UNIVERSITY OF CAMBRIDGE

Methods for Using Biomarker Information in Randomized Clinical Trials

Jixiong Wang

Hughes Hall

This dissertation is submitted on September, 2020 for the degree of Doctor of Philosophy.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other university or similar institution except as declared in the text. This dissertation does not exceed the prescribed limit of 60,000 words.

Chapter 2 is joint work with Ashish Patel. The work in Chapter 2 has been published in *Biometrics* [76]. The work in Chapter 3 has also been submitted for publication. A manuscript will be prepared to describe the methodology, simulation and real data application results presented in Chapter 4.

Jixiong Wang
April, 2021

Abstract

Methods for Using Biomarker Information in Randomized Clinical Trials

Jixiong Wang

Advances in high-throughput biological technologies have led to large numbers of potentially predictive biomarkers becoming routinely measured in modern clinical trials. Biomarkers which influence treatment efficacy may be used to find subgroups of patients who are most likely to benefit from a new treatment. Consequently, there is a growing interest in better approaches to identify biomarker signatures and utilize the biomarker information in clinical trials.

The first focus of this thesis is on developing methods for detecting biomarker-treatment interactions in large-scale trials. Traditional interaction analysis, using regression models to test biomarker-treatment interactions one biomarker at a time, may suffer from poor power when there is a large multiple testing burden. I adapt recently proposed two-stage interaction detecting procedures for application in randomized clinical trials. I propose two new stage 1 multivariate screening strategies using lasso and ridge regressions to account for correlations among biomarkers. For these new multivariate screening strategies, I prove the asymptotic between-stage independence, required for family-wise error rate control. Simulation and real data application results are presented which demonstrate greater power of the new strategies compared with previously existing approaches.

The second focus of this thesis is on developing methods for utilizing biomarker information during the course of a randomized clinical trial to improve the informativeness of results. Under the adaptive signature design (ASD) framework, I propose two new classifiers that more efficiently leverage biomarker signatures to select a subgroup of patients who are most likely to benefit from the new treatment. I provide analytical arguments and demonstrate through simulations that these two proposed classification criteria can provide at least as good, and sometimes significantly greater power than the originally proposed ASD classifier.

Third, I focus on an important issue in the statistical analysis of interactions for binary outcomes, which is pertinent to both topics above. Testing for biomarker-treatment

interactions with logistic regression can suffer from an elevated number of type I errors due to the asymptotic bias of the interaction regression coefficient under model misspecification. I analyze this problem in the randomized clinical trial setting and propose two new debiasing procedures, which can offer improved family-wise error rate control in various simulated scenarios.

Finally, I summarize the main contributions from the work above, discuss some practical limitations as well as their real world value, and prioritize future directions of research building upon the work in this thesis.

Acknowledgements

There are many people without whom I would not have come this far. First and foremost, I would like to thank my supervisors Dr. Paul Newcombe and Professor James Wason for their unwavering support throughout my PhD study. Paul has donated enormous time, effort and expertise for the duration of this work and has been a constant source of invaluable advice over the last three years. James' sharp intuition and incredible insights have always enabled me to think deeply and innovatively, and continuously improve the quality of my work. Their kindness has made my PhD study a much more enjoyable experience. I am also grateful to Ashish Patel, whom I collaborated with on theoretical work. His sound comments and rigorousness have been invaluable throughout the process.

I am indebted to my wife, Huiying Zhu, whose love has always kept me going. Without her support, I am convinced that I would not be able to finish this work today. Special thanks to my elder daughter Yino, whose sense of humor has made me laugh every day, and my little daughter Yimo, whose crying and smiling made me not feel alone on countless working nights. Thanks to my dear parents - Qiang Wang and Qi Pan - and other family members for their love, care and support. Finally, to my grandma in heaven, I will never forget the beautiful summer day when you were playing with my kids in the backyard of our Cambridge home - which is like it happened yesterday.

Contents

1 Introduction **15**
1.1 Motivation . . . 15
1.2 Methods for detecting biomarker-treatment interactions . . . 16
1.2.1 Standard one-biomarker-at-a-time interaction tests . . . 17
1.2.2 Gene-environment interaction studies . . . 18
1.2.2.1 Case-only tests . . . 19
1.2.2.2 Two-stage interaction tests . . . 20
1.3 Multiple testing correction procedures . . . 22
1.3.1 Bonferroni correction . . . 23
1.3.2 Šidák correction . . . 23
1.3.3 Holm-Bonferroni method and other family-wise error rate controlling procedures . . . 24
1.3.4 Benjamini-Hochberg method and other false discovery rate controlling procedures . . . 24
1.4 Sparse regression methods . . . 25
1.4.1 Ridge regression . . . 25
1.4.2 Lasso regression . . . 26
1.5 Adaptive signature design (ASD) for randomized clinical trials . . . 26
1.5.1 Cross-validated ASD design . . . 27
1.6 Trial data sets used as case studies through this thesis . . . 31
1.6.1 START trial . . . 31
1.6.2 STOPAH trial . . . 32
1.6.3 PREVAIL trial . . . 33
1.7 Thesis overview . . . 34

2 Sparse regression screening procedures in a two-stage interaction detecting framework **35**
2.1 Introduction . . . 35
2.2 Lack of applicability of responder-only tests in randomized clinical trials . . . 36
2.3 Two-stage interaction tests in randomized clinical trials . . . 37

2.4 New stage 1 sparse regression screening procedures accounting for biomarker-biomarker correlations . . . 38
2.4.1 Ridge regression screening . . . 38
2.4.2 Lasso regression screening . . . 39
2.5 Asymptotic between-stage independence for stage 1 sparse regression screening 39
2.6 Weighted false discovery rate controlling procedures in a two-stage interaction detecting framework . . . 42
2.7 Simulation studies . . . 43
2.8 Data applications . . . 51
2.8.1 START trial . . . 51
2.8.2 STOPAH trial . . . 52
2.8.3 PREVAIL trial . . . 53
2.8.4 Empirical between-stage correlation . . . 54
2.9 Discussion . . . 55

3 Adaptive signature design using biomarker-treatment interaction information to maximize treatment effect test statistics 59
3.1 Introduction . . . 59
3.2 Methods . . . 61
3.2.1 Multivariate risk difference (MRD) classifier . . . 61
3.2.2 Multivariate gradient-based (MGB) classifier . . . 62
3.2.3 Relationships between the ASD, MRD and MGB classifiers . . . 64
3.3 Simulation studies . . . 65
3.4 Data applications . . . 72
3.4.1 START trial . . . 72
3.4.2 STOPAH trial . . . 73
3.5 Discussion . . . 76

4 De-biased logistic regression biomarker-treatment interaction estimator under model misspecification 81
4.1 Introduction . . . 81
4.2 Asymptotic bias of the gene-environment interaction estimator under model misspecification . . . 82
4.3 Asymptotic bias of the biomarker-treatment interaction estimator under model misspecification . . . 84
4.4 De-biased biomarker-treatment interaction estimator . . . 85
4.5 Simulation studies . . . 88
4.6 Data applications . . . 95
4.6.1 PREVAIL trial . . . 95

4.7 Discussion . . . 97

5 Conclusions and future directions **99**
5.1 Conclusions . . . 99
5.2 Future directions . . . 102
5.2.1 Extension of the ASD framework to account for biomarker-biomarker correlations . . . 102
5.2.2 Extension of the two-stage interaction testing framework to binary outcomes with family-wise error rate control . . . 103
5.3 Concluding remark . . . 105

Bibliography **107**

A Derivations **117**
A.1 Power of case-only interaction tests . . . 117
A.2 Between-stage independence proof: Murcray et al. . . . 118
A.3 Between-stage independence proof: Dai et al. . . . 121
A.4 Discussion of alternative family-wise error rate controlling methods . . . 122
A.5 Proof of independence between stage 1 sparse regression screening and stage 2 standard interaction tests . . . 123
A.5.1 Proof of Lemma 2.5.2 . . . 123
A.5.2 Proof of Corollary 2.5.2.1 . . . 124
A.5.3 Proof for the lasso screening test . . . 124
A.6 Calculating the test statistic gradient for the MGB classifier . . . 126
A.7 Extending the MGB classifier . . . 126
A.8 A proposed framework for detecting biomarker-treatment interactions using Bayesian variable selection . . . 127
A.9 Asymptotic score functions for logistic regression . . . 129
A.9.1 Approximating the sigmoid function . . . 129
A.9.2 Asymptotic score functions . . . 132
A.10 De-biased biomarker-treatment interaction estimator under alternative treatment coding . . . 134

B Numerical analysis **135**
B.1 Relationship between the MRD and MGB classifiers . . . 135
B.2 Relationship between the ASD and MRD classifiers . . . 136

C Additional simulation results **139**
C.1 Sparse regression screening procedures . . . 139
C.2 The MGB classifier using Bayesian variable selection . . . 141

C.3 De-biasing procedures under alternative treatment coding 142

Chapter 1

Introduction

1.1 Motivation

Recent developments in medicine have seen a shift toward targeted therapeutics [13]. It has been shown that individual variability can often contribute to differences in response to the same treatment [41]. For example, patients with leukemia respond to the treatment with all-trans retinoic acid if they have the PML-RARA translocation and patients with breast cancer benefit from the targeted antibody drug trastuzumab if the gene ERBB2 is over-expressed [64]. Conversely, use of some drugs can lead to increased risk to patients with specific genetic variants, e.g. a strong association of carbamazepine-induced Stevens-Johnsons syndrome and the human leukocyte antigen-B (HLA-B)*1502 allele was reported [48]. Another example is that the Class II allele HLA-DRB1*07:01 has been associated with lapatinib-induced liver injury [58]. Detecting such interactions between biomarkers and treatments in randomized clinical trials is of growing interest.

Discovering biomarker-treatment interactions helps identify predictive biomarkers[1] [62]: Biomarkers which influence treatment efficacy can be used to find subgroups of patients who are most likely to benefit from the new treatment, as well as to predict subgroup treatment effects [78]. When there are known predictive biomarkers, new adaptive design approaches can be used in settings where there are genetically-driven subgroups to improve efficiency [80]. Furthermore, the discovery of novel biomarker-treatment interactions may result in the identification of new disease susceptibility loci, providing insights into the biology of diseases. Such outcomes are very much aligned with the goals of precision medicine: to enable the provision of "the right drug at the right dose to the right patient" [13].

The first focus of this thesis is on developing methods for detecting biomarker-treatment interactions in large-scale studies of human populations. This is a non-trivial task, which

[1] A predictive biomarker is used to predict the effect of a therapeutic intervention. In contrast, a prognostic biomarker is used to predict patients' disease progression regardless of treatment.

faces several challenging problems [51]. Traditional interaction analysis, using regression models to test biomarker-treatment interactions one biomarker at a time, may suffer from poor power when there is a large multiple testing burden, for example when performing such analysis on a genome-wide scale for genetic biomarkers. Standard genotyping microarrays measure half a million or more variants and, when combined with whole genome imputation, can lead to millions of biomarkers to consider. Another type of omics, metabolomics - the measurement of metabolite concentrations in the body - may have a more direct effect on drug efficacy and is also becoming increasingly widely assayed [2, 33, 74]. As -omics technologies continue to drop in price and become routinely measured in clinical trials, interaction testing frameworks which are designed to scale to large numbers of covariates will become ever more important. Another limitation of the traditional approach testing each biomarker at a time is that it fails to model correlations between covariates. When there exists strong multicollinearity, confounding due to correlations can lead to a lot of false positives, i.e. the precision to detect true signals can be significantly reduced.

The second topic of this thesis is to develop methods for utilizing biomarker information during the course of a randomized trial. In molecularly targeted cancer drugs, therapies are often effective only for a subset of patients [34]. Genomic technologies, such as microarrays and single-nucleotide polymorphism genotyping, provide rich biomarker panels from which to develop potential signatures[2] to discriminate the subset of patients, who will most likely benefit from a targeted therapy. In the context of clinical trials, the predictive biomarkers hold great potential to improve trial efficacy [83, 80]. However, there has so far only been limited work on optimal methods for identifying and utilizing predictive biomarkers to improve efficiency of randomized clinical trials.

1.2 Methods for detecting biomarker-treatment interactions

This section provides a formal description of the biomarker-treatment interaction detecting problem, and introduces some approaches in existing literature. The challenges of detecting biomarker-treatment interactions in large-scale studies of human populations are also discussed. This motivates the development of new methods in the subsequent chapters.

[2]Biomarker signature is the behavior or the pattern of a set of biomarkers that can be used to make prediction. For example, a predictive biomarker signature can determine patients who are more likely to respond to a specific treatment.

1.2.1 Standard one-biomarker-at-a-time interaction tests

In the context of randomized clinical trials, one can test each biomarker in turn for a biomarker-treatment interaction using the following generalized linear model

$$G\{E(Y_i \mid X_{ij}, T_i)\} = \beta_{0_j} + \beta_{X_j} X_{ij} + \beta_T T_i + \beta_{X_j \times T} X_{ij} \times T_i \tag{1.1}$$

with Y_i denoting the response outcome, T_i the binary treatment-control indicator, and $X_{i1}, \ldots, X_{im}$ representing the values of m biomarkers, for the ith patient. G is a canonical link function that depends on which parametric distribution the outcome follows. The null hypothesis $H_{0_j} : \beta_{X_j \times T} = 0$ could be tested for each $j = 1, \ldots, m$, e.g. using a Wald test with the Bonferroni correction applied to preserve the family-wise error rate (the probability of at least one type I error).

The intercept term β_{0_j} is the expected value of the linear predictor $G\{\cdot\}$ when all covariates are 0; the coefficient β_{X_j} can be interpreted as the addictive factor by which the expected linear predictor changes given a unit increase of the corresponding covariate ("main effect"), when there is no interaction between X_{ij} and T_i; lastly the interaction coefficient $\beta_{X_j \times T}$ measures the departure from these corresponding main effects. Another way to interpret the interaction term $\beta_{X_j \times T} X_{ij} \times T_i$ in (1.1) is that the effect of T_i upon the linear predictor now involves X_{ij}, which can be expressed as $\beta_T + \beta_{X_j \times T} X_{ij}$, since we are able to write the model form as

$$G\{E(Y_i \mid X_{ij}, T_i)\} = \beta_{0_j} + \beta_{X_j} X_{ij} + (\beta_T + \beta_{X_j \times T} X_{ij}) \times T_i$$

In contrast, when there is no interaction between X_{ij} and T_i, i.e. $\beta_{X_j \times T} = 0$, the treatment effect is a constant β_T, which does not depend on the value of X_{ij}.

In the generalized linear model (1.1), G is a canonical link function: To evaluate a quantitative outcome in linear regression, an identity function is used [85]; for a binary response outcome, G can be a logit function [40]; in a cohort study, a log-linear model is typically used to estimate relative risks [7].

The number of biomarkers m to be considered in modern clinical trials is potentially large. For example, standard genome-wide association study (GWAS) micro-arrays measure several hundred thousand variants scattered throughout the genome. Next generation sequencing, which is becoming increasingly prevalent, can measure the genotype at every single nucleotide, leading to millions of variants. Given a desired overall family-wise error rate $\overline{\alpha}$, a Bonferroni correction [21] requires an adjusted significance level for each individual test to be $\overline{\alpha}/m$. Thus, the traditional interaction analysis can significantly lack power, because this forms a multiple testing problem when there is a large number of biomarkers to be considered [31]. As there is a trade-off between the type I error rate and the power of a hypothesis test, other less stringent correction methods can be more suitable

in scenarios where multiple testing adjustment is considered less important (depending on the end goal of predictive biomarker discovery).

Another drawback of the traditional approach is that the univariate tests can result in a lot of false positives when there exist substantial correlations between biomarkers. Multivariate modeling allows testing each predictor, while accounting for correlations with the other predictors. However, for a high-dimension, low-sample-size data set, a traditional multivariate regression analysis is not feasible because the maximum likelihood solution is not uniquely defined. This motivated the work in Chapter 2 to seek modern penalized regression procedures such as ridge and lasso regressions to model correlated high-dimensional data. These techniques have proven useful for feature selection in genomics [82, 49, 55], but there has so far been little exploration of their utility in the context of randomized clinical trials to identify features associated with the trial response.

1.2.2 Gene-environment interaction studies

The topic of detecting biomarker-treatment interactions in randomized clinical trials is closely related to gene-environment interaction studies, which have been a focus of genetic epidemiology for years. There is now a significant literature on statistical methods for discovering gene-environment interactions, which exploit aspects of the study design to improve power thus mitigating the multiple testing burden. These include case-only tests [59], empirical Bayes [52], Bayesian model averaging [45], and two-stage tests with different screening procedures [43, 53, 38, 30, 79]. Some of these methods seek to leverage the reasonable assumption of independence between most environment variables and genotypes at the population level. To alleviate the multiple testing burden, two-stage methods use independent information from the data to perform a screening test to select a subset of genetic biomarkers, and then only test interactions within this reduced set. Since there is a clear analogy to gene-environment interaction problems, in this thesis, I examine how existing gene-environment interaction testing methods may be modified so that they are transferable to the biomarker-treatment setting [16, 18, 19, 78]. Table 1.1 summarizes shared and, more importantly, key different features between these two types of studies.

Table 1.1: Comparing gene-environment interaction studies and biomarker-treatment interaction studies

Gene-Environment Interaction Studies	**Biomarker-Treatment Interaction Studies**
Multiple testing issues.	
Importance of accounting for correlations between covariates.	
Environmental factors are not necessarily independent of genetic variants, although it is usually assumed.	Biomarker-treatment independence is guaranteed through randomization.
Disease outcomes are usually rare events, under this and other assumptions, case-only methods give valid inference and can improve power.	Treatment responders are not usually rare, except for some scenarios, e.g. prevention trials.

In the following subsections, I introduce some existing gene-environment interaction testing methods. Their applicability to predictive biomarker discovery in randomized clinical trials will be further discussed in Chapter 2.

1.2.2.1 Case-only tests

Analogously to the model of the form (1.1), in gene-environment interaction studies, one can test each genetic biomarker (e.g. gene expression or single-nucleotide polymorphism) for a gene-environment interaction using the model:

$$G\{E(Y_i \mid X_{ij}, e_i)\} = \beta_{0_j} + \beta_{X_j} X_{ij} + \beta_e e_i + \beta_{X_j \times e} X_{ij} \times e_i \tag{1.2}$$

where Y_i denotes the response outcome, e_i the environmental factor, and $X_{i1}, \ldots, X_{im}$ represent the values of m genetic biomarkers, for the ith patient. G is a canonical link function. The null hypothesis $H_{0_j} : \beta_{X_j \times e} = 0$ could be tested for each $j = 1, \ldots, m$ with the Bonferroni correction applied for family-wise error rate control. This type of approach is sometimes referred to as "case-control" analysis when Y_i is binary.

To improve power, an alternative type of approach, i.e. "case-only" analysis [59], was proposed for detecting gene-environment interactions under the assumptions: 1) the environmental factor and genetic variants occur independently; 2) the disease is rare and individuals are either "affected" (cases) or "unaffected" (controls), i.e. the outcome is binary. This type of test corresponds to the logistic regression model of the following form:

$$logit\{E(e_i \mid X_{ij}, R_i = 1)\} = \gamma_{0_j} + \gamma_{X_j} X_{ij} \tag{1.3}$$

where R_i denotes the ith patient's binary response outcome. Notice that we use Y_i in (1.2) to represent the general type of response outcome (e.g. continuous, binary or categorical) and R_i in (1.3) for the binary response outcome, to highlight that the case-only analysis is only applicable for binary outcomes. The gene-environment independence condition and the rareness of response events are assumed, i.e. $pr(R_i = 1 \mid X_{ij}) \approx 0$. Under such conditions, it can be shown that the estimator of γ_{X_j} in equation (1.3) is a consistent estimator of the interaction coefficient $\beta_{X_j \times e}$ in equation (1.2) for the standard interaction test. The original mathematical derivation of this estimator and what conditions its consistency relies on was shown in [59]. A re-worked, more detailed derivation is presented in Appendix A.1 for reference. When the response is not rare, the estimator is biased in either a positive or negative direction depending on the sign of the true interaction effect.

In gene-environment interaction studies, a case-only analysis can be substantially more powerful than a case-control analysis. Cases within the current data set were oversampled relative to their prevalence in the population. Thus, given the disease is rare, a case-only test is equivalent to a comparable case-control test with infinitely many controls. Conceptually, when genetic biomarkers are independent of the environmental factor within the source population, gene-environment interaction will induce gene-environment association within the oversampled cases, which is captured by the model (1.3).

Another assumption the case-only tests rely on is gene-environment independence. If this condition is violated at the population level, the case-only estimator will be biased, because now the gene-environment association within cases is not only introduced by the interaction but also the gene-environment association from the source population. This motivated proposals of empirical Bayes [52] and also Bayes model averaging [45] approaches combining case-only and case-control tests, based on prior evidence of gene-environment dependence.

1.2.2.2 Two-stage interaction tests

To alleviate the multiple testing burden, two-stage approaches have been gaining traction for interaction testing. Two-stage approaches use a screening test as a filtering stage (stage 1) to select a subset of genetic biomarkers, and then in stage 2, only test interactions within the reduced set of genetic biomarkers (the Bonferroni Correction only accounts for this reduced set of genetic biomarkers, resulting in a less stringent significance level used in each single test), thus increasing power. Typically, a powerful but less robust test is used in stage 1, with a robust test being used in stage 2 {e.g. using the one-biomarker-at-a-time model (1.2)}. To preserve the overall family-wise error rate, two-stage approaches rely on the stage 1 screening tests being independent of the final stage 2 tests. More detail on how this can be established follows.

In gene-environment studies, several testing procedures have been proposed to be used

to prioritize genetic biomarkers in the screening stage. Most of them fall into one of three categories:

1. Testing marginal effects of genetic biomarkers (marginal association screening tests) [43].

2. Testing correlations between genetic biomarkers and environmental factors of interest (case-only style gene-environment association screening tests) [53].

