10.54	1.000
0.007	998	0.01	0.22	0.22	<0.01	0.19	0.19	-0.20	4.88	4.78	-1.72	41.03	40.09	1.000	1000	0.71	10.46	10.51	1.000
0.009	994	0.01	0.24	0.23	<0.01	0.19	0.19	-0.10	4.70	4.78	-0.94	39.40	40.02	1.000	1000	0.98	10.37	10.47	1.000

[a] Bias, SD and SE are in the scale of 100 times.

[b] Estimated standard deviation of ln(RR)

[c] Estimated standard error of ln(RR)

[d] This is the type I error under the scenario of no difference.

Table 12. Simulation results comparing the proposed method without random effects to the ordinary test for proportions under different scenarios with N=500.[a]

	Proposed method														Two-sample proportion test				
		p_1			$p_r^{(1)}$			$p_2^{(1)}$			$RR^{(1)}$				$RR^{(1)}$				
Difference	Simulation Est/Power	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Power	Simu-lation	Bias	Est. SD[b]	Est. SE[c]	Power[d]
None: $\left(p_1, p_2^{(1)}\right)=$ (0.46,0.46)																			
Cluster effect:																			
0.0001	999	0.13	2.23	2.23	0.02	3.16	3.29	0.13	4.49	4.63	0.25	10.98	11.20	0.048	1000	0.28	11.00	11.23	0.050
0.003	1000	0.12	2.26	2.23	0.08	3.17	3.29	0.18	4.45	4.63	0.38	10.98	11.20	0.048	1000	0.38	10.94	11.22	0.049
0.005	1000	0.13	2.24	2.23	0.11	3.17	3.29	0.21	4.43	4.63	0.42	10.90	11.20	0.044	1000	0.42	10.87	11.21	0.045
0.007	1000	0.12	2.22	2.23	0.14	3.17	3.29	0.26	4.44	4.63	0.55	10.90	11.20	0.049	1000	0.55	10.85	11.20	0.048
0.009	999	0.12	2.24	2.23	0.18	3.21	3.29	0.31	4.47	4.63	0.65	10.91	11.20	0.044	1000	0.65	10.85	11.19	0.044
Mild: $\left(p_1, p_2^{(1)}\right)=$ (0.38,0.46)																			
Cluster effect:																			
0.0001	999	0.09	2.15	2.17	0.08	3.45	3.62	0.27	4.97	5.09	0.81	14.95	15.16	0.298	1000	0.80	12.24	12.52	0.286
0.003	999/998	0.11	2.19	2.17	0.14	3.50	3.62	0.25	4.93	5.09	0.73	15.05	15.14	0.310	1000	0.74	12.32	12.51	0.297
0.005	1000	0.13	2.17	2.17	0.18	3.53	3.62	0.33	4.95	5.09	0.89	15.02	15.13	0.317	1000	0.89	12.28	12.49	0.310
0.007	1000	0.15	2.18	2.17	0.20	3.53	3.62	0.36	4.92	5.09	0.87	14.91	15.12	0.310	1000	0.87	12.18	12.48	0.305
0.009	998	0.16	2.20	2.17	0.24	3.54	3.62	0.40	4.93	5.08	0.94	14.86	15.11	0.313	1000	0.97	12.14	12.46	0.305
Severe: $\left(p_1, p_2^{(1)}\right)=$ (0.30,0.46)																			
Cluster effect:																			
0.0001	998	0.09	2.07	2.05	0.05	3.97	4.07	0.21	5.64	5.73	0.93	21.61	21.93	0.774	999	0.93	14.01	14.33	0.745
0.003	1000	0.10	2.07	2.05	0.15	4.02	4.07	0.27	5.66	5.73	1.13	21.73	21.92	0.772	1000	1.13	14.10	14.31	0.752
0.005	1000	0.10	2.04	2.05	0.16	4.03	4.07	0.33	5.64	5.73	1.32	21.77	21.92	0.781	1000	1.32	14.04	14.29	0.763
0.007	1000	0.12	2.06	2.05	0.20	4.07	4.07	0.36	5.64	5.72	1.33	21.79	21.90	0.784	1000	1.33	14.05	14.28	0.763
0.009	1000	0.13	2.08	2.05	0.25	4.08	4.07	0.44	5.64	5.72	1.53	21.81	21.89	0.783	1000	1.53	14.07	14.25	0.766

[a] Bias, SD and SE are in the scale of 100 times.