3. A combination of 1 and 2 [30, 38].

A common stage 1 screening test used in two-stage interaction testing is a marginal association test [43]. The marginal effect of a biomarker on the outcome can be modeled in a regression model of the form

$$G\{E(Y_i \mid X_{ij})\} = \delta_{0_j} + \delta_{X_j} X_{ij} \tag{1.4}$$

The screening procedure is conducted by testing the null hypothesis $H_{0_j} : \delta_{X_j} = 0$ for $j = 1, \ldots, m$, with a pre-specified significance level $\alpha_1 \in (0, 1)$. In stage 2, one then tests interactions using the one-biomarker-at-a-time model (1.2) within the set of genetic biomarkers selected at stage 1 (those with null hypotheses rejected at the specified significance level). Another way to utilize stage 1 information is to test all m genetic biomarkers in stage 2 using weighted significance levels, that add up to the targeted error rate $\overline{\alpha}$, based on ranked genetic biomarkers from stage 1. One possible weighting scheme [39] is: the B most significant biomarkers, i.e. with lowest p-values in stage 1, are compared with an adjusted significance level $(\overline{\alpha}/2)/B$, the next $2B$ biomarkers are compared with $(\overline{\alpha}/4)/(2B)$, ..., the next $2^k B$ biomarkers are compared with $(\overline{\alpha}/2^{k+1})/(2^k B)$, and so on.

The motivation of conducting marginal association tests to screen for candidate interaction tests is that we expect a genetic biomarker that has an interaction with the environmental factor for the disease will also show some level of marginal association with the response. This is usually true because: 1) when there is an interaction effect, we may expect a marginal effect in the same direction; 2) if there is an interaction which is not accounted for in a model, e.g. of the form (1.4), then it will appear as a marginal effect. However, it is also possible that the biomarker's main association with response and the interaction effect may be in opposite directions, such that the overall marginal effect cancels out. When this is the case, a marginal screening strategy would fail due to the first stage test statistic having low power.

In the context of gene-environment interaction studies of binary outcomes, an alternative type of screening is testing the correlation between a genetic biomarker and the

environmental factor of interest [53]:

$$logit\{E(e_i \mid X_{ij})\} = \omega_{0_j} + \omega_{X_j} X_{ij}$$

In contrast to a traditional case-only interaction test of the form (1.3), the use of a combination of cases and controls is necessary to preserve the independence between stage 1 and stage 2. Like the case-only analysis, this type of screening requires case-control sampling for a rare response endpoint and gene-environment independence. There exist other related proposals combining this type of screening and marginal association screening tests [30, 38].

To preserve the overall family-wise error rate, a key requirement to apply the two-stage approach is the independence between stage 1 and 2 tests. Both Murcray et al. [53] and Dai et al. [17] proved that: If stage 1 and 2 test statistics are asymptotically independent and m^* denotes the number of stage 1 selected biomarkers, then using a Bonferroni adjusted significance level $\alpha = \overline{\alpha}/m^*$ at stage 2 to test interactions within the reduced set is sufficient to preserve the overall family-wise error rate of the two-stage procedure under $\overline{\alpha}$.

Murcray et al. [53] provided the proof of the between-stage independence for the case-only style gene-environment association screening tests. The proof relied on an analysis of a contingency table for binary response outcomes and the use of the delta method to derive a joint distribution of stage 1 and 2 test statistics from the distribution of table variables. More generally, Dai et al. [17] provided a unified approach to proving the asymptotic between-stage independence by evaluating the covariance matrix between stage 1 and 2 test statistics directly. Specifically, their between-stage independence proof for the marginal association screening tests also applied to generalized linear models. In Appendix A.2 and A.3, I provide re-worked, more detailed derivations of these proofs.

In Chapter 2, I discuss adaption of the two-stage approach for detecting biomarker-treatment interactions in randomized clinical trials. A substantive part of the work in Chapter 2 is proving that the critical between-stage independence assumption could be maintained in a randomized trial setting.

1.3 Multiple testing correction procedures

The traditional interaction analysis typically applies the Bonferroni correction to control the family-wise error rate. This approach can significantly lack power, because this creates a multiple testing problem when there is a large number of biomarkers to be considered. With regard to our interest in high-dimensional interaction testing it is worth considering whether other procedures to declare statistical significance are able to provide

improved efficiency. This section introduces some frequently used procedures for controlling family-wise error rates and false discovery rates when conducting multiple hypothesis tests.

1.3.1 Bonferroni correction

Given the desired overall significance level $\overline{\alpha}$, let the significance level of each test be

$$\alpha = \overline{\alpha}/m$$

where m is the total number of hypothesis tests. Free of dependence and distributional assumptions, we can say the overall family-wise error rate, defined as the probability of at least one type I error, is controlled under $m\alpha = \overline{\alpha}$, which follows from Boole's inequality [32].

A great flexibility the Bonferroni correction [21] provides is that rather than testing each hypothesis at the $\overline{\alpha}/m$ level, the hypotheses may be tested at any other combination of levels that add up to $\overline{\alpha}$, provided that the level of each test is determined before looking at the data. This technique, known as "alpha splitting", in the context of feature selection, can be used to funnel power into features we are more interested in.

However, the correction is criticized for being very stringent when there are a large number of positively correlated tests. This point can be illustrated using the following case: For testing $m = 10,000$ covariates which are completely correlated to each other (one covariate determines all the other covariates), a single hypothesis can only be declared significant at $\alpha = \overline{\alpha}/m = 0.000005$ using a Bonferroni correction to target an overall family-wise error rate under 0.05. With totally correlated tests, the family-wise error rate is actually controlled under 0.000005, instead of the desired 0.05.

1.3.2 Šidák correction

A slightly less conservative correction [66] can be obtained under the assumption of independence between individual tests, by solving the following equation

$$\alpha = 1 - (1 - \overline{\alpha})^{1/m}$$

In the previous case, $\overline{\alpha} = 0.05$ and $m = 10,000$, the Šidák corrected significance level is approximately 0.00000513 which is only slightly less stringent than the Bonferroni correction. The Šidák correction is conservative when tests are positively dependent. In contrast, the correction can be liberal for tests that are negatively dependent.

1.3.3 Holm-Bonferroni method and other family-wise error rate controlling procedures

A more complex step-down[3] procedure [37] which is uniformly more powerful than the Bonferroni correction and controls the family-wise error rate at $\overline{\alpha}$ is described as below:

1. Order the p-values of the m hypotheses in ascending order: $p_{(1)}, p_{(2)}, ..., p_{(m)}$.
2. For a given significance level $\overline{\alpha}$, find the minimal index k such that

$$p_{(k)} > \frac{\overline{\alpha}}{m+1-k}$$

3. Reject the first $k-1$ null hypotheses $H_{(1)}, ..., H_{(k-1)}$ and accept $H_{(k)}, ..., H_{(m)}$.

In a very similar method, the Hochberg procedure [35], rejection of $H_{(1)}, ..., H_{(k)}$ is made after finding the maximal index k such that $p_{(k)} \leq \overline{\alpha}/(m+1-k)$. This method is more powerful than the Holm-Bonferroni procedure, but requires the hypotheses to be independent or under certain forms of positive dependence.

1.3.4 Benjamini-Hochberg method and other false discovery rate controlling procedures

In scenarios when controlling the family-wise error rate is too stringent and not necessary, controlling the false discovery rate, which is defined as the expected proportion of "discoveries" (rejected null hypotheses) that are false, can be a more reasonable goal. Generally, false discovery rate controlling procedures have greater power, at the cost of an increased number of type I errors.

A step-up[4] procedure, the Benjamini-Hochberg method [3], controls the false discovery rate at $\overline{\alpha}$ as described below:

1. Given p-values in ascending order: $p_{(1)}, p_{(2)}, ..., p_{(m)}$, find the largest k such that

$$p_{(k)} \leq \frac{k}{m}\overline{\alpha} \tag{1.5}$$

2. Reject the first k null hypotheses $H_{(1)}, ..., H_{(k)}$ and accept $H_{(k+1)}, ..., H_{(m)}$.

This method is valid when the m tests are independent or positively dependent.

[3]Start with the most significant hypothesis and examine less significant hypotheses in subsequent steps, in comparison with "single-step" methods (e.g. Bonferroni correction) comparing test statistics to their critical values simultaneously.

[4]Start with the least significant hypothesis and examine more significant hypotheses in subsequent steps.

A slightly more complicated Benjamini-Hochberg-Yekutieli procedure [4] allows arbitrary dependence by changing (1.5) to the following condition:

$$p_{(k)} \leq \frac{k}{m \cdot c(m)} \overline{\alpha}$$

where $c(m) = \Sigma_{i=1}^{m} 1/i$.

1.4 Sparse regression methods

One significant drawback of existing two-stage interaction testing procedures is that biomarkers are only tested one at a time, which ignores correlations between biomarkers. Thus, the number and locations of interaction signals may be unclear when there are substantial biomarker-biomarker correlations, since one true effect may be repeatedly represented in the results from testing multiple correlated proxies univariately. In a high-dimensional, low-sample-size data set, a traditional ordinary least squares multivariate regression analysis testing each predictor, while accounting for correlations with the other predictors, is not feasible. Therefore I will consider modern sparse regression methods to model correlated high-dimensional data. There are multiple examples where the use of these techniques has proven useful in the analysis of high-dimensional biomarker data [82, 49, 55].

1.4.1 Ridge regression

Ridge regression [36] is a multivariate analysis method, which applies regularization to avoid overfitting in high-dimensional, low-sample-size problems. Typically, the objective of ridge regression is to minimize a loss function L_n along with an L_2 regularization term:

$$L_n(\boldsymbol{\delta}) + \lambda_n ||\boldsymbol{\delta}||_2^2$$

where $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_m)^T$ is a vector of model coefficients, n is the sample size, m is the total number of covariates and $||\boldsymbol{\delta}||_2^2 = \sum_{j=1}^{m} \delta_j^2$. Through penalizing coefficient magnitudes, ridge regression can achieve a better bias-variance trade-off for high-dimensional problems. Ridge shrinks all the estimated coefficients towards zero, but will not set them exactly to zero. Another property of ridge is the grouping effect, such that strongly correlated covariates tend to have similar estimated coefficients. The optimal value of the regularization parameter λ_n can be chosen using cross-validation based on prediction error (sum of squared errors for linear regression, deviance for logistic regression).

1.4.2 Lasso regression

Lasso [73] was introduced in order to improve the prediction accuracy and interpretability of regression models by altering the model fitting process to select only a subset of the provided covariates for use in the final model rather than using all of them. It introduces L_1 penalties on coefficients, thus only retaining those stronger covariates, to avoid overfitting when facing a high-dimensional, low-sample-size problem. Typically, the objective of lasso is to minimize a loss function along with an L_1 regularization term:

$$L_n(\boldsymbol{\delta}) + \lambda_n ||\boldsymbol{\delta}||_1$$

where $||\boldsymbol{\delta}||_1 = \sum_{j=1}^{m} |\delta_j|$. As compared with ridge regression, which uses a L_2 regularization term, lasso does not only shrink coefficients towards 0, it also forces small coefficients to exactly 0, thus explicitly selecting a simpler model with fewer covariates. This behavior is different from that of ridge, because the L_1 penalty is singular at origin while the L_2 penalty is not. Lasso also does not have a grouping effect like ridge regression. In contrast, it tends to select only one from each group of highly correlated covariates. Fitting a lasso model can be efficiently done through the pathwise coordinate descent method [27], and the optimal value of λ_n can be chosen using cross-validation.

There exist a number of variants in order to overcome the limitations of the original lasso regression. For example, elastic net regularization [87] introduces an additional L_2 penalty, allowing the method to have a grouping effect, which can select highly correlated covariates together. Another generalization, group lasso [84], was proposed to ensure that a group of certain covariates are either included or excluded in the model together. This phenomenon is sometimes useful, e.g. for detecting biomarker-treatment interactions, a biomarker's main effect term and its corresponding interaction term are usually desired to be selected or excluded in the model together [47].

1.5 Adaptive signature design (ASD) for randomized clinical trials

One motivation of finding biomarker-treatment interactions is to identify subgroups of patients who will likely benefit from a new treatment and also predict subgroup treatment effects. In the context of clinical trials, these predictive biomarkers with biomarker-treatment interactions hold great potential to improve trial efficiency [83, 80]. However, there has so far only been limited work on optimal methods for utilizing predictive biomarkers during the course of a randomized trial.

The adaptive signature design (ASD) was proposed as a solution for developing and testing a biomarker signature all within the same trial [25]. The approach employs two

stages: Stage 1 uses a proportion of patients to develop a classifier that can predict whether a patient is more likely to benefit from the new treatment by finding a subset of "sensitive" biomarkers which have significant biomarker-treatment interactions. In stage 2, the established classifier is applied to the remaining patients to identify a "sensitive" subgroup who will likely benefit from the new treatment compared to the control. Since the classification process is carried out after the trial is complete, it neither restricts the entry of patients nor guides the treatment allocation. The final analysis under an ASD tests the treatment effect in both the sensitive subset of patients, as well as in all the patients. The treatment effect is considered efficacious if either of the tests is significant. The ASD is potentially useful, when there exists heterogeneity in the expected treatment effect and no pre-specified biomarker signature is available before the trial. An extension of the ASD achieves better power by employing a cross-validation procedure to make more efficient use of the available data [26]. In Chapter 3, I develop new classifiers within the cross-validated ASD framework to improve trial efficiency. Next, I describe how the cross-validated ASD works in detail.

1.5.1 Cross-validated ASD design

Following the original setting [25], our use of the cross-validated ASD is described in the context of a binary response end point. For the ith patient, we model their binary outcome, R_i, as a function of covariates $X_{i1}, \ldots, X_{im}$ and the treatment assignment variable T_i, using logistic regression. Therefore

$$logit\{E(R_i \mid X_{i1}, \ldots, X_{im}, T_i)\} = \beta_0 + \beta_T T_i + \sum_{j=1}^{m} (\beta_{X_j} X_{ij} + \beta_{X_j \times T} X_{ij} \times T_i) \qquad (1.6)$$

where β_0 is the intercept, β_T is the treatment main effect, β_{X_j} is the jth biomarker's main effect, and the coefficient $\beta_{X_j \times T}$ measures its biomarker-treatment interaction effect. If for some j, $\beta_{X_j \times T}$ is not zero, then a patient's treatment effect depends on their value of the jth biomarker X_j. Those patients, whose individual biomarker profiles mean they respond more positively to the new treatment, form a potentially identifiable "sensitive" subgroup.

In a K-fold cross-validated ASD, a proportion of $(K-1)/K$ of all the patients (development cohort) are used to develop a classifier for identifying the sensitive subgroup (stage 1), then the classifier is applied to the remaining $1/K$ patients (validation cohort) to select a subset of patients who are more likely to benefit from the new treatment (stage 2). This procedure is repeated K times over K non-overlapping validation cohorts (with corresponding development cohorts). At the end, each patient is either selected to the sensitive subgroup or not according to their individual biomarker profiles and the finalized classifier. A test for treatment effect is then carried out within this subgroup, for example

using a Fisher's exact test. Because standard asymptotic theory does not apply when the analysis sample has been defined in this way using cross-validation, as opposed to through random sampling from the underlying population, a permutation method is recommended to obtain a valid p-value [26]. The permutation derived one-sided p-value is defined as:

$$\frac{1 + \text{number of permutations where } T^* \leq T}{1 + \text{number of permutations}}$$

where T is the statistic for the observed data and T^* is the statistic for the data with permuted treatment labels.

Since the adaptive signature inferential procedure is not carried out until the end of the trial, the stage 1 developed classifier neither restricts entry of patients during stage 2 nor has bearing on the randomized treatment allocation. This preserves the design's ability to evaluate the new treatment effect within all the eligible patients. The overall procedure is considered positive if either the test in the whole sample or the subgroup test is significant. The overall type I error rate is controlled by distributing type I error between the two tests. For example, for an overall significance level $\alpha = 0.05$, Freidlin and Simon [25] used $\alpha_1 = 0.04$ for the whole-group test and $\alpha_2 = 0.01$ for the subgroup test.

There are potentially many algorithms for developing the classifier to identify the sensitive subgroup. Freidlin and Simon [25], Freidlin et al. [26] used an approach based on an individual's predicted odds ratio between the treatment and the control arms. The approach proceeds as follows:

1. Using stage 1 data, the following logistic regression model is fit in turn for each biomarker j

$$logit\{E(R_i \mid X_{ij}, T_i)\} = \beta_{0_j} + \beta_{X_j} X_{ij} + \beta_T T_i + \beta_{X_j \times T} X_{ij} \times T_i$$

   The null hypothesis $\beta_{X_j \times T} = 0$ is tested, e.g. using a Wald test with a significance level μ.

2. Classify stage 2 patients based on the m^* biomarkers with significant biomarker-treatment interactions found in stage 1: A patient is designated sensitive if the predicted odds ratio $\hat{pr}(R_i = 1 \mid X_{ij}, T_i = 1)\hat{pr}(R_i = 0 \mid X_{ij}, T_i = 0)/\{\hat{pr}(R_i = 0 \mid X_{ij}, T_i = 1)\hat{pr}(R_i = 1 \mid X_{ij}, T_i = 0)\}$ exceeds a threshold γ for at least G of the m^* biomarkers.

This classification method requires a set of three tuning parameters: μ (the stage 1 biomarker-treatment interaction significance level), γ (the stage 2 odds ratio threshold), and G (the threshold on the number of biomarkers with significant predicted marginal odds ratios which exceed γ). In the cross-validated ASD, tuning parameter selection is

embedded into each loop of the cross-validation using each development cohort. From a list of prespecified sets of each parameter value, the combination of values achieving the smallest subgroup treatment effect test p-value is selected. We will refer to this tuning parameter selection procedure as "inner" cross-validation, in contrast to the "outer" cross-validation procedure which ensures independent patients are used for training the classifier and testing treatment in the sensitive subgroup. This nested cross-validation procedure can be computationally expensive, when the tuning procedure considers a large number of potential sets for the three parameters. Figure 1.1 illustrates this nested cross-validation procedure for one permutation.

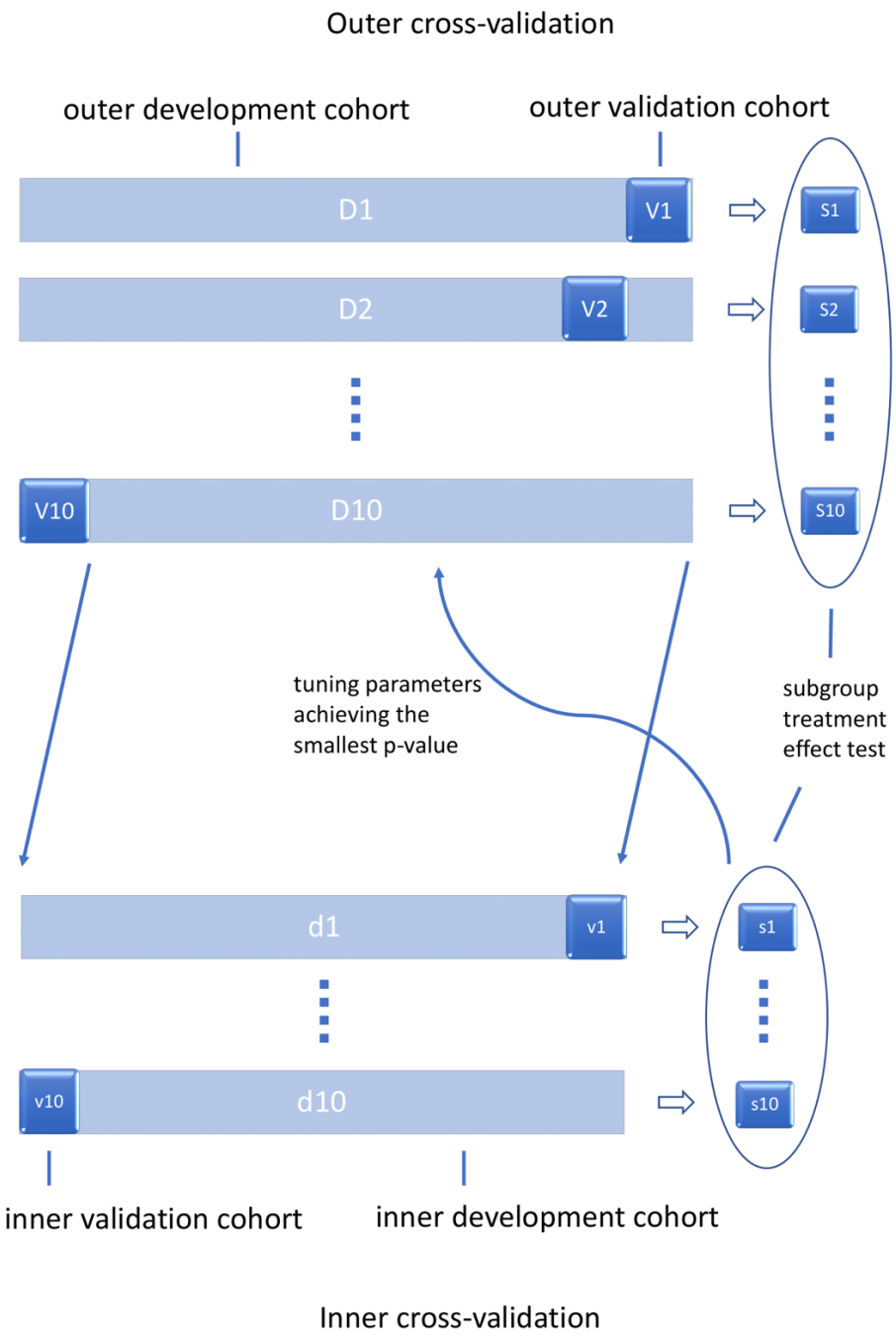

Figure 1.1: Nested cross-validation procedure: Di is the ith outer development cohort; Vi is the ith outer validation cohort; Si is the ith outer selected subset from Vi; di is the ith inner development cohort of an outer development cohort; vi is the ith inner validation cohort of an outer development cohort; si is the ith inner selected subset from vi.

1.6 Trial data sets used as case studies through this thesis

1.6.1 START trial

The first real data set is from the Systematic Therapy of At Risk Teens (START trial) [24], which is composed of 684 participants aged from 11 to 17 with antisocial behavior, half of whom were treated with management as usual (the control arm) and the rest were treated with multisystemic therapy followed by management as usual (the treatment arm). The primary binary outcome of this trial is whether or not a young person was placed in specialist accommodation for young offenders at 18 months post randomization (1 means at home, 0 means out-of-home placement, thus an odds ratio larger than 1 means benefiting more from the new treatment). For this primary outcome, the trial does not show significant difference between the multisystemic therapy group vs the management-as-usual group. A secondary outcome of this trial is a continuous response endpoint, the 18 months' follow-up outcome from Inventory of Callous and Unemotional Traits. The result for this secondary outcome shows multisystemic therapy was detrimental to trial participants compared with management as usual. The findings of this trial do no recommend multisystemic therapy to be used over management as usual in young offenders as the intervention.

In the analysis conducted in the later chapters, I excluded covariates with more than 10% missing data and used mean imputation to replace missing values for covariates with less than 10% missing data. As a result, 75 covariates (demographics, baseline questionnaires, offending history and psychiatric diagnoses) were included in the analysis. A correlation plot is shown in Figure 1.2. There exist several clusters of correlated covariates. The ratio of highly correlated covariates (defined as correlating to at least one another covariate with $\rho \geq 0.6$) to all covariates is 0.324.

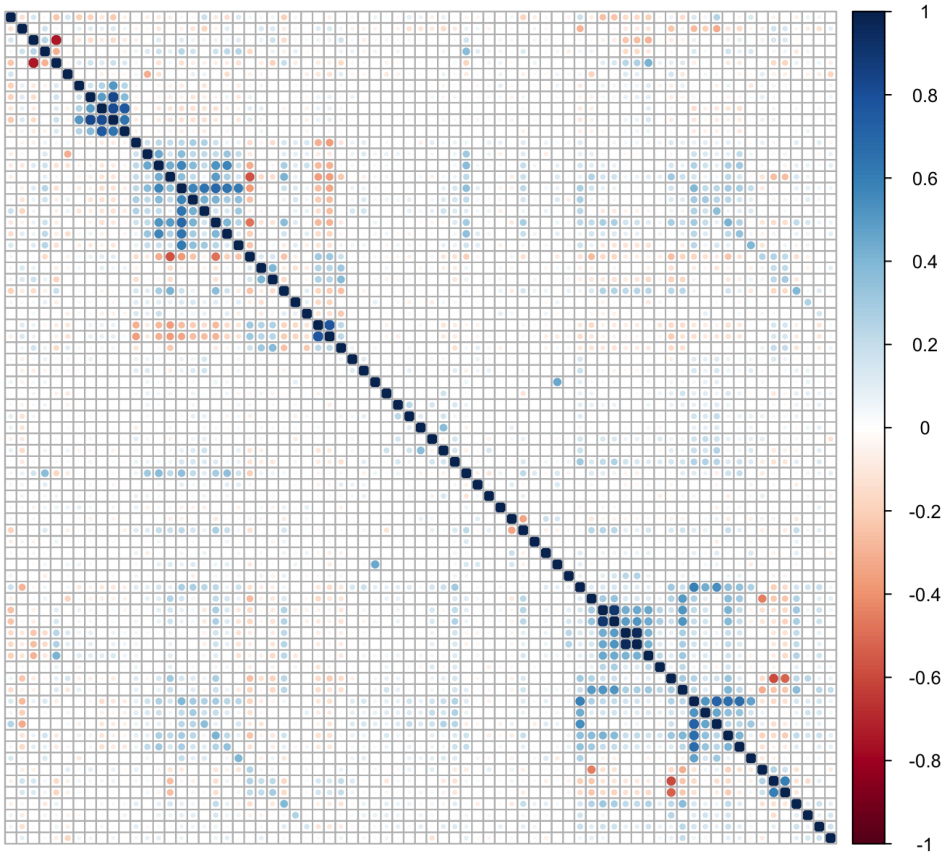

Figure 1.2: Correlation heat map: 75 covariates from the START trial data.

1.6.2 STOPAH trial

The second data set is from a randomized clinical trial evaluating the treatment of alcoholic hepatitis upon steroid response (STOPAH trial) [72]. The data set consists of $1,068$ subjects. In this 2×2 factorial trial, each patient was randomized twice: The first randomization was between with and without prednisolone ($534 : 534$) and the second was between with and without pentoxifylline ($537 : 531$). The 28-day mortality was used as the binary response endpoint. The trial finds: 1) Prednisolone reduced 28-day mortality within trial participants but the association did not reach significance; 2) pentoxifylline did not improve patients' survival at 28 days.

In the analysis conducted in the later chapters, I excluded biomarkers with more than 10% missing data and used mean imputation to replace missing values for biomarkers with less than 10% missing data. As a result, 40 covariates (a small number of which were demographic variables) were included in the analysis conducted in all the later chapters. There exist several clusters of highly correlated covariates, as shown in Figure 1.3. The ratio of highly correlated covariates (defined as correlating to at least one another covariate with $\rho \geq 0.6$) to all covariates is 0.600.

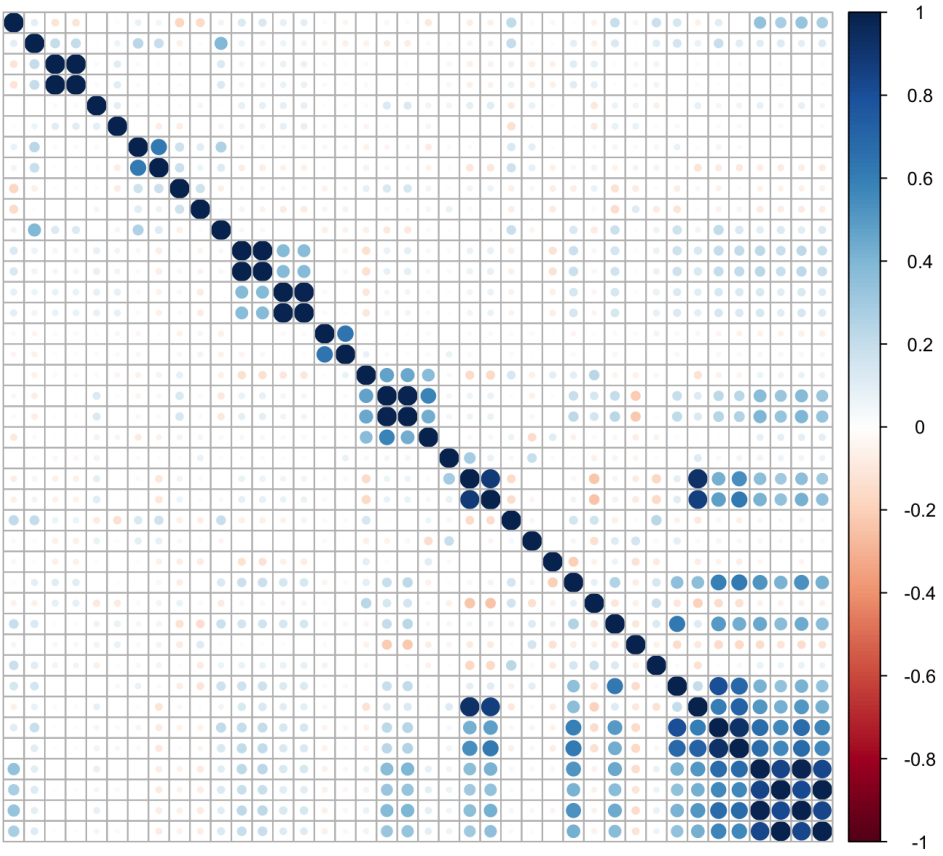

Figure 1.3: Correlation heat map: 40 covariates from the STOPAH trial data.

1.6.3 PREVAIL trial

The third trial data set, which has high-dimensional gene expression biomarkers, is the publicly available PREVAIL trial [54]. The data set is a phase II randomized trial evaluating the efficacy of lactoferrin as a preventative measure for hospital-acquired infections. Gene expression data are available for 61 patients at the National Center for Biotechnology Information (NCBI) website (GSE118657). Of all the 61 patients, 32 were in the lactoferrin group, and the remaining were in the placebo group. The trial did not find significant difference in clinical outcomes between lactoferrin vs placebo groups. In the analyses within this thesis, the Sequential Organ Failure Assessment (SOFA) score measuring change in organ function post-randomization, was used as the continuous response endpoint, and ICU (intensive care unit) mortality was used as the binary response endpoint. From a total of $49,495$ genes, I restricted my analysis to the $10,000$ probes with the highest standard deviations. There exists substantial correlation among these biomarkers. The ratio of highly correlated covariates (defined as correlating to at least one another covariate with $\rho \geq 0.6$) to all covariates is 0.991.

1.7 Thesis overview

This thesis is organized as follows. The focus of Chapter 2 is on improving two-stage approaches for detecting biomarker-treatment interactions. I propose two novel screening tests to be used within a two-stage framework, which utilize ridge and lasso regressions to model correlated high-dimensional data. I provide a proof of asymptotic independence between the stage 1 ridge regression screening and stage 2 standard one-biomarker-at-a-time interaction test statistics. Furthermore, it is shown by simulations and real data applications that my newly proposed methods can provide better performance than traditional one-biomarker-at-a-time approaches for correlated biomarkers. Chapter 2 ends with a proposal to incorporate weighted false discovery rate controlling procedures in the two-stage interaction detecting framework and demonstrates the performance by simulations. Chapter 3 focuses on how to utilize biomarker-treatment interaction information in clinical trial designs. In the ASD (adaptive signature design) framework, I propose two new types of classifiers for selecting sensitive patients who likely benefit from a treatment. I explore both theoretically and through simulations how the two proposed classifiers compare to the existing classifier that is currently used within the ASD. Chapter 4 returns to the topic of detecting biomarker-treatment interactions. It attempts to solve a generic interaction testing issue which is relevant to the work in both Chapter 2 and 3: When testing for interactions under logistic regression for binary outcomes, the interaction effect estimate can be biased when the biomarker is either indirectly or directly associated with the outcome, since this implies a form of model misspecification. This issue can lead to an elevated family-wise error rate. I propose two de-biasing approaches for the standard one-biomarker-at-a-time interaction tests and demonstrate the performance by simulations. Lastly, Chapter 5 concludes the thesis with a summary of my contributions and potential future directions of this research.

Chapter 2

Sparse regression screening procedures in a two-stage interaction detecting framework

2.1 Introduction

High-dimensional biomarkers such as genomics are increasingly being measured in randomized clinical trials. Consequently, there is a growing interest in developing methods that improve the power to detect biomarker-treatment interactions. In the different but related context of gene-environment interaction studies, there is now a significant literature of statistical methods, which exploit aspects of the study design to improve power thus mitigating the multiple testing burden. These include case-only tests [59], empirical Bayes [52], Bayesian model averaging [45], and two-stage tests with different screening procedures [43, 53, 38, 30, 79]. To alleviate the multiple testing burden, two-stage methods use independent information from the data to perform a screening test to select a subset of genetic biomarkers that are more likely to have statistical interactions, and then only test interactions within this reduced set. In the wider context of variable selection, screening strategies have also been explored to focus algorithms on a reduced search space [23, 77]. To our knowledge, all current screening strategies use one-biomarker-at-a-time tests. These univariate screening tests will result in a lot of false positives at stage 1 when there exist substantial correlations between biomarkers. False positives at stage 1 will harm power of the overall two-stage procedure because now stage 2 multiple testing correction needs to account for more biomarkers that have passed stage 1.

In this chapter, I propose two novel screening tests to be used within the two-stage framework for detecting biomarker-treatment interactions, which utilize ridge and lasso regressions to model correlated high-dimensional data at stage 1. I prove these two-stage methods incorporating stage 1 sparse regression screening procedures are able to

preserve the overall family-wise error rate given independence between the treatment and biomarkers. Furthermore, it is shown by simulations and real data applications that the new methods can provide better performance than traditional one-biomarker-at-a-time approaches for correlated biomarkers.

2.2 Lack of applicability of responder-only tests in randomized clinical trials

As previously discussed in Section 1.2.1, the traditional interaction analysis can lack power when the number of biomarkers to be considered is large. A Bonferroni correction is often used for family-wise error rate control. With regard to our interest in high-dimensional interaction testing, it is worth exploring whether other procedures are able to provide improved efficiency. In Appendix A.4, I demonstrate some alternative family-wise error rate controlling methods (Šidák correction [66], Holm-Bonferroni procedure [37] and Hochberg procedure [35]) can only provide a small improvement across the randomized clinical trial settings I consider in this thesis: when the number of biomarkers is large and only a small subset of biomarkers have true interactions with treatment.

In gene-environment studies, case-only tests were proposed to improve power under the assumption of gene-environment independence and the rareness of response events. Section 1.2.2.1 introduced case-only tests in detail. Analogously to the model (1.3), we consider adapting this type of test for detecting biomarker-treatment interactions in randomized clinical trials:

$$logit\{E(T_i \mid X_{ij}, R_i = 1)\} = \gamma_{0_j} + \gamma_{X_j} X_{ij} \tag{2.1}$$

with R_i denoting the binary response outcome, T_i the binary treatment-control indicator, and $X_{i1}, \ldots, X_{im}$ representing the values of m biomarkers, for the ith patient. Generally, randomized clinical trials for rare responses are not conducted, because only very large trials are powered to show a treatment effect for a rare response outcome. Violation of the rare response assumption prevents applying such approaches to finding biomarker-treatment interactions in principle. Secondly, although biomarker-treatment independence is guaranteed in randomized clinical trials, responder-only tests cannot outperform standard interaction tests (1.1), as the trial population represents the entire data set that has been randomized and thus responders (corresponding to "cases") are not "oversampled". To conclude, responder-only tests are not applicable for detecting biomarker-treatment interactions in principle.

2.3 Two-stage interaction tests in randomized clinical trials

To alleviate the multiple testing burden, as discussed in Section 1.2.2.2, two-stage gene-environment interaction testing approaches have been proposed. A two-stage interaction testing approach uses a screening stage to select a subset of biomarkers, and then in stage 2, only tests biomarker-treatment interactions within the reduced set of biomarkers. The multiple correction procedure (e.g. Bonferroni correction) at stage 2 only needs to account for this reduced set of biomarkers, instead of all the biomarkers, thus increasing power. To preserve the overall family-wise error rate of the two-stage approach, both Murcray et al. [53] and Dai et al. [17] proved that the stage 1 screening test statistic needs to be (asymptotically) independent of the stage 2 test statistic. Further detail of existing two-stage approaches for detecting gene-environment interactions was given in Section 1.2.2.2. We now consider in more detail, the applicability of this framework for detecting biomarker-treatment interactions in data from randomized clinical trials.

A marginal association test of the form (1.4), testing the marginal effect of a biomarker on the outcome, is typically used as the stage 1 screening test. This type of screening test can readily be used when applying this framework to detect biomarker-treatment interactions in randomized clinical trials. By using the marginal association screening test, we expect that a biomarker's marginal association with the response is informative for the existence of a biomarker-treatment interaction.

An alternative type of screening that has been proposed in the gene-environment testing literature, is to test the correlation between a genetic biomarker and the environmental factor of interest. This has been discussed in detail in Section 1.2.2.2. If applied to detect biomarker-treatment interactions, this type of screening test corresponds to the model of the following form:

$$logit\{E(T_i \mid X_{ij})\} = \omega_{0_j} + \omega_{X_j} X_{ij}$$

However, such a screening procedure is not generally applicable in randomized clinical trials, where the rare response condition does not hold and the trial population represents the entire data set thus responders (cases) are not "oversampled". Indeed, biomarker-treatment independence induced by randomization dictates $\omega_{X_j} = 0$ in the whole sample. The stage 1 test statistic is not informative for the interaction parameter being non-zero at all.

Thus, I recommend the marginal screening test of the form (1.4), i.e. testing for associations between biomarkers and the outcome, to be used in the two-stage interaction testing framework for detecting biomarker-treatment interactions in randomized clinical

trials.

2.4 New stage 1 sparse regression screening procedures accounting for biomarker-biomarker correlations

A limitation in the existing literature on two-stage interaction testing approaches is that biomarkers are tested one at a time. When there are substantial correlations between biomarkers, stage 1 one-biomarker-at-a-time tests can result in a lot of false positives. This reduces power of two-stage interaction testing approaches, because false positives from the stage 1 screening tests have a negative impact on the multiple testing adjustment at stage 2 interacting testing.

To account for biomarker-biomarker correlations, I propose new stage 1 multivariate screening tests of the following form

$$G\{E(Y_i \mid X_{i1}, \ldots, X_{im})\} = \delta_0 + \delta_T T_i + \sum_{j=1}^{m} \delta_{X_j} X_{ij} \tag{2.2}$$

This multivariate version of the marginal association screening test also includes the treatment main effect term. This is necessary to preserve the independence between stage 1 screening and stage 2 interaction tests as described later. When $n < m$, as will usually be the case with high-dimensional biomarker data, fitting this multivariate model directly is infeasible. Thus, penalized regression methods are introduced to address this issue.

2.4.1 Ridge regression screening

To fit the above multivariate model, I use ridge regression, with a L_2 regularization term $||\boldsymbol{\delta}||_2^2 = \delta_T^2 + \sum_{j=1}^{m} \delta_{X_j}^2$ to fit the model (2.2). Ridge regression was introduced in detail in Section 1.4.1 of the introduction chapter.

Ridge shrinks all the estimated coefficients towards zero, but will not set them exactly to zero. For use in a two-stage interaction testing strategy, I propose ordering the biomarkers based on the ridge coefficients obtained from stage 1, and then use the resulting ranking to determine varying significance thresholds across buckets of markers during stage 2 one-at-a-time interaction tests according to the weighting scheme [39]: the B most significant biomarkers, i.e. with smallest estimated ridge coefficients in stage 1, are compared with an adjusted significance level $(\overline{\alpha}/2)/B$, the next $2B$ biomarkers are compared with $(\overline{\alpha}/4)/(2B)$, ..., the next $2^k B$ biomarkers are compared with $(\overline{\alpha}/2^{k+1})/(2^k B)$, and so on.

The L_2 penalty is a smooth function of $\boldsymbol{\delta}$ at the origin, which allows us to prove the

between-stage independence as explained in Section 2.5. However, It is known that ridge regression has a tendency to average effects across strongly correlated covariates. This phenomenon is not desirable for a screening strategy since it could inflate the number of non-interacting biomarkers being put forward to stage 2.

2.4.2 Lasso regression screening

Lasso is an alternative method to fit the model (2.2) with a L_1 regularization term $||\boldsymbol{\delta}||_1 = |\delta_T| + \sum_{j=1}^{m} |\delta_{X_j}|$. Lasso selects a subset of biomarkers which can be next tested by the stage 2 interaction testing procedure.

Lasso does not exhibit the grouping effect, which in principle makes it more desirable for a screening strategy. However, as lasso uses a L_1 penalty which is not a smooth function, it is challenging to prove the between-stage independence requirement to preserve the overall family-wise error rate in two-stage approaches using current asymptotic theory.

2.5 Asymptotic between-stage independence for stage 1 sparse regression screening

In this section, I show that independence between stage 1 and stage 2 test statistics holds for stage 1 ridge regression screening tests.

For the ith subject, let Y_i denote the outcome variable, $\boldsymbol{X}_i = (T_i, X_{i1}, \ldots, X_{im})^T$ be a vector of the binary treatment-control indicator and the values of m biomarkers. Consider the proposed stage 1 marginal association screening test based on the multivariate model of the form

$$G\{E(Y_i \mid \boldsymbol{X}_i)\} = \boldsymbol{X}_i^T \boldsymbol{\delta}$$

where $\boldsymbol{\delta} = (\delta_T, \delta_{X_1}, \ldots, \delta_{X_m})^T$ and G is a canonical link function. The model underlying the stage 2 standard one-biomarker-at-a-time interaction test is of the form

$$G\{E(Y_i \mid \boldsymbol{V}_{ij})\} = \boldsymbol{V}_{ij}^T \boldsymbol{\beta}_j \quad (j = 1, \ldots, m)$$

where $\boldsymbol{V}_{ij} = (X_{ij}, T_i, X_{ij}T_i)^T$ and $\boldsymbol{\beta}_j = (\beta_{X_j}, \beta_{T_j}, \beta_{X_j \times T})^T$. The above forms ignore intercepts without loss of generality. Homogeneity of variance is assumed, i.e. $var(Y_i \mid \boldsymbol{X}_i)$ and $var(Y_i \mid \boldsymbol{V}_{ij})$ are constants. I first show the property of between-stage asymptotic independence for the stage 1 multivariate regression marginal association estimator without regularization.

Theorem 2.5.1. *For any $j = 1, \ldots, m$, if X_{ij} is independent of T_i, and, $E(T_i) = 0$ or $E(X_{ij}) = 0$ (i.e. T_i or X_{ij} is centered around 0), then under the null hypothesis $\beta_{X_j \times T} = 0$,*

$$cov\{n^{1/2}(\hat{\delta}^0_{X_j} - \delta_{X_j}), n^{1/2}(\hat{\beta}_{X_j \times T} - \beta_{X_j \times T})\} \xrightarrow{p} 0$$

where $\hat{\delta}^0_{X_j}$ and $\hat{\beta}_{X_j \times T}$ are the maximum likelihood estimators for unknown parameters δ_{X_j} and $\beta_{X_j \times T}$ respectively without regularization (i.e. $\lambda_n = 0$).

Proof. Based on the unified approach to proving the between-stage asymptotic independence by Dai et al. [17], we need to evaluate the covariance matrix $\boldsymbol{A}_1^{-1}\boldsymbol{B}\boldsymbol{A}_2^{-1}$, where

$$\begin{aligned}
\boldsymbol{A}_1 &= E[(\boldsymbol{X}_i\boldsymbol{X}_i^T)\{Y_i - E(Y_i \mid \boldsymbol{X}_i)\}^2] \\
\boldsymbol{B} &= E[(\boldsymbol{X}_i\boldsymbol{V}_{ij}^T)\{Y_i - E(Y_i \mid \boldsymbol{X}_i)\}\{Y_i - E(Y_i \mid \boldsymbol{V}_{ij})\}] \\
\boldsymbol{A}_2 &= E[(\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T)\{Y_i - E(Y_i \mid \boldsymbol{V}_{ij})\}^2]
\end{aligned}$$

Of this $(m+1) \times 3$ matrix product $\boldsymbol{A}_1^{-1}\boldsymbol{B}\boldsymbol{A}_2^{-1}$, the $(j+1)$th element of the last column is the value which $cov\{n^{1/2}(\hat{\delta}^0_{X_j} - \delta_{X_j}), n^{1/2}(\hat{\beta}_{X_j \times T} - \beta_{X_j \times T})\}$ converges to in probability. We need to show this limiting value is zero. In fact, I am able to prove a stronger result that the last column of $\boldsymbol{A}_1^{-1}\boldsymbol{B}\boldsymbol{A}_2^{-1}$ are all zeros as follows.