[b] Estimated standard deviation of ln(RR)

[c] Estimated standard error of ln(RR)

[d] This is the type I error under the scenario of no difference.

Table 13. Simulation results comparing the proposed method with random effects to the ordinary test for proportions under different scenarios with N=500.[a]

	Proposed method														Two-sample proportion test				
		p_1			$p_r^{(1)}$			$p_2^{(1)}$			$RR^{(1)}$				$RR^{(1)}$				
Difference	Simulation Est/Power	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Power	Simu-lation	Bias	Est. SD[b]	Est. SE[c]	Power[d]
None: $\left(p_1, p_2^{(1)}\right)=$ (0.46,0.46)																			
Cluster effect: 0	873	-0.48	2.39	2.43	-1.80	3.46	3.52	-3.62	5.59	4.83	-6.75	12.62	11.26	0.071	1000	0.33	11.07	11.23	0.055
0.0001	882	-0.47	2.37	2.45	-1.79	3.44	3.53	-3.84	5.43	4.85	-7.25	12.18	11.24	0.066	1000	0.28	11.00	11.23	0.050
0.001	890	-0.45	2.39	2.45	-1.80	3.46	3.53	-3.71	5.45	4.85	-7.00	12.18	11.25	0.065	1000	0.25	10.99	11.24	0.050
0.003	893	-0.44	2.38	2.44	-1.58	3.56	3.52	-3.52	5.48	4.83	-6.61	12.32	11.24	0.073	1000	0.38	10.94	11.22	0.049
0.005	883	-0.39	2.40	2.43	-1.56	3.52	3.51	-3.43	5.45	4.83	-6.53	12.13	11.24	0.062	1000	0.42	10.87	11.21	0.045
0.007	891	-0.41	2.34	2.44	-1.42	3.55	3.52	-3.07	5.48	4.84	-5.70	12.16	11.27	0.058	1000	0.55	10.85	11.20	0.048
0.009	880	-0.43	2.35	2.43	-1.35	3.56	3.51	-2.94	5.41	4.84	-5.36	12.22	11.29	0.053	1000	0.65	10.85	11.19	0.044
0.01	900	-0.35	2.34	2.43	-1.29	3.52	3.50	-2.91	5.35	4.82	-5.45	12.11	11.26	0.054	1000	0.71	10.86	11.18	0.042
Mild: $\left(p_1, p_2^{(1)}\right)=$ (0.38,0.46)																			
Cluster effect: 0	913/866	-0.48	2.28	2.43	-2.08	3.94	3.90	-4.41	6.24	5.32	-9.98	16.99	15.08	0.219	1000	0.80	12.34	12.52	0.283
0.0001	914/861	-0.54	2.30	2.40	-2.11	3.85	3.88	-4.26	6.10	5.30	-9.81	16.80	15.12	0.230	1000	0.80	12.24	12.52	0.286
0.001	930/864	-0.50	2.27	2.40	-1.99	3.96	3.87	-4.18	6.22	5.30	-9.30	16.95	15.12	0.231	1000	0.74	12.20	12.52	0.289
0.003	913/852	-0.50	2.32	2.39	-1.88	3.96	3.86	-4.14	6.18	5.29	-9.19	16.90	15.12	0.254	1000	0.74	12.32	12.51	0.297
0.005	903/850	-0.40	2.33	2.38	-1.76	3.98	3.85	-3.81	6.32	5.28	-8.58	17.26	15.10	0.241	1000	0.89	12.28	12.49	0.310
0.007	904/857	-0.36	2.32	2.36	-1.66	3.93	3.82	-3.54	6.22	5.26	-7.98	17.05	15.11	0.257	1000	0.87	12.18	12.48	0.305
0.009	901/863	-0.32	2.31	2.36	-1.55	3.92	3.83	-3.37	6.11	5.27	-7.66	16.78	15.11	0.257	1000	0.97	12.14	12.46	0.305
0.01	907/870	-0.30	2.31	2.35	-1.45	3.88	3.81	-3.16	6.06	5.26	-7.14	16.82	15.12	0.259	1000	1.09	12.20	12.45	0.307
Severe: $\left(p_1, p_2^{(1)}\right)=$ (0.30,0.46)																			
Cluster effect: 0	937/923	-0.49	2.19	2.26	-2.44	4.44	4.33	-4.89	7.11	5.92	-13.65	24.46	21.75	0.696	1000	1.04	14.06	14.32	0.752
0.0001	950/934	-0.48	2.19	2.24	-2.45	4.39	4.32	-4.93	7.16	5.90	-13.77	24.80	21.74	0.690	1000	0.93	14.01	14.33	0.745
0.001	958/947	-0.45	2.17	2.24	-2.36	4.47	4.30	-4.70	7.00	5.90	-13.12	24.43	21.75	0.702	1000	1.14	13.85	14.31	0.758
0.003	934/920	-0.45	2.21	2.23	-2.28	4.48	4.30	-4.61	7.11	5.90	-12.87	24.66	21.76	0.698	1000	1.13	14.10	14.31	0.752
0.005	947/935	-0.41	2.17	2.21	-2.10	4.59	4.27	-4.32	7.19	5.86	-12.09	24.93	21.76	0.704	1000	1.32	14.04	14.29	0.763
0.007	945/937	-0.36	2.19	2.21	-2.00	4.57	4.27	-4.13	7.09	5.87	-11.60	24.81	21.76	0.712	1000	1.33	14.05	14.28	0.763
0.009	926/918	-0.32	2.20	2.19	-1.86	4.55	4.25	-3.84	6.98	5.86	-10.83	24.45	21.78	0.718	1000	1.53	14.07	14.25	0.766
0.01	925/919	-0.27	2.19	2.20	-1.76	4.51	4.25	-3.76	6.91	5.86	-10.77	24.34	21.74	0.704	1000	1.72	14.07	14.23	0.773