We simplify the expression of $\boldsymbol{B}$ as

$$\begin{aligned}
\boldsymbol{B} &= E[(\boldsymbol{X}_i\boldsymbol{V}_{ij}^T)\{Y_i^2 - Y_iE(Y_i \mid \boldsymbol{X}_i) - Y_iE(Y_i \mid \boldsymbol{V}_{ij}) + E(Y_i \mid \boldsymbol{X}_i)E(Y_i \mid \boldsymbol{V}_{ij})\}] \\
&= E[(\boldsymbol{X}_i\boldsymbol{V}_{ij}^T)E\{Y_i^2 - Y_iE(Y_i \mid \boldsymbol{X}_i) - Y_iE(Y_i \mid \boldsymbol{V}_{ij}) + E(Y_i \mid \boldsymbol{X}_i)E(Y_i \mid \boldsymbol{V}_{ij}) \mid \boldsymbol{X}_i\}] \\
&= E(\boldsymbol{X}_i\boldsymbol{V}_{ij}^T)var(Y_i \mid \boldsymbol{X}_i)
\end{aligned}$$

which uses the law of iterated expectations, the fact that $\boldsymbol{X}_i$ includes $\boldsymbol{V}_{ij}$ under the null hypothesis $\beta_{X_j \times T} = 0$, and assumes homogeneity of variance, i.e. $var(Y_i \mid \boldsymbol{X}_i)$ is a constant.

Similarly, we have $\boldsymbol{A}_1 = E(\boldsymbol{X}_i\boldsymbol{X}_i^T)var(Y_i \mid \boldsymbol{X}_i)$ and $\boldsymbol{A}_2 = E(\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T)var(Y_i \mid \boldsymbol{V}_{ij})$. Thus,

$$\boldsymbol{A}_1^{-1}\boldsymbol{B}\boldsymbol{A}_2^{-1} \propto E(\boldsymbol{X}_i\boldsymbol{X}_i^T)^{-1}E(\boldsymbol{X}_i\boldsymbol{V}_{ij}^T)E(\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T)^{-1} \tag{2.3}$$

I will show that the last column of the right hand side of (2.3) above are all zeros, by first

considering the second and the third terms, which may be expressed as:

$$
\begin{aligned}
\underset{(m+1)\times 3}{E(\boldsymbol{X}_i\boldsymbol{V}_{ij}^T)} &= \begin{Bmatrix} E(T_iX_{ij}) & E(T_i^2) & E(T_i^2X_{ij}) \\ E(X_{i1}X_{ij}) & E(T_iX_{i1}) & E(T_iX_{i1}X_{ij}) \\ \vdots & \vdots & \vdots \\ E(X_{im}X_{ij}) & E(T_iX_{im}) & E(T_iX_{im}X_{ij}) \end{Bmatrix} \\
\underset{3\times 3}{E(\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T)^{-1}} &= \begin{Bmatrix} E(X_{ij}^2) & E(T_iX_{ij}) & E(T_iX_{ij}^2) \\ E(T_iX_{ij}) & E(T_i^2) & E(T_i^2X_{ij}) \\ E(T_iX_{ij}^2) & E(T_i^2X_{ij}) & E(T_i^2X_{ij}^2) \end{Bmatrix}^{-1} \\
&= \frac{1}{det\{E(\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T)\}} \begin{Bmatrix} \cdot & \cdot & E(T_iX_{ij})E(T_i^2X_{ij}) - E(T_i^2)E(T_iX_{ij}^2) \\ \cdot & \cdot & E(T_iX_{ij})E(T_iX_{ij}^2) - E(X_{ij}^2)E(T_i^2X_{ij}) \\ \cdot & \cdot & E(X_{ij}^2)E(T_i^2) - E(T_iX_{ij})^2 \end{Bmatrix}
\end{aligned}
$$

Thus, for the $(m+1)\times 3$ matrix product of these terms $E(\boldsymbol{X}_i\boldsymbol{V}_{ij}^T)E(\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T)^{-1}$, the $(k+1)$th element $(k=1,\ldots,m)$ of the last column is

$$
\begin{aligned}
& \frac{1}{det\{E(\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T)\}} \left\{ E(X_{ik}X_{ij}),\quad E(T_iX_{ik}),\quad E(T_iX_{ik}X_{ij}) \right\} \\
& \cdot \begin{Bmatrix} E(T_iX_{ij})E(T_i^2X_{ij}) - E(T_i^2)E(T_iX_{ij}^2) \\ E(T_iX_{ij})E(T_iX_{ij}^2) - E(X_{ij}^2)E(T_i^2X_{ij}) \\ E(X_{ij}^2)E(T_i^2) - E(T_iX_{ij})^2 \end{Bmatrix} \\
= & \frac{1}{det\{E(\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T)\}} E(T_i)var(T_i)E(X_{ij})\{E(X_{ik}X_{ij})E(X_{ij}) - E(X_{ik})E(X_{ij}^2)\} \\
= & \ 0
\end{aligned}
$$

The result above uses the independence between T_i and X_{ij}, and the assumption $E(T_i)=0$ or $E(X_{ij})=0$. Similarly, the first element of the last column is also zero.

Premultiplying $E(\boldsymbol{X}_i\boldsymbol{V}_{ij}^T)E(\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T)^{-1}$ by $E(\boldsymbol{X}_i\boldsymbol{X}_i^T)^{-1}$ completes the right side of (2.3). Since the last column of $E(\boldsymbol{X}_i\boldsymbol{V}_{ij}^T)E(\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T)^{-1}$ are all zeros, the last column of $E(\boldsymbol{X}_i\boldsymbol{X}_i^T)^{-1}E(\boldsymbol{X}_i\boldsymbol{V}_{ij}^T)E(\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T)^{-1}$ in (2.3) are also all zeros. Thus, I have proved that $cov\{n^{1/2}(\hat{\delta}^0_{X_j}-\delta_{X_j}), n^{1/2}(\hat{\beta}_{X_j\times T}-\beta_{X_j\times T})\}$ converges to zero in probability and this result holds for any $j=1,\ldots,m$. □

Previous work [17] has demonstrated that stage 1 univariate marginal association screening tests are independent with the stage 2 one-biomarker-at-a-time interaction tests. Theorem 2.5.1 extends this to show independence still holds when stage 1 tests are extended to a multivariate regression. My proof relies on: 1) the inclusion of the treatment main effect in the multivariate regression of the form (2.2); 2) an assumption of independence between the treatment assignment and biomarker values, which is valid in the context of a

randomized clinical trial; and 3) homogeneity of variance, which is a common assumption for continuous outcomes in linear regression.

Next I establish the asymptotically linear form of the ridge estimator to show it only differs from the non-penalized estimator in a constant.

Lemma 2.5.2. *Under standard regularity conditions [75, p. 51-52] and if* $\lambda_n = O(n^{1/2})$, *i.e.* $\lim_{n\to\infty} \lambda_n/n^{1/2} = \lambda_0 \geq 0$, *then*

$$n^{1/2}(\hat{\boldsymbol{\delta}}^{\lambda} - \boldsymbol{\delta}) \to \mathcal{N}(-2\lambda_0\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}, \sigma^2\boldsymbol{\Sigma}^{-1})$$

in distribution, where $\hat{\boldsymbol{\delta}}^{\lambda}$ *is the ridge estimator,* $\mathcal{N}$ *is a normal distribution,* σ *and* Σ *are a constant and an invertible constant matrix.*

Based on the asymptotic distributions derived in Lemma 2.5.2 above and Theorem 2.5.1, I am able to prove the asymptotic independence between the stage 1 ridge marginal association screening estimator and the stage 2 one-at-a-time interaction estimator in the following corollary.

Corollary 2.5.2.1. *For any* $j = 1, \ldots, m$, *if* X_{ij} *is independent of* T_i, *and,* $E(T_i) = 0$ *or* $E(X_{ij}) = 0$ *(i.e.* T_i *or* X_{ij} *is centered around* 0*), then under the null hypothesis* $\beta_{X_j \times T} = 0$,

$$cov\{n^{1/2}(\hat{\delta}^{\lambda}_{X_j} - \delta_{X_j}), n^{1/2}(\hat{\beta}_{X_j \times T} - \beta_{X_j \times T})\} \xrightarrow{p} 0$$

where $\hat{\delta}^{\lambda}_{X_j}$ *is the maximum likelihood estimator with the ridge penalty.*

Proofs of Lemma 2.5.2 and Corollary 2.5.2.1 are given in Appendix A.5. In addition, a sketch proof for lasso regression is also provided in Appendix A.5, which shows the asymptotic independence between the stage 1 lasso marginal association screening selector and the stage 2 one-at-a-time interaction estimator.

2.6 Weighted false discovery rate controlling procedures in a two-stage interaction detecting framework

The existing two-stage interaction testing literature has focused on controlling family-wise error rates. Sometimes, in the context of predictive biomarker detection in randomized clinical trials, controlling the family-wise error rate is too stringent and not necessary. Decisions around approving a new drug require strict error control. However, when prediction or pure exploration is the end goal of predictive biomarker discovery, false positives have less direct impact on patient safety. In these cases, controlling the false

discovery rate, which is defined as the expected proportion of "discoveries" (rejected null hypotheses) that are false, can be a more reasonable goal. Generally, false discovery rate controlling procedures lead to greater power, at the cost of an increased number of type I errors.

Thus, I propose incorporating the weighted false discovery rate controlling procedures provided by Ramdas et al. [61] into the two-stage framework. The algorithm works as below:

1. For all m biomarkers, define positive prior weights $\{w_j\}_{j=1}^m$ with the normalization condition $\sum_{j=1}^m w_j = m$, and assign weights based on the m^* biomarkers which have passed stage 1 screening: If the jth biomarker passed the screening, assign $w_j = m/m^*$, otherwise, $w_j = 0$. An alternative way is to use ordered biomarkers based on stage 1 p-values and assign weights like the weighted hypothesis testing described previously: B most significant biomarkers with weights $m/2/B$, $2B$ with $m/4/2B$, and so on.

2. Given p-values $\{p_j\}_{j=1}^m$ from stage 2 interaction tests, make adjustments using prior weights $\bar{p}_j = p_j/w_j$.

3. Use a false discovery rate controlling procedure, e.g. the Benjamini-Hochberg (BH) procedure [3], on the adjusted p-values $\{\bar{p}_j\}_{j=1}^m$, to control the overall false discovery rate.

Similarly to controlling family-wise error rates in two-stage approaches, the weighted false discovery rate controlling procedures also require independence between stage 1 and stage 2 tests {Proposition 2 and Lemma 1(b) in Ramdas et al. [61]}.

2.7 Simulation studies

To evaluate performance of my proposed biomarker-treatment interaction testing procedures described above, I generated simulated data sets, with $m = 1,000$ biomarkers each. Data were simulated under the model $Y_i = \beta_0 + \beta_T T_i + \sum_{j=1}^m (\beta_{X_j} X_{ij} + \beta_{X_j \times T} X_{ij} \times T_i) + \varepsilon_i$, where the treatment main effect was set to $\beta_T = 0.5$ and the intercept to $\beta_0 = 0$. All $1,000$ biomarkers were partitioned into 50 clusters of correlated biomarkers, containing 20 biomarkers each. We denote the clusters $C_1 = \{X_1, \ldots, X_{20}\}$, $C_2 = \{X_{21}, \ldots, X_{40}\}$, and so on. One biomarker in the first cluster was ascribed a main effect and an interaction effect, i.e. $\beta_{X_1} = 0.5$ and $\beta_{X_1 \times T} = 1$. Four other biomarkers in four other different clusters were ascribed main effects on the trait without interactions, i.e. $\beta_{X_{21}} = \beta_{X_{41}} = \beta_{X_{61}} = \beta_{X_{81}} = 1.5$. All other biomarkers do not have direct effects on the outcome. Each biomarker X_j was generated from a standard normal distribution $\mathcal{N}(0, 1)$ and the binary treatment assignment

was drawn from a $Bernoulli(0.5)$ distribution, while ε_i was generated from a normal distribution with standard deviation 5. The residual standard deviation was chosen such that the proportion of variance explained by the true model is 0.292, which is realistic for various biomarkers and traits. I considered two types of correlation patterns among biomarkers: 1) The 20 biomarkers within each cluster are correlated with each other ($\rho = 0.6$), but there are no correlations between biomarkers in different clusters; 2) all biomarkers are independent of one another ($\rho = 0$). For each scenario, $1,000$ replicate data sets were generated to estimate power and family-wise error rates. Power for all the approaches listed below is defined according to the idea of "cluster discoveries" in Brzyski et al. [8] as pr(reject at least one H_0^j for any $X_j \in C_i \mid$ at least one H_1^k is true for any $X_k \in C_i$), where H_0^j is the null hypothesis for X_j and H_1^k is the alternative hypothesis for X_k.

Five different screening procedures were compared:

1. "Univariate screening (threshold)": A selection of biomarkers to take forward to stage 2 was based on significance in a regression of response on the biomarkers one at a time, of the form (1.4). A significance level $\alpha_1 = 0.05$ was used without adjustment for each stage 1 biomarker test.

2. "Univariate screening (rank)": All biomarkers were taken forward to stage 2, and the stage 1 p-value ranking was used to conduct a stage 2 weighted hypothesis test described in Section 1.2.2.2 with $B = 5$ (a number recommended by [30]).

3. "Ridge screening (rank)": Ridge regression was used to estimate marginal effects at stage 1. Then all biomarkers were ordered based on these stage 1 coefficients and the rank would be used by the stage 2 weighted hypothesis test with $B = 5$. The optimal λ_n was chosen to minimize predictive errors under 5-fold cross-validation. The R package **glmnet** [27] was used.

4. "Lasso screening": A selection of biomarkers was based on the result from a lasso multivariate regression for estimating marginal effects. The optimal λ_n was chosen to minimize predictive errors under 5-fold cross-validation. The R package **glmnet** was used.

5. "No screening": A standard single-step interaction test of the form (1.1), targeting an overall family-wise error rate $\overline{\alpha} = 0.05$, was performed as a baseline comparator with a Bonferroni correction applied with $m = 1,000$.

The standard interaction tests were also performed as the stage 2 tests for all the four two-stage approaches (1 - 4) described above.

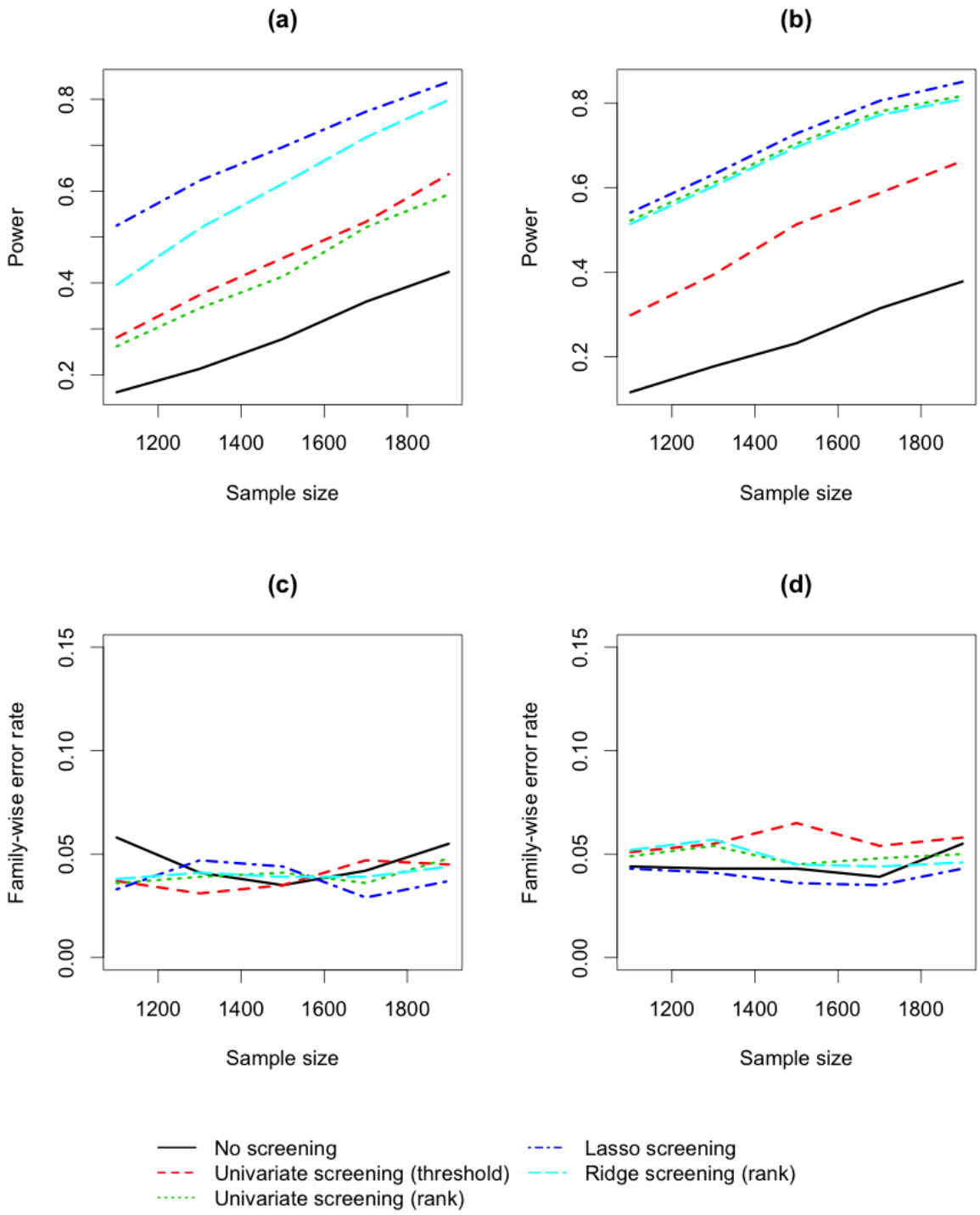

Figure 2.1: Comparison of two-stage interaction tests with different screening procedures in simulated data. The four panels represent: (a) power, highly correlated biomarkers ($\rho = 0.6$), (b) power, independent biomarkers ($\rho = 0$), (c) family-wise error rate, highly correlated biomarkers ($\rho = 0.6$), (d) family-wise error rate, independent biomarkers ($\rho = 0$).

In Figure 2.1(a), with highly correlated biomarkers, the ridge and lasso screening pro-

cedures demonstrated substantially higher power than the univariate screening procedures, showing a clear benefit of accounting for correlations between the biomarkers at stage 1. For the univariate screening procedures, the five biomarkers with true marginal signals and all the biomarkers associated with them, including $X_1, \ldots, X_{100}$, were likely to be retained after screening in the "threshold" approach or land into the top buckets at stage 2 in the "rank" approach. In contrast, the sparse regression screening procedures considered the effect of each biomarker, adjusted for all other biomarkers, and therefore tended to ascribe less evidence to biomarkers whose marginal associations were exaggerated by correlation with the true signal(s). Thus, much fewer biomarkers (five including X_1, X_{21}, X_{41}, X_{61}, X_{81}), i.e. the true simulated marginal signals, tended to pass the screening or land in the top buckets at stage 2. This enhanced the power of the overall two-stage approach compared with using the univariate screening procedures, because the multiple testing correction at stage 2 funneled more power into each "promising" biomarker which has passed stage 1 screening or landed into the top buckets. In Figure 2.1(b), with independent biomarkers, where the multivariate regression is not required for unbiased effect estimation at stage 1, the univariate screening using weighted hypothesis tests and the sparse regression screening procedures performed similarly. All four two-stage tests outperformed the single-step interaction test by providing better power at the same family-wise error rate level. Figure 2.1(c) and (d) showed that all five procedures controlled the family-wise error rates around the desired level of 0.05, in either the scenario with highly correlated biomarkers or the scenario with independent biomarkers.

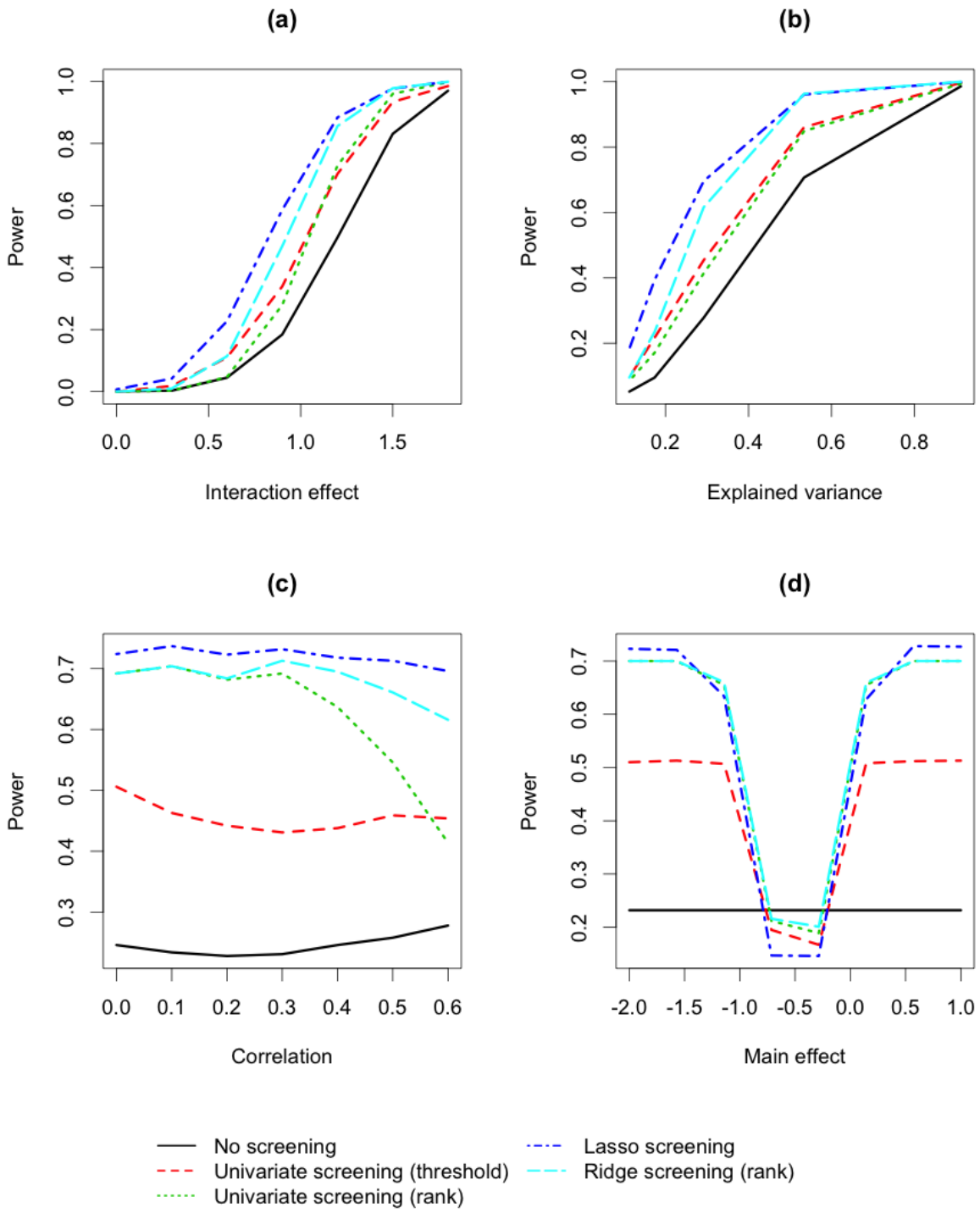

Figure 2.2: Comparison of two-stage interaction tests with different screening procedures in simulated data. The four panels represent: (a) highly correlated biomarkers ($\rho = 0.6$), changing the interaction effect of the interacting biomarker $\beta_{X_1 \times T}$, (b) highly correlated biomarkers ($\rho = 0.6$), changing the standard deviation of the normal distribution from which ε_i was drawn and consequently the variance explained in the outcome, (c) changing the biomarker-biomarker correlation, (d) independent biomarkers ($\rho = 0$), changing the main effect of the interacting biomarker β_{X_1}.

In Figure 2.2(a), I used the base scenario with one biomarker having an interaction (biomarker-biomarker correlation $\rho = 0.6$, sample size of $1,500$) described above, and changed only the interaction effect of the interacting biomarker $\beta_{X_1 \times T}$. Figure 2.2(a) showed that when the true interaction effect was too small, all the procedures failed to identify the interaction. When the interaction effect was large enough, all the procedures were able to find the interaction. In the wide spectrum between the two extremes, the sparse regression screening strategies performed consistently the best among these methods, followed by the two univariate screening procedures.

In Figure 2.2(b), we compared power of the different screening strategies while varying the proportion of explained variation by the true model. Specifically, I changed the standard deviation of the normal distribution from which ε_i was drawn from. For this exploration, biomarkers were set to be correlated at 0.6 and the sample size to $1,500$. Figure 2.2(b) showed that when the true model explained either a low or high proportion of the variance, all the methods tended to perform similarly to each other. Again, in the wide spectrum between the two extremes, the comparison was rather consistent: The sparse regression screening strategies performed best, followed by the two univariate screening procedures, with the single-step interaction test always resulting in the lowest power.

In Figure 2.2(c), I changed only the correlation among biomarkers to examine how it would affect the power comparison of these screening procedures. The sample size was fixed at $1,500$. It is shown that with the increasing correlation, power of the univariate screening procedure (rank) decreases and the benefit using the sparse regression screening strategies increases. This phenomenon is consistent with our observation from the base scenarios shown in Figure 2.1, as previously explained.

In Figure 2.2(d), I simulated scenarios with one biomarker having an interaction, no correlations among the biomarkers, and changed only the main effect of the interacting biomarker β_{X_1}, i.e. main effects of the other four biomarkers were the same as the base scenario. The sample size was fixed at $1,500$. Figure 2.2(d) reveals that there are some special cases where all two-stage approaches give lower power than a standard single-step interaction test. In the cases where power of the two-stage approaches was lower, the main and interaction effect parameters were in opposite directions, which reduces the strength of the marginal association for true interactions.

Next, I examine the weighted false discovery rate controlling procedures (described in Section 2.6), which I incorporated into the two-stage approaches. This was done for the two base simulation scenarios: the one with highly correlated biomarkers and the other with independent biomarkers. Shown in Table 2.1 and Table 2.2, all the weighted BH procedures correctly achieved false discovery rates controlled around 0.05. As expected, they provided increased power but incurred a higher family-wise error rate compared with corresponding family-wise error rate controlling procedures.

Table 2.1: Comparison of two-stage tests with different screening procedures in simulated data. ($\rho = 0.6$, sample size of $1,500$)

Family-wise error rate controlling method applied at stage 2			
	Power	Family-wise error rate	False discovery rate
No screening	0.278	0.035	0.0300
Univariate screening (threshold)	0.454	0.035	0.0234
Univariate screening (rank)	0.414	0.041	0.0328
Lasso screening	0.696	0.044	0.0278
Ridge screening (rank)	0.616	0.039	0.0266
False discovery rate controlling method applied at stage 2			
	Power	Family-wise error rate	False discovery rate
No screening	0.287	0.078	0.0424
Univariate screening (threshold)	0.479	0.124	0.0361
Univariate screening (rank)	0.424	0.078	0.0452
Lasso screening	0.700	0.079	0.0423
Ridge screening (rank)	0.622	0.077	0.0381

Table 2.2: Comparison of two-stage tests with different screening procedures in simulated data. ($\rho = 0$, sample size of $1,500$)

Family-wise error rate controlling method applied at stage 2			
	Power	Family-wise error rate	False discovery rate
No screening	0.232	0.043	0.0400
Univariate screening (threshold)	0.513	0.065	0.0495
Univariate screening (rank)	0.700	0.045	0.0295
Lasso screening	0.728	0.036	0.0220
Ridge screening (rank)	0.700	0.045	0.0307
False discovery rate controlling method applied at stage 2			
	Power	Family-wise error rate	False discovery rate
No screening	0.236	0.051	0.0430
Univariate screening (threshold)	0.529	0.101	0.0640
Univariate screening (rank)	0.706	0.082	0.0475
Lasso screening	0.732	0.061	0.0337
Ridge screening (rank)	0.706	0.089	0.0517

Lastly, I simulated a high-dimensional scenario ($n < m$), where each data set has $5,000$ highly correlated biomarkers. All $5,000$ biomarkers were partitioned into 250 clusters of biomarkers, containing 20 biomarkers each. The 20 biomarkers within each cluster are

correlated with each other ($\rho = 0.6$), but there are no correlations between biomarkers in different clusters. One biomarker in the first cluster was ascribed a main effect and an interaction effect, i.e. $\beta_{X_1} = 0.5$ and $\beta_{X_1 \times T} = 1$. Four other biomarkers in four other different clusters were ascribed main effects on the trait without interactions, i.e. $\beta_{X_{21}} = \beta_{X_{41}} = \beta_{X_{61}} = \beta_{X_{81}} = 1.5$. All other biomarkers do not have direct effects on the outcome.

Figure 2.3(a) shows that the ridge and lasso screening procedures accounting for biomarker-biomarker correlations at stage 1 still demonstrated substantially higher power than the univariate screening procedures. Lasso screening was more powerful than ridge. This is because the weighting strategy used by the "rank" procedure of ridge screening is more conservative than lasso screening which explicitly selects a subset of biomarkers for stage 2 testing. Figure 2.3(b) shows that family-wise error rates of all the procedures were controlled around 0.05 as expected.

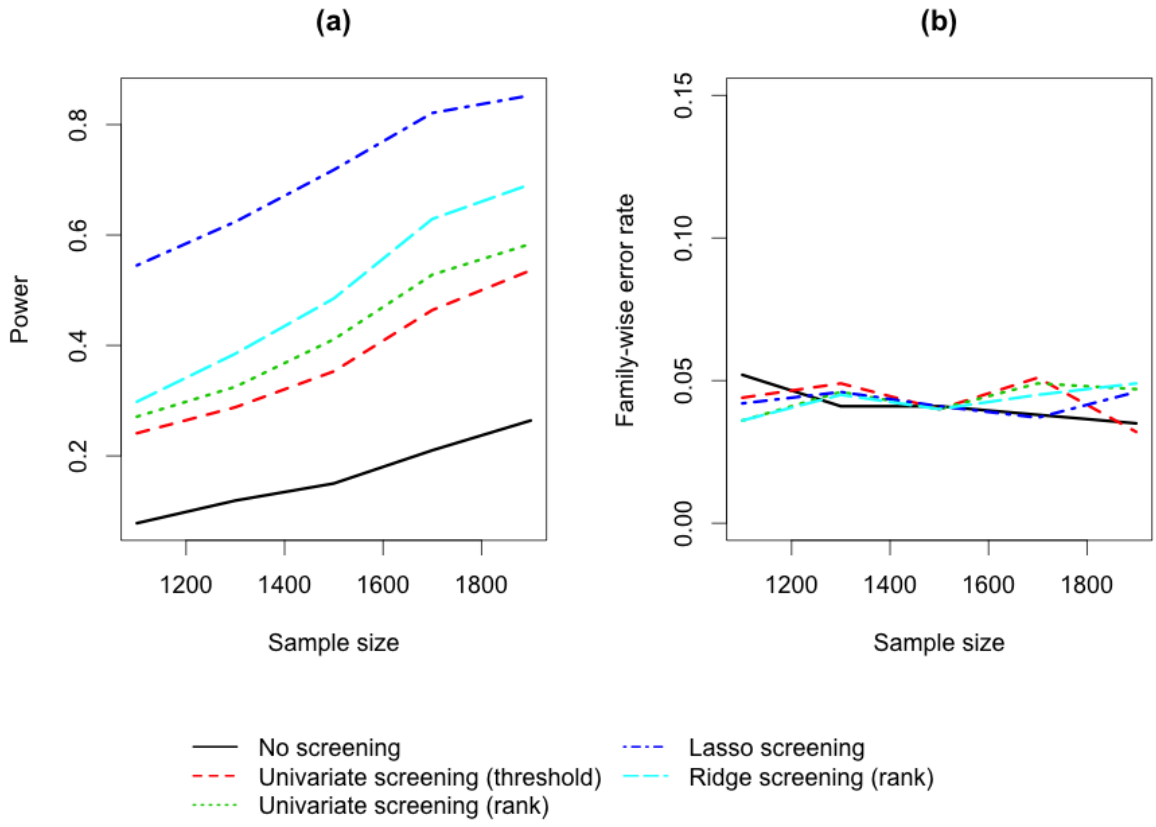

Figure 2.3: Comparison of two-stage interaction tests with different screening procedures in simulated data. The two panels represent: (a) power, 5,000 highly correlated biomarkers ($\rho = 0.6$), (b) family-wise error rate, 5,000 highly correlated biomarkers ($\rho = 0.6$).

In Appendix C.1, I provide simulation results of additional scenarios in which I

changed the main effects of the four biomarkers $\beta_{X_{21}}, \beta_{X_{41}}, \beta_{X_{61}}, \beta_{X_{81}}$. Relative patterns of performance among the screening strategies were consistent with the results described above, demonstrating further robustness of our methods and findings.

2.8 Data applications

In addition to validating my methods through simulations, I exemplified these approaches in three real data applications.

2.8.1 START trial

The START trial data and any required pre-processing were described in Section 1.6.1. I included 684 participants with 75 covariates in the analysis. I performed all the five interaction detecting procedures described in the previous section targeting a family-wise error rate of 0.05 and did not find any significant interactions. The top covariates output by the two "rank" approaches (the univariate screening and the ridge screening) are presented in Table 2.3, which shows that the top ranked covariates from these two procedures are similar in this data set where covariates have low correlation. This is because the ridge screening procedure mainly offers a benefit through the use of the multivariate model at stage 1 accounting for biomarker-biomarker correlations.

Table 2.3: Top covariates from different stage 1 marginal screening procedures: A covariate is highlighted in bold when the two screening procedures disagree in its rank. A covariate is underlined when it does not appear in the top 10 covariates of the other screening procedure.

	START trial	
	Univariate screening	Ridge screening
1	Total Inventory of Callous and Unemotional Traits	Total Inventory of Callous and Unemotional Traits
2	Total Antisocial Beliefs and Attitudes Scale	Total Antisocial Beliefs and Attitudes Scale
3	Strengths & Difficulties Conduct Problems Score	Strengths & Difficulties Conduct Problems Score
4	Strengths & Difficulties ProSocial Behaviour Score	Strengths & Difficulties ProSocial Behaviour Score
5	Strengths & Difficulties Hyperactivity Score	Strengths & Difficulties Hyperactivity Score
6	Volume of self reported delinquency excluding violence towards siblings	Volume of self reported delinquency excluding violence towards siblings
7	Strengths & Difficulties Total Difficulties Score	Strengths & Difficulties Total Difficulties Score
8	IQ	IQ
9	**Variety of self reported delinquency excluding violence towards siblings**	**Parental reported total Inventory of Callous and Unemotional Traits**
10	**Parent reported Strengths & Difficulties Conduct Problems Score**	**Alabama Positive Parental Involvement Score**

2.8.2 STOPAH trial

The 2×2 factorial STOPAH trial data and any required pre-processing were described in Section 1.6.2. In this application with binary outcomes, I applied my approaches to detect predictive biomarkers of steroid response in the treatment of alcoholic hepatitis. I included $1,068$ subjects with 40 covariates (a small number of which were demographic variables) in the analysis for detecting interaction with treatment (pentoxifylline or prednisolone). All five methods described in Section 2.7 did not find any significant biomarker-treatment interactions targeting a family-wise error rate of 0.05. Table 2.4 summarizes the top biomarkers from two "rank" procedures (univariate screening and ridge screening): The results are quite different between the ridge regression screening and the univariate screening, likely owing to the moderate correlation among the biomarkers.

Table 2.4: Top covariates from different stage 1 marginal screening procedures: A covariate is highlighted in bold when the two screening procedures disagree in its rank. A covariate is underlined when it does not appear in the top 10 covariates of the other screening procedure.

	STOPAH trial (pentoxifylline vs non pentoxifylline)	
	Univariate screening	Ridge screening
1	**Max GAHS**	**Age**
2	**MELD**	**Max GAHS**
3	**GAHS**	**MELD**
4	UNOS MELD	UNOS MELD
5	**Max GAHS - Categorical**	**GAHS**
6	**GAHS - Categorical**	**WHO Performance Status - Worst**
7	**Creatinine**	**Hepatic Encephalopathy**
8	**Discriminant Function**	**Urea**
9	**WHO Performance Status - Worst**	**Hepatic Encephalopathy - Worst**
10	**Urea**	**WHO Performance Status**

	STOPAH trial (prednisolone vs non prednisolone)	
	Univariate screening	Ridge screening
1	**Max GAHS**	**MELD**
2	**MELD**	**Max GAHS**
3	**GAHS**	**UNOS MELD**
4	**UNOS MELD**	**GAHS**
5	**Max GAHS - Categorical**	**Age**
6	**GAHS - Categorical**	**Creatinine**
7	**Creatinine**	**Urea**
8	**Discriminant Function**	**WHO Performance Status - Worst**
9	**WHO Performance Status - Worst**	**Hepatic Encephalopathy**
10	**Urea**	**Hepatic Encephalopathy - Worst**

2.8.3 PREVAIL trial

In the third application, I applied my approaches retrospectively to a phase II trial data set, which has high-dimensional gene expression biomarkers. The data set and any required pre-processing were described in Section 1.6.3. I restricted the analysis to $10,000$ probes with the highest standard deviations and the sample size is 61. The Sequential Organ Failure Assessment (SOFA) score was used as the continuous response endpoint. All five methods described in Section 2.7 did not find any significant biomarker-treatment interactions targeting a family-wise error rate of 0.05. A list of the top biomarkers from two

"rank" procedures (univariate screening and ridge screening) are presented in Table 2.5. The rankings of selected covariates are quite different between the ridge regression screening and the univariate screening procedures, likely owing to the high correlation among the biomarkers.

Table 2.5: Top covariates from different stage 1 marginal screening procedures: A covariate is highlighted in bold when the two screening procedures disagree in its rank. A covariate is underlined when it does not appear in the top 10 covariates of the other screening procedure.

	PREVAIL trial	
	Univariate screening	Ridge screening
1	**11715617_a_at**	**11715488_s_at**
2	**11749774_x_at**	**11715489_a_at**
3	**11725694_at**	**11739745_a_at**
4	**11746124_x_at**	**11749774_x_at**
5	**11739745_a_at**	**11746124_x_at**
6	11747047_a_at	11747047_a_at
7	**11715488_s_at**	**11728717_at**
8	**11720970_at**	**11725694_at**
9	**11751473_a_at**	**11716479_s_at**
10	**11756156_s_at**	**11752423_a_at**

2.8.4 Empirical between-stage correlation

In Section 2.5, I proved that stage 1 sparse regression screening and stage 2 interaction test statistics are asymptotically independent to each other for continuous outcomes. I calculated empirical correlations between stage 1 ridge screening and stage 2 interaction test statistics applied in the above three trial data sets (the lasso results are ignored because few stage 1 covariates have non-zero stage 1 lasso coefficients). Table 2.6 summarizes results from Pearson correlation tests, which shows that the empirical correlations between stages are close to zero. In all cases, with either continuous or binary outcomes, the 95% confidence interval contains zero. It is worth noting that even if my theoretical results in Section 2.5 only hold for continuous outcomes, application of my methods to the STOPAH trial data set with binary outcomes also yields nearly zero empirical correlation.

Table 2.6: Empirical correlation between stage 1 ridge screening and stage 2 interaction test statistics

	Response type	Estimate	p-value	95% confidence interval
START	continuous	0.044	0.711	$(-0.188, 0.271)$
PREVAIL	continuous	0.001	0.938	$(-0.019, 0.020)$
STOPAH (pentoxifylline)	binary	0.104	0.523	$(-0.214, 0.402)$
STOPAH (prednisolone)	binary	0.008	0.960	$(-0.304, 0.319)$

2.9 Discussion

Progress towards an era of precision medicine is accelerating due to the increasing availability of personal multi-omic data. Consequently, there is a growing need for studying biomarker-treatment interactions in large-scale studies of human populations. In this chapter, I have reviewed a variety of interaction testing methods and discussed which of them are applicable for detecting biomarker-treatment interactions in randomized clinical trials. Specifically, I showed that two-stage approaches, incorporating a marginal effect screening test to select a subset of biomarkers in stage 1, can provide greater power than a standard single-step interaction test in various scenarios.

The performance of a traditional one-biomarker-at-a-time screening test [43] tends to deteriorate when there is substantial correlation between covariates. Thus, I introduced two novel screening procedures using sparse regression methods, lasso and ridge regressions, to account for biomarker-biomarker correlations. One key requirement of applying two-stage approaches is independence between stage 1 and 2 tests, in order to preserve the overall family-wise error rate. In this chapter, I presented a proof for asymptotic independence between stage 1 sparse screening tests and stage 2 standard one-biomarker-at-a-time interaction tests. While asymptotic independence has been shown previously in the context of stage 1 univariate screening [53, 17], I have extended this theory to show the independence also holds for the use of multivariate models at stage 1, and furthermore, that it holds when inferring parameter estimates from penalized regression. To the best of our knowledge, this is the first time a theoretical basis is provided for incorporating sparse regression methods into two-stage interaction testing approaches. I demonstrated, in various simulated scenarios of highly correlated biomarkers, that these sparse regression screening methods can perform better than the traditional one-biomarker-at-a-time screening procedure. I exemplified my proposed methods for detecting biomarker-treatment interactions in three real trial data sets in which I evaluated the empirical between-stage correlations to verify my theory held out in practice.

I argue that a further limitation in the existing literature on two-stage interaction

testing for application to randomized clinical trials is the historical focus on family-wise error rates. Controlling the false discovery rate is less stringent than controlling the family-wise error rate. When prediction or exploration is considered as an end goal of detecting biomarker-treatment interactions, false positives are less likely to lead to severe safety risks. Controlling the false discovery rate has consequently been promoted as a more pragmatic choice, especially in genome-wide studies [68, 8]. Therefore, I also proposed adapting weighted false discovery rate controlling procedures [61] in the two-stage frameworks and demonstrated their effectiveness in simulations. Generally, false discovery rate controlling procedures have greater power, at the cost of an increased number of type I errors.

I showed that there exist special cases where my proposed two-stage screening strategy offers no benefit, e.g. when the main effect of a biomarker and its interaction effect with the treatment to the response are in opposite directions, such that the marginal effect cancels out. Different weighted hypothesis testing strategies differ in how much stage 1 information is used in the following stage 2 tests. I suggest exploring how these weighting schemes affect the power in different scenarios as a future topic for investigation.

It is known that ridge regression has a tendency to average effects across strongly correlated covariates. This phenomenon is not desirable for a screening strategy since it could inflate the number of non-interacting biomarkers being put forward to stage 2. Lasso, as an alternative sparse regression technique, does not exhibit this effect-averaging behavior. However, in the lasso screening procedures, covariates with small marginal effects may be dropped from further consideration, which is a clear disadvantage. Thus, to alleviate these issues, other formulations of sparse regression methods, which could be used for multivariate interaction analysis, are worth exploring in future research. For example, elastic net [87] and Bayesian variable selection [57] can be used as the screening procedure at stage 1; group lasso can be used as a single-stage or stage 2 interaction detecting method [84].