[a] Bias, SD and SE are in the scale of 100 times. [b] Estimated standard deviation of ln(RR)
[c] Estimated standard error of ln(RR). [d] This is the type I error under the scenario of no difference.

BIBLIOGRAPHY

AHRQ (Agency for Healthcare Research and Quality). (2009) AHRQ *Effective Health Care glossary* http://effectivehealthcare.ahrq.gov/tools.cfm?tooltype=glossary&TermID=118 (accessed April 4, 2009)

Alessandrini, E. A., Lavelle, J. M., Grenfell, S. M., Jacobstein, C. R., & Shaw, K. N. (2004). Return visits to a pediatric emergency department. *Pediatric emergency care*, 20(3), 166-171.

Becker, D. M., Segal, J., Vaidya, D., Yanek, L. R., Herrera-Galeano, J. E., Bray, P. F., ... & Faraday, N. (2006). Sex differences in platelet reactivity and response to low-dose aspirin therapy. *JAMA: The Journal of the American Medical Association*, 295(12), 1420-1427.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Chapman & Hall/CRC.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Brown, H., & Prescott, R. (2006). *Applied mixed models in medicine*. Wiley.

Cai, T., Tian, L., Wong, P. H., & Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2), 270-282.

Chadwick, B. A., & Heaton, T. B. (1999). *Statistical handbook on the American family*. Greenwood Publishing Group.

Chatterjee, S., Laudato, M., & Lynch, L. A. (1996). Genetic algorithms and their statistical applications: an introduction. *Computational Statistics & Data Analysis*, 22(6), 633-651.

Clarke, S. C., & Wilson, B. F. (1994). The relative stability of remarriages: A cohort approach using vital statistics. *Family Relations*, 305-310.