Since the main goal of employing the sparse regression screening procedures in stage 1 is to account for biomarker-biomarker correlations, some less computationally intensive multiple testing correction methods for correlated tests might be beneficial [56, 14, 29]. However, applying such methods, which calculate an "effective" number of independent tests [56, 29], to the single-step interaction test in a limited set of simulations did not offer any power improvement when controlling for the same family-wise error rate (results shown in Appendix C.1). I suggest further investigation into how to incorporate these methods into the two-stage interaction framework including a formal justification of the family-wise error rate control as a topic of future work.

My theoretical work only guarantees family-wise error rate control when using linear regression. A related technical issue was demonstrated by Sun et al. [69] that, for logistic regression, the interaction estimator under model misspecification can be biased when

the biomarker is associated either indirectly or directly with the outcome. This is a generic issue to interaction modeling using logistic regression, but could manifest in my framework as an elevated family-wise error rate at stage 2 one-biomarker-a-time tests. Chapter 4 discusses this problem in detail and proposes a de-biasing approach for the standard one-biomarker-a-time tests. However, the extent to which this bias might inflate family-wise error rates when applying the two-stage framework using logistic regression, and potential corrections, will be the topic of future work.

In summary, I adapted recently proposed two-stage methods for biomarker-treatment interaction testing in the randomized clinical trial setting, and proposed two novel screening tests using sparse modern regression techniques, lasso and ridge regressions, to account for biomarker-biomarker correlations. The simulation and real application results suggest use of sparse regression techniques in two-stage approaches can provide increased power for detecting interactions in randomized clinical trials.

Chapter 3

Adaptive signature design using biomarker-treatment interaction information to maximize treatment effect test statistics

3.1 Introduction

Modern medicine has seen an increasing interest in the development of targeted therapies [13]. Significant heterogeneity in response to treatments, resulting from individual genetic variability, has been found in many diseases. In molecularly targeted cancer drugs, therapies are often effective only for a subset of patients [34]. As a result, there is increasing attention toward discovery of biomarkers to identify subgroups of patients likely to benefit from a treatment [78]. In the context of clinical trials, these predictive biomarkers hold great potential to improve trial efficiency [83, 80]. When a treatment works only for a subgroup of patients, the overall treatment effect within the whole population might be low and undetected by a trial with moderate sample size. Identifying and targeting subgroups that likely benefit from the treatment will mean the power to show the effectiveness of the treatment may be much higher. Despite work on methods for identifying and utilizing predictive biomarkers, I will describe below a case for further work in this area.

Genomic technologies, such as microarrays and single-nucleotide polymorphism genotyping, provide rich biomarker panels from which to develop potential signatures to discriminate the subset of patients, who will most likely benefit from a targeted therapy. However, due to vast potential numbers of candidate biomarkers across different -omics platforms and the small sample size of early phase I and II trials, it is very challenging to develop reliable predictive signatures before a phase III trial, especially when the biological interplay between the treatment and the disease is not well understood [20]. The adaptive

signature design (ASD) was proposed as a solution for developing and testing a biomarker signature all within the same trial [25]. The approach employs two stages: stage 1 to use a proportion of patients to develop a signature to predict whether a patient is more likely to benefit from the new treatment, and stage 2 to use this signature to identify a "sensitive" subgroup from the remaining patients. More detail on how the ASD framework works was given in Section 1.5 of the introduction chapter.

In the signature development phase, the previously proposed ASD framework [25, 26] fits a univariate model (e.g. logistic regression) for each biomarker and selects "sensitive" biomarkers which exhibit significant interaction effects. In a subsequent classifier development phase, the marginal odds ratios for each selected sensitive biomarker are predicted using results from fitted univariate models in the previous phase. Specifically, the classification of a patient into a sensitive subgroup is based on whether or not the number of predicted marginal odds ratios for each "sensitive" biomarker carried by the patient that are significant exceeds a pre-specified threshold. Most subsequent literature focused on improving the signature development. When there is a large number of candidate biomarkers, a pre-selection procedure was shown to be beneficial in the ASD setting [12]. Sparse regression techniques such as lasso and ridge regressions were also recommended to be incorporated into the estimation of model parameters [86]. In the ASD setting, several papers [9, 10, 11] explored the use of Bayesian methods, and demonstrated that incorporating prior information improved estimation of marginal odds ratios for each biomarker. In addition to the binary outcome considered in the original ASD, methods to predict survival outcomes using the subset of selected biomarkers have also been proposed [50].

However, there has so far been limited work on optimizing the classification algorithm, which is a key ingredient of all these frameworks. In particular, it is unlikely that classifying patients according to simple univariate odds ratios is optimal. This motivates us to propose two new types of classifiers to be used in the ASD framework: one that thresholds the predicted risk difference, and one that classifies patients according to the expected change of the treatment effect test statistic. The prediction of risk differences and expected test statistic changes in both of my proposed classifiers are based on fitting a multivariate regression model using all stage 1 selected sensitive biomarkers at once. I explore both theoretically and through simulations how the two proposed classifiers compare to the existing classifier that is currently used within the ASD. I also illustrate application of my methods in two real clinical trial data sets.

3.2 Methods

The original ASD framework and its cross-validated extension were described in detail in Section 1.5. To recap, in order to perform a K-fold cross-validated ASD, a proportion of $(K-1)/K$ of trial participants form a development cohort to identify a biomarker signature. Then the signature is applied to the remaining $1/K$ participants (validation cohort) to select a sensitive subset of patients, who are more likely to benefit from the new treatment. The above procedure is repeated K times over K pairs of development vs validation cohorts. All K sensitive subsets of patients, selected from the K non-overlapping validation cohorts respectively, form the final subgroup. Presence of treatment effect is then tested within this subgroup (referred to as "subgroup test"). Since this test statistic is obtained by testing within a sample selected by cross-validation, as opposed to a standard trial population, the distribution of test statistic cannot be derived by standard asymptotic theory. A permutation-based method is recommended to approximate the empirical distribution of this test statistic and derive the p-value. A treatment effect test is also carried out within the whole sample (referred to as "whole-group test"). An ASD claims the treatment effect is efficacious if either the subgroup test or the whole-group test is significant. An overall significance level (e.g. 0.05) is distributed between the two tests (e.g. 0.01 for the subgroup test and 0.04 for the whole-group test).

The original ASD requires three parameters to develop the signature and classify patients: one for selecting predictive biomarkers and two for thresholding the predicted odds ratios between the new and control arms. Parameter tuning based on cross-validation (referred to as "inner cross-validation") is embedded into each loop of the cross-validation training the classifier using each development cohort (referred to as "outer cross-validation", of which the procedure is described in the previous paragraph). This nested cross-validation procedure was described in detail in Section 1.5 and illustrated in Figure 1.1.

In this section, within the cross-validated ASD framework, I develop two new types of classifiers and discuss theoretically how they compare to the classifier that is currently used in the ASD.

3.2.1 Multivariate risk difference (MRD) classifier

I propose the following classification method based on predicted risk differences:

1. Stage 1 as described in Section 1.5.1 remains the same. However, I also use stage 1 data to fit a multivariate regression model including the m^* sensitive biomarkers

that displayed significant univariate interactions:

$$logit\{E(R_i \mid X_{i(1)}, \ldots, X_{i(m^*)}, T_i)\} = \beta_0 + \beta_T T_i + \sum_{j=1}^{m^*} (\beta_{X_{(j)}} X_{i(j)} + \beta_{X_{(j)} \times T} X_{i(j)} \times T_i) \tag{3.1}$$

where R_i is the binary response outcome, T_i is the binary treatment assignment indicator, and $X_{i(1)}, \ldots, X_{i(m^*)}$ are the values of m^* sensitive biomarkers.

2. In stage 2, a patient is designated sensitive if the predicted risk difference $\hat{pr}(R_i = 1 \mid X_{i(1)}, \ldots, X_{i(m^*)}, T_i = 1) - \hat{pr}(R_i = 1 \mid X_{i(1)}, \ldots, X_{i(m^*)}, T_i = 0)$ exceeds a threshold γ.

I name the method employing this type of classifier the MRD (multivariate risk difference) design. Multivariate risk (or utility) differences are commonly used in precision medicine to model treatment effect heterogeneity and predict individual treatment responses for optimal therapy decisions [60, 65]. Particularly, in the context of clinical trials, multivariate risk differences are estimated to help identify individuals who are more likely to benefit from the new intervention [44, 46]. Within the ASD framework, in one cross-validation iteration, we use a portion of data to fit the regression model and make prediction of risk differences for the remaining patients to select a "sensitive" subgroup. Compared with the original ASD classifier, this method does not need the tuning parameter G, which thresholds the number of biomarkers with significant predicted marginal odds ratios which exceed γ. However, the price is that a multivariate regression model needs to be fit using stage 1 data. This multivariate model considers the joint effect of all sensitive biomarkers, thus potentially allowing a more sensitive summary measure for each patient than considering each sensitive biomarker separately. When the number of selected biomarkers is large, e.g. in a high-dimensional setting where $m^* > n$, fitting this multivariate model directly can become computationally prohibitive. I provide possible solutions to address this issue in the discussion section. Although one could also predict odds ratios by fitting this multivariate model, in Section 3.2.3, following a theoretical argument, I demonstrate how thresholding the predicted risk difference is better able to maximize the subgroup test statistic compared to thresholding the predicted odds ratio.

3.2.2 Multivariate gradient-based (MGB) classifier

To maximize the subgroup test statistic, there exists a trade-off between treatment effect and sample size. If a classifier is too stringent when selecting patients with a high probability of responding to the new treatment, it will tend to select a small, highly specific subgroup. Conversely, if a classifier is too liberal, the subgroup specific treatment

effect will be diluted. Thus, this trade-off inspired us to develop a classifier that directly maximizes the predicted test statistic change. I start my derivation by defining the contingency table over the selected sensitive subgroup as:

Table 3.1: 2×2 contingency table in the sensitive subgroup

	$R=1$	$R=0$
$T=1$	n_{11}	n_{10}
$T=0$	n_{01}	n_{00}

Note that the numbers in Table 3.1 depend on the classifier definition. If we wish to test the hypothesis whether the odds ratio is larger than one or not within this sensitive subgroup, the test statistic of Woolf's method [81] is a function of the above cell values

$$Z(n_{11}, n_{10}, n_{01}, n_{00}) = \frac{\log(\frac{n_{11}n_{00}}{n_{10}n_{01}})}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}} \tag{3.2}$$

I would like to estimate how this test statistic changes after newly classifying an additional individual as sensitive or not. Thus I express the gradient of this function as

$$\nabla Z = (d_{11}, d_{10}, d_{01}, d_{00})^T = \left(\frac{\partial Z}{\partial n_{11}}, \frac{\partial Z}{\partial n_{10}}, \frac{\partial Z}{\partial n_{01}}, \frac{\partial Z}{\partial n_{00}}\right)^T \tag{3.3}$$

which is a vector of partial derivatives that shows how the test statistic changes as a (continuous) change in the cell counts. Then, according to a particular patient's expected response with and without treatment, along with the treatment allocation fraction, We can calculate the expectation of the test statistic change when adding them into the table, assuming we know the current cell values. By the law of total expectation, the expectation of the test statistic change is

$$E = (d_{11}, d_{10}, d_{01}, d_{00})(p_{i11}, p_{i10}, p_{i01}, p_{i00})^T \tag{3.4}$$

where $p_{itr} = pr(T_i = t, R_i = r \mid X_{i(1)}, \ldots, X_{i(m^*)}) = pr(R_i = r \mid X_{i(1)}, \ldots, X_{i(m^*)}, T_i = t)pr(T_i = t)$. Leveraging these analytical results, the gradient-based classifier works as follows:

1. Stage 1 as described in Section 3.2.1 remains the same. In addition, stage 1 data is also used to form a 2×2 contingency table like Table 3.1 to estimate the gradient ∇Z. More precisely, assuming stage 1 sample set $\{(K-1)/K$ of all the patients in K-fold cross-validation$\}$ as S_1 and the overall sample set as S, we initialize the contingency table for the sensitive subgroup as $\hat{n}_{tr} = \sum_{i \in S_1} I(T_i = t, R_i = r) \cdot (|S| - |S_1|)/|S|$, where I is the indicator function. Intuitively, this assumes all the stage 2 patients

are initially selected into the sensitive subgroup, with numbers in the contingency table predicted with stage 1 data. Next, I use the predicted cell numbers $\hat{n}_{tr}$ to estimate the gradient ∇Z expressed by equation (3.3), of which the detailed formula is provided in Appendix A.6.

2. For the ith patient from stage 2 data, the four predicted probabilities $\hat{p}_{i11}, \hat{p}_{i10}, \hat{p}_{i01}, \hat{p}_{i00}$ can be calculated based on coefficients obtained from fitting the model (3.1) with stage 1 data. Using equation (3.4) to predict the test statistic change E specific to patient i, they are designated sensitive if the predicted value $\hat{E}$ is positive. Conversely, the patient is excluded from the sensitive subset if $\hat{E}$ is negative, since their inclusion is predicted to dilute the subgroup effect such that the test statistic would decrease.

I refer to this method the MGB (multivariate gradient-based) design. The classification procedure is carried out for all stage 2 patients simultaneously. This preserves independence between patient selection and treatment allocation. An advantage of this classification process is that it only requires one tuning parameter μ for selecting significant biomarker-treatment interactions in stage 1. This alleviates the computational burden compared with the original ASD which needs to tune three parameters.

3.2.3 Relationships between the ASD, MRD and MGB classifiers

In this section, I show how my two proposed classifiers - the risk difference classifier and the gradient-based classifier - relate to one another. I also show, numerically, how they will result in similar performance over a wide range of scenarios.

Considering first the gradient-based classifier, the expected test statistic change of equation (3.4) can be easily re-expressed as:

$$E = (d_{00} - d_{01}) \left(\frac{d_{11} - d_{10}}{d_{00} - d_{01}} p_{i11} - p_{i01} \right) + d_{10} p_t + d_{00}(1 - p_t)$$

where $p_t = pr(T_i = 1)$. A predicted positive test statistic change (i.e. $\hat{E} > 0$) therefore implies:

$$\frac{\hat{d}_{11} - \hat{d}_{10}}{\hat{d}_{00} - \hat{d}_{01}} \hat{p}_{i11} - \hat{p}_{i01} > \frac{\hat{d}_{10}\hat{p}_t + \hat{d}_{00}(1 - \hat{p}_t)}{\hat{d}_{01} - \hat{d}_{00}} \tag{3.5}$$

This inequality can be simplified to

$$A_1 \hat{p}_{i11} - \hat{p}_{i01} > A_2 \tag{3.6}$$

where $A_1 = (\hat{d}_{11} - \hat{d}_{10})/(\hat{d}_{00} - \hat{d}_{01})$ and A_2 equals to the right side of (3.5). In Appendix A.6,

I show that both A_1 and A_2 are asymptotically constants, relative to $\hat{p}_{i11}$ and $\hat{p}_{i01}$.

In comparison, the risk difference classifier thresholds $\hat{p}(R_i = 1 \mid X_{ij_1}, \ldots, X_{ij_{m^*}}, T_i = 1) - \hat{p}(R_i = 1 \mid X_{ij_1}, \ldots, X_{ij_{m^*}}, T_i = 0)$, which can be expressed as

$$\frac{\hat{p}_{i11}}{\hat{p}_t} - \frac{\hat{p}_{i01}}{1-\hat{p}_t} = \frac{1}{1-\hat{p}_t}\left(\frac{1-\hat{p}_t}{\hat{p}_t}\hat{p}_{i11} - \hat{p}_{i01}\right) = A_3(A_4\hat{p}_{i11} - \hat{p}_{i01})$$

where $A_3 = 1/(1-\hat{p}_t)$ and $A_4 = (1-\hat{p}_t)/\hat{p}_t$, both of which are asymptotic constants, relative to $\hat{p}_{i11}$ and $\hat{p}_{i01}$. Assuming the gradient-based classification criterion implied by inequality (3.6) is optimal in the sense of maximizing the subgroup test statistic, the risk difference classifier can achieve the equivalent criterion by using a threshold of A_2A_3, i.e.

$$A_3(A_4\hat{p}_{i11} - \hat{p}_{i01}) = A_3(A_1\hat{p}_{i11} - \hat{p}_{i01}) > A_2A_3$$

provided $A_1 = A_4$. I conduct a numerical analysis and show that when the probability of treatment assignment p_t is 0.5, $A_1 \approx A_4$ for a variety of scenarios. The results are provided in Appendix B.1. This implies the risk difference and gradient-based classifiers can achieve similar performance in practice, when the former has a sufficiently large space of tuning parameters to search for the optimal risk difference threshold.

I take a further step to compare the risk difference classifier with the classifier used by the original ASD. For a particular risk difference threshold to select a sensitive subgroup, in general, the same subgroup will not be discriminable using an odds ratio threshold. This is because there will always be individuals with different predicted odds ratios who exhibit the same risk difference for the new treatment versus control, and therefore would be selected into (or out of) the sensitive subgroup together under a risk difference classifier, but not necessarily under an odds ratio classifier. A numerical example is provided in Appendix B.2. Thus, if the gradient-based and risk difference classifiers always have the potential to select an optimal sensitive subgroup in the sense of maximizing the test statistic, it follows that classification by thresholding the odds ratio is sub-optimal since it may be unable to discriminate the same (optimal) subgroup.

3.3 Simulation studies

I conducted a simulation study to evaluate performance of my proposed methods described above. I generated $1,000$ replicate data sets, each of which has $m = 100$ biomarkers. Data were simulated under the model (1.6), where the treatment main effect was set to $\beta_T = \log(1.5)$ and the intercept $\beta_0 = 0$. Three biomarkers were ascribed main effects and interaction effects, i.e. $\beta_{X_1} = \beta_{X_2} = \beta_{X_3} = \log(1.5)$ and $\beta_{X_1 \times T} = \beta_{X_2 \times T} = \beta_{X_3 \times T} = \log(2.5)$. Seven other biomarkers were ascribed main effects on the trait without

interactions, i.e. $\beta_{X_4} = \beta_{X_5} = \beta_{X_6} = \beta_{X_7} = \beta_{X_8} = \beta_{X_9} = \beta_{X_{10}} = \log(1.5)$. All other biomarkers do not have direct effects on the response. The binary treatment assignment was drawn from a $Bernoulli(0.5)$ distribution and each biomarker X_j was generated from a standard normal distribution $\mathcal{N}(0, 1)$. All biomarkers were simulated as independent of one another - in practice this would be the case for a collection of -omics biomarkers corresponding to different gene regions, transcripts or proteins. A 0.01-level Fisher's exact test is carried out within the subgroup, and a 0.04-level Fisher's exact test is conducted within the whole sample (using the R function fisher.test, one-sided). For the cross-validated ASD framework with each of the three classifiers (the original ASD classifier, the MRD classifier and the MGB classifier), I use 10-fold outer cross-validation for selecting the sensitive subgroup and 10-fold inner cross-validation for selecting tuning parameters. The tuning sets of parameters are provided in Table 3.2. As suggested by Cherlin and Wason [12], in practice, I only use one inner cross-validation fold to select parameters within each outer cross-validation fold to save computational time. The permutation method repeats the whole process for the additional 99 (the number used by Freidlin et al. [26]) permuted data sets to obtain a valid p-value for subgroup treatment effect analysis.

Table 3.2: The tuning sets of parameters used by the inner cross-validation procedure

	μ	γ	G
ASD	$(0.05, 0.20, 0.35, 0.50)$	$\exp(0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8)$	$(1, 2, 3)$
MRD	$(0.05, 0.20, 0.35, 0.50)$	$(0.00, 0.03, 0.06, 0.09, 0.12, 0.15, 0.18, 0.21)$	NA
MGB	$(0.05, 0.20, 0.35, 0.50)$	NA	NA

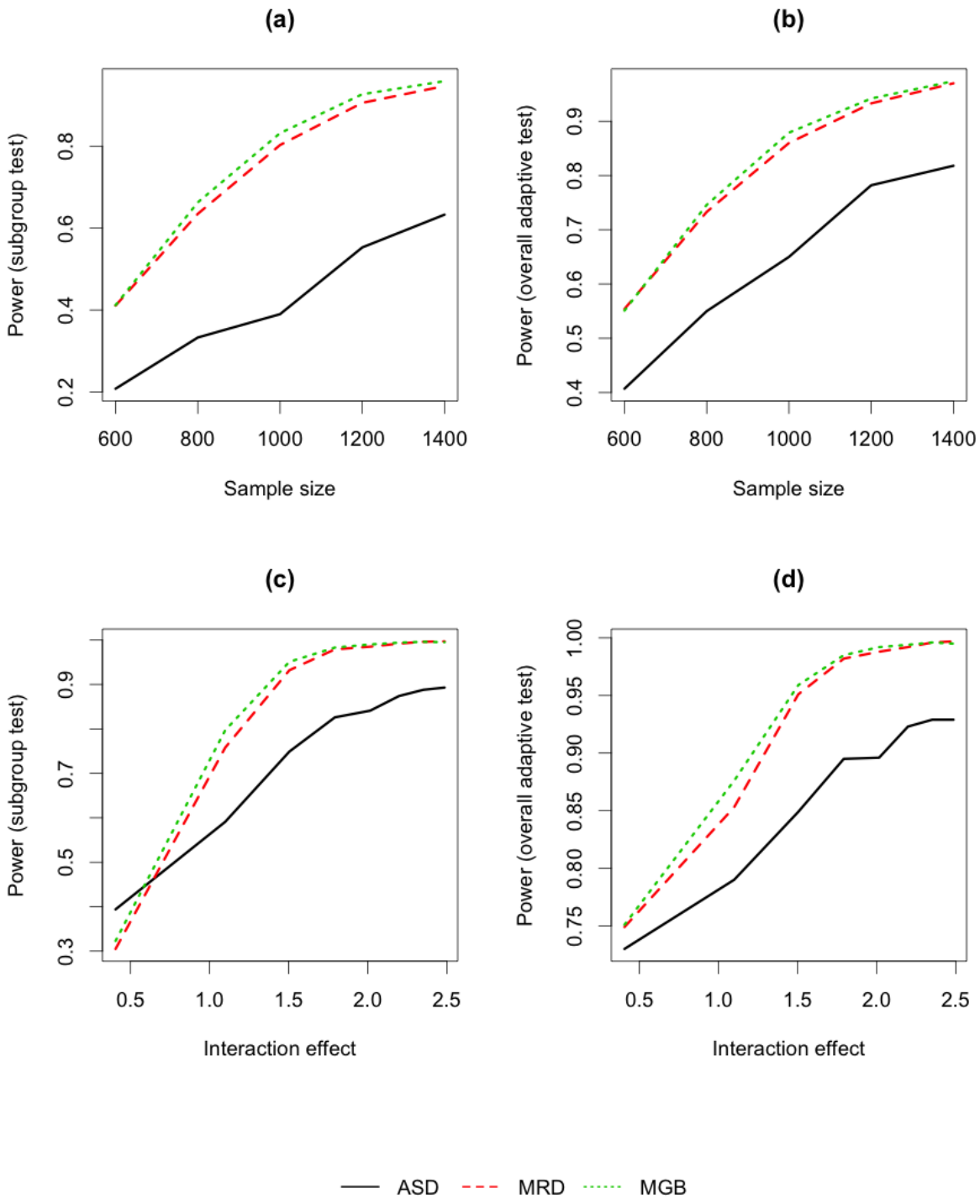

Figure 3.1: Comparison of adaptive signature designs in simulated data. The four panels represent power of: (a) the subgroup 0.01-level test with increasing sample size, (b) the overall adaptive test (the subgroup 0.01-level test and the overall 0.04-level test) with increasing sample size, (c) the subgroup 0.01-level test with increasing interaction effect, (d) the overall adaptive test with increasing interaction effect.

Table 3.3: Comparison of adaptive signature designs in simulated data when sensitive patients exist: sample size of $1,000$, $\beta_T = \log(1.5)$, three biomarkers have interaction effects of $\beta_{X_1 \times T} = \beta_{X_2 \times T} = \beta_{X_3 \times T} = \log(2.5)$, $1,000$ repetitions for power estimation

	ASD	MRD	MGB
Power (subgroup 0.01-level test)	0.390	0.803	0.832
Power (overall adaptive test)	0.650	0.860	0.879
Power (overall non-adaptive 0.05-level test)		0.632	
Average subgroup size	709	463	541
Average subgroup odds ratio	1.46	2.21	2.07
Average overall odds ratio		1.31	
Stage 2 computational time (hours)	15.6	17.0	12.6

Figure 3.1(a) and (b) compare power of subgroup 0.01-level tests and overall adaptive design tests among the three classifiers. The proposed MRD and MGB designs demonstrate substantial power increases over the original ASD. Consistently with my argument in Section 3.2.3, the risk difference classifier performs similarly to the gradient-based classifier across all simulation scenarios. Table 3.3 lists detailed results for when the sample size is $1,000$. The adaptive design incorporating a subgroup 0.01-level test using any of the three classifiers outperforms the non-adaptive design. As explained in Section 3.2.2, when selecting a treatment sensitive subgroup, there exists a trade-off between the degree of increased treatment effect and the size of the subgroup. In this scenario of a trial with sample size $1,000$, the original ASD classifier selects subgroups of average size 709 with an average odds ratio of 1.46, while the MRD classifier is more stringent, and selects averagely 463 patients with a higher average subgroup odds ratio of 2.21. The MGB classifier achieves a similar average odds ratio of 2.07 compared with the MRD classifier but with a larger average subgroup size of 541, which represents a better trade-off between the size of the sensitive group and increased efficacy there-in. Another advantage of the MGB classifier is that, free of tuning two extra parameters, it is the least computationally expensive procedure among the three: its computational time is reduced by 20% to 25% compared with the other two classifiers in stage 2.

Table 3.4: Comparison of adaptive signature designs in simulated data when sensitive patients do not exist and the new treatment is effective within the whole group: sample size of $1,000$, $\beta_T = \log(1.5)$, no biomarkers have interaction effects, $1,000$ repetitions for power and error rate estimation. Subgroup test error rates are counted as the proportion of repetitions that the overall 0.04-level test is non-significant and the subgroup 0.01-level test is significant

	ASD	MRD	MGB
Power (subgroup 0.01-level test)	0.372	0.213	0.218
Power (overall adaptive test)	0.741	0.736	0.740
Power (overall non-adaptive 0.05-level test)		0.758	
Average subgroup size	781	616	583
Average subgroup odds ratio	1.35	1.38	1.38
Average overall odds ratio		1.37	
Stage 2 computational time (hours)	15.1	16.2	12.1
Subgroup error rate	0.008	0.003	0.007

In Table 3.4, I simulated a scenario with no biomarker having any biomarker-treatment interaction. Other settings remain the same as the previous scenario. Thus, the treatment effect is expected to be significant within the whole sample, but no sensitive subgroup exists with a larger treatment effect. The adaptive designs perform similarly to the non-adaptive design in regards to power, as it allocates 80% of the significance level to the overall 0.04-level test and only 20% to the subgroup 0.01-level test. Though the three classifiers still select subgroups of patients, the odds ratios within these subgroups are close to those within the whole sample. In Table 3.4, I list the subgroup test error rates which are counted as the proportion of repetitions that the overall 0.04-level test is non-significant and the subgroup 0.01-level test is significant, which are controlled under 0.01 as expected for all three classifiers.

Table 3.5: Comparison of adaptive signature designs in simulated data when sensitive patients do not exist and the new treatment is not effective within the whole group: sample size of $1,000$, $\beta_T = 0$, no biomarkers have interaction effects, $1,000$ repetitions for power and type I error rate estimation

	ASD	MRD	MGB
Type I error rate (subgroup 0.01-level test)	0.014	0.011	0.014
Type I error rate (overall adaptive test)	0.044	0.042	0.044
Type I error rate (overall non-adaptive 0.05-level test)		0.047	
Average subgroup size	410	461	498
Average subgroup odds ratio	0.953	1.02	1.02
Average overall odds ratio		1.01	
Stage 2 computational time (hours)	15.1	16.1	12.0

In Table 3.5, I simulated a null scenario in which the treatment effect is zero and no biomarker-treatment interactions exist. The subgroup type I error rates are controlled under 0.01 successfully and the overall type I error rates are controlled under 0.05 as expected for both the overall adaptive and non-adaptive design tests. The overall type I error rate is probably below nominal level as there exists correlation between the overall test statistic and the subgroup test statistic.

Next, I simulated additional scenarios with one biomarker having a main effect log(1.5) and interaction effects varying from log(1.5) to log(12). Nine other biomarkers were ascribed only main effects log(1.5). Figure 3.1(c) and (d) show that when the interaction effect is small, the three classification methods cannot develop a reliable signature, thus meaning all the subgroup tests are not powerful. When the interaction effect increases, power of all the three designs increases. The MRD and MGB classifiers perform consistently better than the original ASD classifier, demonstrating the robustness of my methods.

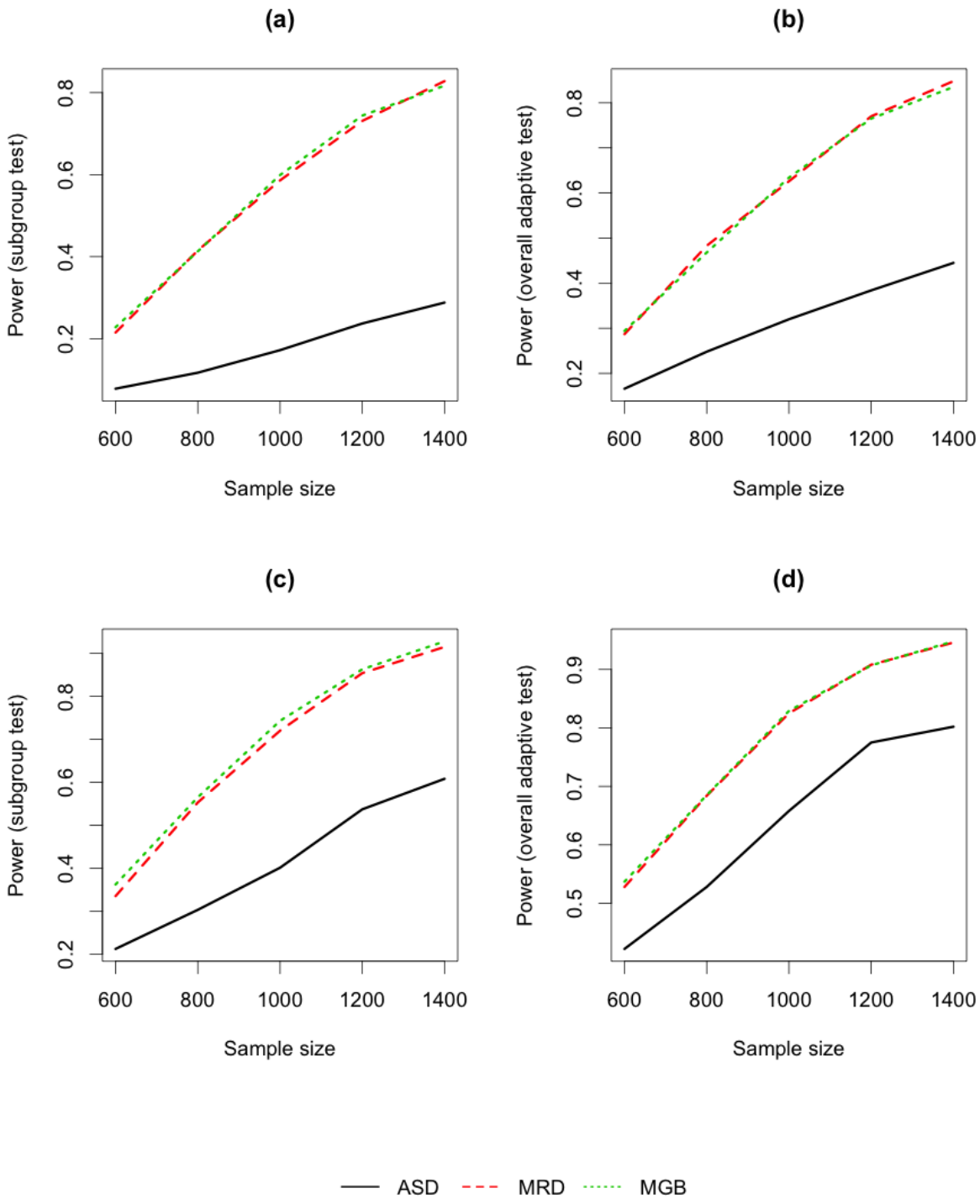

Figure 3.2: Comparison of adaptive signature designs in simulated data. The four panels represent: (a) (b) the data-generating model includes higher-order interactions $X_1 \times X_2 \times T, X_1 \times X_3 \times T, X_2 \times X_3 \times T$ but the analysis model does not, (c) (d) the data-generating model includes a higher-order interaction $X_1 \times X_2 \times X_3 \times T$ but the analysis model does not.

Lastly, I simulated scenarios where data-generating models are not in line with analysis models. In Figure 3.2(a, b) and (c, d), I added $\log(1.5)X_1 \times X_2 \times T + \log(1.5)X_1 \times X_3 \times T + \log(1.5)X_2 \times X_3 \times T$ and $\log(1.5)X_1 \times X_2 \times X_3 \times T$ respectively into the data-generating model (1.6) (other parameters remain the same with the setting described in the first paragraph of this section), while the fitted model does not include these higher-order interaction terms. Figure 3.2 shows that relative patterns of performance among the classifiers were consistent with the results described previously, demonstrating further robustness of our methods and findings.

3.4 Data applications

In addition to examining performance of my methods through simulation studies, I further compare the different approaches in two real randomized clinical trial data applications.

3.4.1 START trial

The START trial data and the pre-processing steps applied were described in Section 1.6.1. The primary binary outcome was used and 675 participants with 75 covariates were included in the analysis. I applied my methods to assess if the new treatment is superior to the control within the whole sample or within the sensitive subgroup selected by the inferred signature. I used 5-fold outer and inner cross-validation, and permuted data sets 999 times to obtain the adjusted subgroup test p-value.

Table 3.6: Comparison of adaptive signature designs: START trial, sample size of 675

	ASD	MRD	MGB
p-value (subgroup test)	0.294	0.514	0.535
p-value (overall test)		0.813	
Subgroup size	423	291	308
Subgroup odds ratio	1.14	0.945	0.927
Overall odds ratio		0.832	

Table 3.7: 2×2 contingency tables of different adaptive signature designs: START trial

	Whole trial sample	
	Response	No response
Treatment	297	43
Control	299	36
	Sensitive subgroup (ASD)	
	Response	No response
Treatment	184	22
Control	191	26
	Sensitive subgroup (MRD)	
	Response	No response
Treatment	128	18
Control	128	17
	Sensitive subgroup (MGB)	
	Response	No response
Treatment	134	19
Control	137	18

Table 3.6 summarizes results of applying the methods described in the previous section and Table 3.7 gives numbers of responders and non-responders between arms within sensitives subgroups selected by the three classifiers respectively. Neither the overall 0.05-level nor 0.04-level whole-group test were significant with the p-value of 0.813. The subgroups found by the three classifiers had odds ratios close to 1, indicating that the classifier did not develop a signature with evidence of being sensitive.

3.4.2 STOPAH trial

The STOPAH trial data and any required pre-processing were described in Section 1.6.2. In this application with 28-day mortality as the binary response endpoint, I applied my approaches to assess: 1) if the treatment with pentoxifylline is superior to the control without pentoxifylline; and 2) if the treatment with prednisolone is superior to the control without prednisolone. I used 5-fold outer and inner cross-validation, and permuted data sets 999 times to obtain the adjusted subgroup test p-value.

Table 3.8: Comparison of adaptive signature designs: STOPAH trial, sample size of $1,068$

Pentoxifylline vs no pentoxifylline			
	ASD	MRD	MGB
p-value (subgroup test)	0.378	0.591	0.608
p-value (overall test)		0.468	
Subgroup size	432	424	457
Subgroup odds ratio	1.02	0.913	0.914
Overall odds ratio		1.03	
Prednisolone vs no prednisolone			
	ASD	MRD	MGB
p-value (subgroup test)	0.257	0.251	0.199
p-value (overall test)		0.034	
Subgroup size	545	463	699
Subgroup odds ratio	1.09	1.13	1.18
Overall odds ratio		1.38	

Table 3.8 summarizes results of applying my methods to the STOPAH trial data. Treatment with pentoxifylline was not significant in the overall test with a p-value of 0.468. The three adaptive designs also did not find pentoxifylline effective within patient subgroups defined by the respective resulting classifiers, all of which still resulted in odds ratios around 1. Treatment with prednisolone was significant in the overall 0.05-level test. However, the subgroup 0.01-level tests were not significant under all the three designs. The odds ratios within the patient subgroups defined under each design were all smaller than the overall odds ratio of 1.38, demonstrating that all designs failed to find a reliable signature. In spite of this, since the overall treatment effect of prednisolone was significant under a significance level 0.04, the three adaptive designs were still able to conclude an effect in the overall patient population. This demonstrates that the adaptive signature design does little to compromise the trial's ability to detect treatment efficacy within the overall sample, even if a sensitive subgroup does not exist. Table 3.9 and 3.10 give numbers of responders and non-responders between arms within sensitives subgroups selected by the three classifiers respectively.

Table 3.9: 2×2 contingency tables of different adaptive signature designs: STOPAH trial, pentoxifylline vs no pentoxifylline

	Whole trial sample	
	Response	No response
Pentoxifylline	451	86
No pentoxifylline	444	87
	Sensitive subgroup (ASD)	
	Response	No response
Pentoxifylline	175	45
No pentoxifylline	168	44
	Sensitive subgroup (MRD)	
	Response	No response
Pentoxifylline	168	46
No pentoxifylline	168	42
	Sensitive subgroup (MGB)	
	Response	No response
Pentoxifylline	183	45
No pentoxifylline	187	42

Table 3.10: 2×2 contingency tables of different adaptive signature designs: STOPAH trial, prednisolone vs no prednisolone

	Whole trial sample	
	Response	No response
Prednisolone	459	75
No prednisolone	436	98
	Sensitive subgroup (ASD)	
	Response	No response
Prednisolone	237	48
No prednisolone	213	47
	Sensitive subgroup (MRD)	
	Response	No response
Prednisolone	291	59
No prednisolone	271	62
	Sensitive subgroup (MGB)	
	Response	No response
Prednisolone	300	51
No prednisolone	290	58

The results of these two real examples in Table 3.6 and 3.8 show that the MRD and MGB designs tend to perform similarly to each other, but differently to the original ASD. This is consistent with my theoretical argument in Section 3.2.3.

3.5 Discussion

This chapter has proposed two classification strategies to be used in the cross-validated adaptive signature framework. The MRD (multivariate risk difference) design relies on thresholding each patient's predicted risk difference. The MGB (multivariate gradient-based) design classifies patients based on the expected change of the treatment effect test statistic within the selected subgroup. Both designs consider the joint effect of all stage 1 selected sensitive biomarkers by fitting a multivariate model. This gives a more informative summary measure for each patient, compared to the classification strategy used in the original ASD (adaptive signature design), which considers a patient's predicted marginal odds ratio for each sensitive biomarker separately. I demonstrated that the MRD and MGB designs perform similarly to each other in a variety of scenarios, and presented an analytical argument that they both perform consistently better than the original ASD. This point is further validated through simulations and real data applications. I showed

that, in various simulated scenarios, the MRD and MGB designs achieve superior power compared to the original ASD. Another limitation of the original ASD is that it requires three tuning parameters to develop the classifier: μ for selecting sensitive biomarkers, γ for thresholding the predicted marginal odds ratio for each sensitive biomarker, and G for thresholding the number of sensitive biomarkers with predicted marginal odds ratios that exceed γ. This is computationally intensive, especially within the cross-validated ASD. In contrast, my MGB design requires the optimization of just a single tuning parameter and consequently, as I have demonstrated, offers a more computationally efficient ASD framework.

In the adaptive signature framework, the classification process is carried out after the completion of the trial, which preserves the ability to detect the overall treatment effect in all eligible patients. Throughout this chapter, my adaptive designs follow the original ASD, with 80% of the overall significance level allocated to the whole-group test and 20% to the subgroup test. This minimizes the risk of missing an overall treatment effect even when a signature is not detectable. Indeed, this property was demonstrated in my analysis of the effect of prednisolone on alcoholic hepatitis in data from the STOPAH trial. Even though no robust biomarker signature was detected, all adaptive signature designs still provided evidence for a significant effect of prednisolone in the combined set of all patients.

One limitation of the ASD is that it does not give a clear answer to the question of whether the use of any detected signature should be recommended in practice, if the treatment is still significant in the overall trial population. When the subgroup test is not significant, the answer is negative, because a reliable signature is not found. The answer is more ambiguous when the whole-group test and the subgroup test are both significant, because we cannot be certain if the significance of the subgroup test is driven by a genuine biomarker signature, or whether it represents a consistent overall treatment effect across subgroups. The reason we cannot be certain whether the signature should be recommended going forward is because the adaptive signature framework does not directly test for a difference in treatment effect between the subgroup and overall trial population. Further investigation is necessary into the comparison between the overall and subgroup treatment effects. This might be done either by using the existing trial data (e.g. a positive interaction effect between the signature and the treatment on the outcome would indicate the usefulness of a detected signature) or conducting a new randomized trial. I wish to pursue this question as a topic of future work. Thus, in the current ASD setting, I can only recommend use of any detected signature in the following scenario: The overall test is not significant while the subgroup test is significant.

My proposed MRD and MGB designs rely on fitting a multivariate model to all "sensitive" biomarkers, that is, which have demonstrated a univariate influence on efficacy. In a high-dimensional setting, where the number of selected biomarkers is larger than

the sample size, this approach may be infeasible using traditional regression methods. One option is that instead of selecting all sensitive biomarkers which exhibit biomarker-treatment interaction effects at some pre-specified significance threshold, I can use a fixed number (smaller than the sample size) of top ranked biomarkers to be selected into the next phase. Doing so would allow control over the number of parameters in the multivariate model to be fitted. Another option is to utilize sparse regression techniques such as lasso, ridge regression, or Bayesian variable selection to make predictions of relevant quantities (e.g. the odds ratio, risk difference or expected test statistic change) under high-dimensional settings, which is another area I wish to explore in future work [47, 76].

One limitation of my proposed MGB design is that the gradient ∇Z of the test statistic is estimated once using stage 1 data, and never updated as I classify patients one by one as "sensitive" or not in stage 2. Intuitively, updating this gradient after each stage 2 patient is processed may provide an "adaptive" classification criterion that better maximizes the subgroup test statistic. However, sequentially updating the classification algorithm in this way would introduce association between patient selection and treatment allocation within the subgroup, because the treatment allocation of the processed stage 2 patients will have bearing on the classification of the remaining unprocessed patients. Thus, different processing sequences of stage 2 patients could result in different sensitive subgroups, thus different subgroup test results, which would require careful thought around how to interpret. Therefore, although developing a sequentially updated classification algorithm is not trivial, I plan to explore the possibility in the future.

My discussion of the MGB in Section 3.2.2 was in terms of Woolf's association test statistic. I chose this statistic since concise analytical forms of its partial derivatives were available, which allowed my analytical arguments around the relative performance of the different adaptive signature designs. In practice, a Fisher's exact or chi-squared test is usually used to analyze a 2×2 contingency table like Table 3.1. Extending the gradient-based classifier to these statistics is possible and I present how to do this in Appendix A.7. The test (3.2) approximates these tests asymptotically and so I expect the comparative performances inferred in Section 3.2.3 to still hold, although extended formal analytical arguments based on these tests are non-trivial and beyond the scope of this thesis. Furthermore, as also demonstrated in Appendix A.7, the gradient-based classifier can be extended to some other types of outcomes, in addition to the binary end point which this chapter focused on.

In summary, I have presented two novel classification algorithms for use within the cross-validated ASD (adaptive signature design) framework: the MRD (multivariate risk difference) classifier and the MGB (multivariate gradient-based) classifier. Through theoretical arguments and simulations, I demonstrated that the MRD and MGB classifiers provided substantially better power than the original ASD univariate classifier. The MGB

classifier also exhibited the most computationally efficient performance. Application of my methods was further illustrated using two real trial data sets.