Cook, D. I., Gebski, V. J., & Keech, A. C. (2004). Subgroup analysis in clinical trials. *Medical Journal of Australia*, 180(6), 289-292.

Deng, K., Pineau, J., & Murphy, S. A. (2012). Active Learning for Developing Personalized Treatment. *arXiv preprint arXiv*:1202.3714.

Foster, J. C., Taylor, J. M., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24), 2867-2880.

Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning.

Gray, J. B., & Fan, G. (2008). Classification tree analysis using TARGET. *Computational Statistics & Data Analysis*, 52(3), 1362-1372.

Grubinger, T., Zeileis, A., & Pfeiffer, K. P. (2011). *evtree: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R* (No. 2011-20).

Holland, J. H. (1975). Adaptation in natural and artificial systems, University of Michigan press. *Ann Arbor, MI*, 1(97), 5.

Huang, Y., Gilbert, P. B., & Janes, H. (2012). Assessing Treatment Selection Markers using a Potential Outcomes Framework. Biometrics.

Imai, K., & Ratkovic, M. Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation.

LeBlanc, M., & Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422), 457-467.

Lipkovich, I., Dmitrienko, A., Denne, J., & Enas, G. (2011). Subgroup identification based on differential effect search—A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in medicine*, 30(21), 2601-2621.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. OUP Oxford.

Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in medicine*, 21(19), 2917-2930.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1), 1-11.

Ronco, C., & Honore, P. M. (2008). Renal support in critically ill patients with acute kidney injury. *N Engl J Med*, 359(1), 82-4.

Rubin, D. B., & van der Laan, M. J. (2007). Statistical Issues and Limitations in Personalized Medicine Research with Clinical Trials. *The International Journal of Biostatistics,* 8(1).

Sacks, F. M., Pfeffer, M. A., Moye, L. A., Rouleau, J. L., Rutherford, J. D., Cole, T. G., ... & Braunwald, E. (1996). The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. *New England Journal of Medicine*, 335(14), 1001-1009.

Shen, H., Herzog, W., Drolet, M., Pakyz, R., Newcomer, S., Sack, P., ... & Shuldiner, A. R. (2009). Aspirin Resistance in healthy drug-naive men versus women (from the Heredity and Phenotype Intervention Heart Study). *The American journal of cardiology*, 104(4), 606-612.

Skinner, K. B., Bahr, S. J., Crane, D. R., & Call, V. R. (2002). Cohabitation, Marriage, and Remarriage A Comparison of Relationship Quality Over Time. *Journal of Family Issues*, 23(1), 74-90.

Sleight, P. (2000). Debate: Subgroup analyses in clinical trials: fun to look at—but don't believe them. *Curr Control Trials Cardiovasc Med*, 1(1), 25-27.

Song, X., & Pepe, M. S. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics*, 60(4), 874-883.

Su, X., Zhou, T., Yan, X., Fan, J., & Yang, S. (2008). Interaction trees with censored survival data. *The international journal of biostatistics*, 4(1), 1-26.

Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research*, 10, 141-158.

Su, X., Meneses, K., McNees, P., & Johnson, W. O. (2011). Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3), 457-474.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2), 331-355.

Wang, R., & Ware, J. H. (2012). Detecting moderator effects using subgroup analyses. *Prevention Science*, 1-10.

Zhou, H., Chen, J., Rissanen, T. H., Korrick, S. A., Hu, H., Salonen, J. T., & Longnecker, M. P. (2007). Outcome-dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology*, 18(4), 461-468.

Zhao, Y., Zeng, D., Socinski, M. A., & Kosorok, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4), 1422-1433.

Zhao, Y., Zeng, D., Rush, A. J., & Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499), 1106-1118.

Zorman, M., Podgorelec, V., Kokol, P., Peterson, M., & Lane, J. (2000). Decision tree's induction strategies evaluated on a hard real world problem. *In Computer-Based Medical Systems, 2000. CBMS 2000. Proceedings. 13th IEEE Symposium on* (pp. 19-24). IEEE