Chapter 4

De-biased logistic regression biomarker-treatment interaction estimator under model misspecification

4.1 Introduction

Heterogeneity in response to a treatment can be explained by the complex relationships between each person's biological characteristics and the treatment. These complex relationships often make it challenging to correctly specify a model. Inference based on misspecified models can result in false conclusions [15]. Therefore, understanding how sensitive results are to misspecified models is important.

The commonly used one-biomarker-at-a-time model for biomarker-treatment interaction analysis, which plays an important role in the frameworks proposed in Chapter 2 and 3, is incorrectly specified when a true interaction signal between another covariate and the treatment is not accounted for. This problem is closely related to the issue of invalid model-based inference when the relationship between the outcome and the environmental factor is misspecified in the model, which has been reported in gene-environment interaction studies [15, 1]. (The relevancy to our work follows from the fact that a binary environmental factor may be thought of as analogous to a treatment intervention.) Tchetgen and Kraft [70] analyzed gene-environment interaction testing using one-biomarker-at-a-time models under model misspecification in the scenario that the genetic biomarker is independent of all other biomarkers, the environmental factor, and also the outcome. In this particular setting, they proved that the estimated gene-environment interaction coefficient is asymptotically unbiased even if the environmental factor is misspecified in the fitted logistic regression model. Rosenblum and Van Der Laan [63] provided a proof of a similar result for the

generalized linear model in the randomized clinical trial setting. However, these theoretical results are constrained by the condition that the genetic biomarker is not associated with the outcome.

Sun et al. [69] showed that, under model misspecification, the estimated gene-environment interaction coefficient is unbiased for linear regression given gene-environment independence. They also showed that the interaction coefficient estimator is generally biased for logistic regression when the genetic biomarker is associated with the outcome directly or indirectly through correlation with confounding covariates. Their results built upon previous work in two respects. First, the result for linear regression does not depend on independence between the genetic biomarker and the outcome. Second, they formally specified the conditions under which the gene-environment interaction parameter estimate is biased.

This chapter is organized as follows. Section 4.2 summarizes the theoretical results of Sun et al. [69] in the gene-environment interaction setting. Section 4.3 applies these results to the randomized clinical trial setting and analyzes asymptotic bias of the biomarker-treatment interaction estimator for linear regression and logistic regression respectively under model misspecification. Section 4.4 derives two de-biasing approaches for testing biomarker-treatment interactions under logistic regression, and Section 4.5 evaluates performance of these de-biased estimators in different simulated scenarios. Section 4.6 illustrates applicability of my proposed methods in a real trial data set. The final section discusses potential future work related to this topic.

4.2 Asymptotic bias of the gene-environment interaction estimator under model misspecification

In this section, I summarize the existing theoretical results concerning asymptotic bias when testing for gene-environment interactions if the outcome-environment relationship is misspecified in the model.

Suppose the true data-generating model is specified by the generalized linear model of the form

$$G\{E(Y_i \mid X_{ij}, e_i, \boldsymbol{Z}_{ij})\} = \beta_{0_j} + \beta_{X_j} X_{ij} + \beta_e f(e_i) + \beta_{X_j \times e} X_{ij} \times h(e_i) + \boldsymbol{Z}_{ij}^T \boldsymbol{\beta}_{Z_j} \quad (4.1)$$

with Y_i denoting the response outcome, e_i the environmental exposure variable, $X_{i1}, \ldots, X_{im}$ representing the values of m biomarkers, for the ith patient. The vector $\boldsymbol{Z}_{ij} = (Z_{i1}, \ldots, Z_{ip_j})^T$ includes additional predictor variables, which could consist of functions of X_{ik} for any $k = 1, \ldots, m$ and $k \neq j$ (e.g. the main effect or interaction terms of X_{ik} other than X_{ij}), the environmental factor e_i, and any other confounding variables. The model coefficients

$\boldsymbol{\beta}_{Z_j} = (\beta_{i1}, \ldots, \beta_{ip_j})^T$ for $\boldsymbol{Z}_{ij}$ are all non-zero, i.e. $\beta_{ik} \neq 0$ for $k = 1, \ldots, p_j$. This means $\boldsymbol{Z}_{ij}$ only retains the predictor variables that have effects on the outcome. f and h are functions of e_i (e.g. e_i^2). G is a canonical link function.

A standard interaction test for the jth biomarker fits the (misspecified) model of the following form

$$G\{E(Y_i \mid X_{ij}, e_i)\} = \delta_{0_j} + \delta_{X_j} X_{ij} + \delta_e e_i + \delta_{X_j \times e} X_{ij} \times e_i \tag{4.2}$$

and the null hypothesis $\delta_{X_j \times e} = 0$ is tested. Notice that I use $\delta_{X_j \times e}$ to denote the interaction coefficient in the fitted (misspecified) model (4.2) and $\beta_{X_j \times e}$ as the interaction coefficient in the true data-generating model (4.1).

Comparing the fitted model (4.2) with the true model (4.1), the environmental exposure factor e_i is misspecified in (4.2) when either f or h is a nonlinear function of e_i, or $\boldsymbol{Z}_{ij}$ includes any term involving a nonlinear function of e_i. It has been reported [15, 1] that, under exposure misspecification, i.e. when either of these conditions holds, the inference on interaction could be invalid, resulting in inflated p-values, and therefore more type I errors. More precisely, when e_i is misspecified and the interaction coefficient is $\beta_{X_j \times e} = 0$ within the true data-generating model (4.1), the interaction estimator $\hat{\delta}_{X_j \times e}$ does not necessarily converge to zero in probability, i.e. the large sample limiting value of the coefficient $\delta_{X_j \times e}$ is not necessarily zero.

Sun et al. [69] summarized the conditions under which the exposure misspecification does not invalidate the interaction inference, i.e. $\delta_{X_j \times e} = 0$ in the fitted (misspecified) model when $\beta_{X_j \times e} = 0$ in the true model. For linear regression, with G set as an identity link function in the models (4.1) and (4.2), the sufficient conditions for valid interaction inference are (both of the following conditions must be met):

1. Gene-environment independence, i.e. $X_{ij} \perp e_i$.
2. At least one of X_{ij} or e_i is independent of each Z_{ik} for all $k = 1, \ldots, p_j$.

For logistic regression, with G set as a logit link function, the sufficient conditions for an interaction not being falsely detected under exposure misspecification are (all of the following conditions must be met):

1. Gene-environment independence, i.e. $X_{ij} \perp e_i$.
2. The jth biomarker X_{ij} is independent of each Z_{ik} for all $k = 1, \ldots, p_j$.
3. The jth biomarker's mains effect is zero, i.e. $\beta_{X_j} = 0$.

These conditions for logistic regression are much more stringent than those for linear regression, therefore the interaction tests for binary outcomes are more susceptible to the issue of bias under model misspecification.

4.3 Asymptotic bias of the biomarker-treatment interaction estimator under model misspecification

I now examine how the issue of bias discussed in the previous section manifests in a randomized clinical trial setting. I assume a true data-generating model of the following form

$$G\{E(Y_i \mid X_{i1}, \ldots, X_{im}, T_i)\} = \beta_0 + \beta_T T_i + \sum_{j=1}^{m} (\beta_{X_j} X_{ij} + \beta_{X_j \times T} X_{ij} \times T_i) \tag{4.3}$$

with Y_i denoting the response outcome, T_i the binary treatment-control indicator, $X_{i1}, \ldots, X_{im}$ representing the values of m biomarkers, for the ith patient. G is the canonical link function.

A common approach for detecting biomarker-treatment interactions tests each biomarker one at a time by fitting the model of the form

$$G\{E(Y_i \mid X_{ij}, T_i)\} = \delta_{0_j} + \delta_{X_j} X_{ij} + \delta_T T_i + \delta_{X_j \times T} X_{ij} \times T_i \tag{4.4}$$

and performing inference on the null hypothesis $\delta_{X_j \times T} = 0$, for each $j = 1, \ldots, m$. This particular setup is a special case of the setting discussed in Section 4.2. Next, I will describe how the relevant theoretical results in Section 4.2 can be translated to this clinical trial setting.

The relationship between the outcome and the treatment variable T_i is misspecified in the fitted model (4.4) when there exists any other biomarker which has an interaction with the treatment, i.e. $\beta_{X_k \times T} \neq 0$ for any $k = 1, \ldots, m$ and $k \neq j$ in the true model (4.3). Notice that if all the biomarkers do not have interaction with the treatment, i.e. the outcome-treatment relationship is correctly specified in (4.4), then no bias will be induced for testing biomarker-treatment interactions.

For linear regression, the sufficient conditions for valid interaction inference ($\delta_{X_j \times T} = 0$ when $\beta_{X_j \times T} = 0$) are achieved when the treatment allocation T_i is independent of both X_{ij} and $X_{ij} \times T_i$ for each $j = 1, \ldots, m$. The condition that T_i is independent of $X_{ij} \times T_i$, i.e. $cov(T_i, X_{ij} \times T_i) = 0$, is met when T_i is independent of X_{ij} and $E(X_{ij}) = 0$.[1] All the standard interaction tests I performed for continuous outcomes in Chapter 2 met these conditions.

For logistic regression, the biomarker-treatment independence should be guaranteed in a randomized trial. However, the third condition for valid interaction inference described in Section 4.2, i.e. $\beta_{X_j} = 0$, can be violated, because the main effect β_{X_j} is not necessarily zero when the interaction effect $\beta_{X_j \times T}$ is zero. This means that any biomarker with

[1] $cov(T_i, X_{ij} \times T_i) = E(X_{ij} T_i^2) - E(T_i) E(X_{ij} T_i) = E(X_{ij}) E(T_i^2) - E(X_{ij}) E(T_i)^2 = 0$ given that $X_{ij} \perp T_i$ and $E(X_{ij}) = 0$.

a non-zero main effect ($\beta_{X_j} \neq 0$ in the true model) will generally result in an invalid interaction test under the model (4.4) ($\delta_{X_j \times T} \neq 0$) even under the null ($\beta_{X_j \times T} = 0$ in the true model). It is not implausible that many treatment interacting biomarkers may also have an independent main effect on the outcome. Thus, the issue of bias could manifest in the standard one-biomarker-at-a-time interaction tests for binary outcomes as an elevated number of type I errors.

This motivated us to develop a de-biased logistic regression biomarker-treatment interaction estimator in this randomized clinical trial setting.

4.4 De-biased biomarker-treatment interaction estimator

In this section, I derive a de-biased estimator for one-biomarker-at-a-time interaction testing under logistic regression.

For logistic regression, we rephrase the true data-generating model (4.3), of which G is a logit link function, as

$$S^{-1}\{E(Y_i \mid \boldsymbol{X}_i)\} = \boldsymbol{X}_i^T \boldsymbol{\beta} \tag{4.5}$$

where the vector $\boldsymbol{X}_i = (1, T_i, X_{i1}, \ldots, X_{im}, X_{i1}T_i, \ldots, X_{im}T_i)^T$ includes the treatment assignment T_i, the values of m biomarkers $X_{i1}, \ldots, X_{im}$, and multiplicative interactions between the treatment and biomarkers $X_{i1}T_i, \ldots, X_{im}T_i$. The coefficient vector $\boldsymbol{\beta} = (\beta_0, \beta_T, \beta_{X_1}, \ldots, \beta_{X_m}, \beta_{X_1 \times T}, \ldots, \beta_{X_m \times T})^T$ defines the corresponding model coefficients for the treatment main effect, biomarkers' main effects and interaction effects. S is the sigmoid function which is the inverse of the logit link function for binary outcomes. The fitted one-biomarker-at-a-time models (4.4) for logistic regression are reformulated as

$$S^{-1}\{E(Y_i \mid \boldsymbol{V}_{ij})\} = \boldsymbol{V}_{ij}^T \boldsymbol{\delta}_j \quad (j = 1, \ldots, m) \tag{4.6}$$

where $\boldsymbol{V}_{ij} = (1, T_i, X_{ij}, X_{ij}T_i)^T$ includes the treatment assignment T_i, the value of jth biomarker, and their multiplicative interaction $X_{ij}T_i$. The vector $\boldsymbol{\delta}_j = (\delta_{0_j}, \delta_T, \delta_{X_j}, \delta_{X_j \times T})^T$ defines the coefficients in this fitted model. I assume treatment $T_i \in \{0, 1\}$ and $X_{ij} \sim \mathcal{N}(0, 1)$ for $j = 1, \ldots, m$ (which it is assumed can be achieved by properly scaling each covariate).

To evaluate asymptotic bias of the biomarker-treatment estimate, I first seek to establish a relationship between the coefficients $\boldsymbol{\delta}_j$ in the (misspecified) model (4.6) and the coefficients $\boldsymbol{\beta}$ in the true model (4.5). The asymptotic mean $\boldsymbol{\delta}_j$ of a maximum likelihood

estimator $\hat{\boldsymbol{\delta}}_j$ is the solution to the following score equations

$$E[\boldsymbol{V}_{ij}\{Y_i - S(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)\}] = \mathbf{0}$$

which may be re-expressed as

$$E[(1, X_{ij}, T_i, X_{ij}T_i)^T\{S(\boldsymbol{X}_i^T\boldsymbol{\beta}) - S(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)\}] = \mathbf{0} \tag{4.7}$$

The above equation involving vectors and matrices defines a system of equations. Solving this equation system requires evaluating the expectation of a sigmoid function, which generally does not have a closed-form expression. A sigmoid function can, however, be accurately approximated by the standard normal cumulative distribution function [6, 67, 22]. Utilizing this technique, I am able to show (in Appendix A.9) that solving the equation system (4.7) is approximately equivalent to solving the following four equations

$$rE(\boldsymbol{X}_i^T\boldsymbol{\beta}) - E(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j) = 0 \tag{4.8}$$
$$rE(X_{ij}\boldsymbol{X}_i^T\boldsymbol{\beta}) - E(X_{ij}\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j) = 0 \tag{4.9}$$
$$r_TE(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1) - E(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1) = 0 \tag{4.10}$$
$$r_TE(X_{ij}\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1) - E(X_{ij}\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1) = 0 \tag{4.11}$$

of which r and r_T are defined as

$$r = \sqrt{\frac{1 + \xi^2 var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)}{1 + \xi^2 var(\boldsymbol{X}_i^T\boldsymbol{\beta})}} \tag{4.12}$$

$$r_T = \sqrt{\frac{1 + \xi^2 var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1)}{1 + \xi^2 var(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1)}} \tag{4.13}$$

where ξ is a parameter used in the approximating function of the sigmoid function. I use $\xi^2 = \pi/8$ in the subsequent simulation studies [6].

Solving the system of equations (4.8) to (4.11) for $\boldsymbol{\delta}_j$ gives the solutions to the biomarker's main effect coefficient δ_{X_j} and the interaction effect coefficient $\delta_{X_j\times T}$ for the (misspecified) model

$$\delta_{X_j} = \frac{r(\beta_{X_j} + p_T\beta_{X_j\times T}) - p_Tr_T(\beta_{X_j} + \beta_{X_j\times T})}{1 - p_T} \tag{4.14}$$

$$\delta_{X_j\times T} = \frac{r_T(\beta_{X_j} + \beta_{X_j\times T}) - r(\beta_{X_j} + p_T\beta_{X_j\times T})}{1 - p_T} \tag{4.15}$$

where $p_T = pr(T_i = 1)$ is the probability of treatment assignment. Under the null $\beta_{X_j\times T} = 0$, formula (4.15) immediately gives $\delta_{X_j\times T} = (r_T - r)\beta_{X_j}/(1 - p_T)$. This implies,

under the null $\beta_{X_j \times T} = 0$, when the main effect is not zero $\beta_{X_j} \neq 0$, the asymptotic limit $\delta_{X_j \times T}$ of the interaction estimator in the fitted one-biomarker-at-a-time model (4.4) is generally not zero. Sun et al. [69] provided a similar result but I establish a quantitative relationship between the biomarker's main effect β_{X_j} and the estimated interaction coefficient $\delta_{X_j \times T}$.

Notice equations (4.14) and (4.15) involve $\beta_{X_j \times T}$, which is the interaction coefficient in the true model (4.3). Next, I seek to derive an unbiased interaction estimator, which will converge to $\beta_{X_j \times T}$. I first solve equations (4.14) and (4.15) for $\beta_{X_j \times T}$. This gives the following expression for $\beta_{X_j \times T}$:

$$\beta_{X_j \times T} = \frac{(r_T^{-1} - r^{-1})\delta_{X_j} + (r_T^{-1} - p_T r^{-1})\delta_{X_j \times T}}{1 - p_T}$$

of which, all the terms on the right side can be estimated from data (as described below). Thus, I propose the following de-biased estimator

$$\widetilde{\beta}_{X_j \times T} = \frac{(\hat{r}_T^{-1} - \hat{r}^{-1})\hat{\delta}_{X_j} + (\hat{r}_T^{-1} - \hat{p}_T \hat{r}^{-1})\hat{\delta}_{X_j \times T}}{1 - \hat{p}_T}$$

where $\hat{\delta}_{X_j}$ and $\hat{\delta}_{X_j \times T}$ are maximum likelihood estimators of δ_{X_j} and $\delta_{X_j \times T}$, which can be obtained by fitting the model (4.4). The treatment assignment probability p_T can be estimated empirically or in most cases will be a known feature of the trial design. Estimating r and r_T defined by equations (4.12) and (4.13) is more challenging, as it involves estimating the variances of two linear predictors in the true model, $var(\boldsymbol{X}_i^T \boldsymbol{\beta})$ and $var(\boldsymbol{X}_i^T \boldsymbol{\beta} \mid T_i = 1)$, and the variances of two linear predictors in the fitted model $var(\boldsymbol{V}_{ij}^T \boldsymbol{\delta}_j)$ and $var(\boldsymbol{V}_{ij}^T \boldsymbol{\delta}_j \mid T_i = 1)$. Fitting model (4.4) gives estimates of $var(\boldsymbol{V}_{ij}^T \boldsymbol{\delta}_j)$ and $var(\boldsymbol{V}_{ij}^T \boldsymbol{\delta}_j \mid T_i = 1)$.

To estimate $var(\boldsymbol{X}_i^T \boldsymbol{\beta})$ and $var(\boldsymbol{X}_i^T \boldsymbol{\beta} \mid T_i = 1)$, I propose fitting the saturated model (4.3) involving the main effect and interaction effect terms of all the biomarkers. When the true data-generating model is correctly expressed by (4.3) and sample size is sufficiently large, this approach will give good estimates of $var(\boldsymbol{X}_i^T \boldsymbol{\beta})$ and $var(\boldsymbol{X}_i^T \boldsymbol{\beta} \mid T_i = 1)$. In high-dimensional settings where $m \gg n$, I propose using group lasso to fit model (4.3). In practice, it is usually not possible to specify the true model correctly, e.g. when there exists some unknown confounder. I examine how our proposed methods perform in these scenarios in the simulation section.

The variance of the estimator $\widetilde{\beta}_{X_j \times T}$ can be obtained in a similar manner:

$$\begin{aligned} & var(\widetilde{\beta}_{X_j \times T}) \\ = \; & \{(\hat{r}_T^{-1} - \hat{r}^{-1})^2 var(\hat{\delta}_{X_j}) + (\hat{r}_T^{-1} - \hat{p}_T \hat{r}^{-1})^2 var(\hat{\delta}_{X_j \times T}) \\ & + 2(\hat{r}_T^{-1} - \hat{r}^{-1})(\hat{r}_T - \hat{p}_T \hat{r}^{-1}) cov(\hat{\delta}_{X_j}, \hat{\delta}_{X_j \times T})\}/(1 - \hat{p}_T)^2 \end{aligned}$$

This completes my derivation of the de-biased interaction estimator.

Throughout this section, we have assumed $T_i \in \{0, 1\}$. An alternatively way is to code treatment as $T_i \in \{-0.5, 0.5\}$, i.e. -0.5 for the control arm and 0.5 for the experiment arm [71]. In Appendix A.10, I give the form of the de-biased interaction estimator under this alternative treatment coding scheme.

4.5 Simulation studies

I generated simulated data sets to evaluate performance of my proposed de-biasing interaction testing procedures for binary outcomes. Data were simulated under the model (1.6), where the treatment main effect was set to $\beta_T = \log(1.5)$ and the intercept to $\beta_0 = 0$. Three biomarkers were ascribed main effects and interaction effects, i.e. $\beta_{X_1} = \beta_{X_2} = \beta_{X_3} = \log(1.5)$ and $\beta_{X_1 \times T} = \beta_{X_2 \times T} = \beta_{X_3 \times T} = \log(3)$. Three other biomarkers were ascribed main effects on the trait without interactions, i.e. $\beta_{X_4} = \beta_{X_5} = \beta_{X_6} = \log(4.5)$. All other biomarkers do not have effects on the outcome. Each biomarker X_j was generated from a standard normal distribution $\mathcal{N}(0, 1)$ and the binary treatment assignment was drawn from a $Bernoulli(0.5)$ distribution. All biomarkers were simulated as independent of one another. I considered two different numbers of simulated features: 1) a moderate number of biomarkers, $m = 100$, and sample size varying from $1,000$ to $2,000$; 2) a relatively larger number of biomarkers, $m = 1,000$, and sample size varying from $1,500$ to $2,500$. For each scenario, $1,000$ replicate data sets were generated to estimate power and family-wise error rates.

Three interaction testing procedures were compared:

1. "No de-biasing": A standard one-biomarker-at-a-time interaction test of the form (4.4) was performed.

2. "Non-penalized de-biasing": The one-at-a-time model of the form (4.4) was fitted for each biomarker $j = 1, \ldots, m$ to obtain the estimated main effect coefficients, interaction coefficients and the standard errors of these estimators. The full model of the form (4.3) was fitted to estimate r and r_T expressed by equations (4.12) and (4.13). The de-biased interaction estimators and their estimated standard errors were calculated as described in the previous section, which were further used to obtain the corrected p-values.

3. "Lasso de-biasing": The only difference from the above "non-penalized de-biasing" procedure is that group lasso was used to fit the full model of the form (4.3) to estimate r and r_T. The R package **glinternet** [47] was used with the penalization parameter λ_n chosen to minimize predictive errors under 5-fold cross-validation.

The Bonferroni correction was applied to all three procedures described above, targeting a family-wise error rate of 0.05.

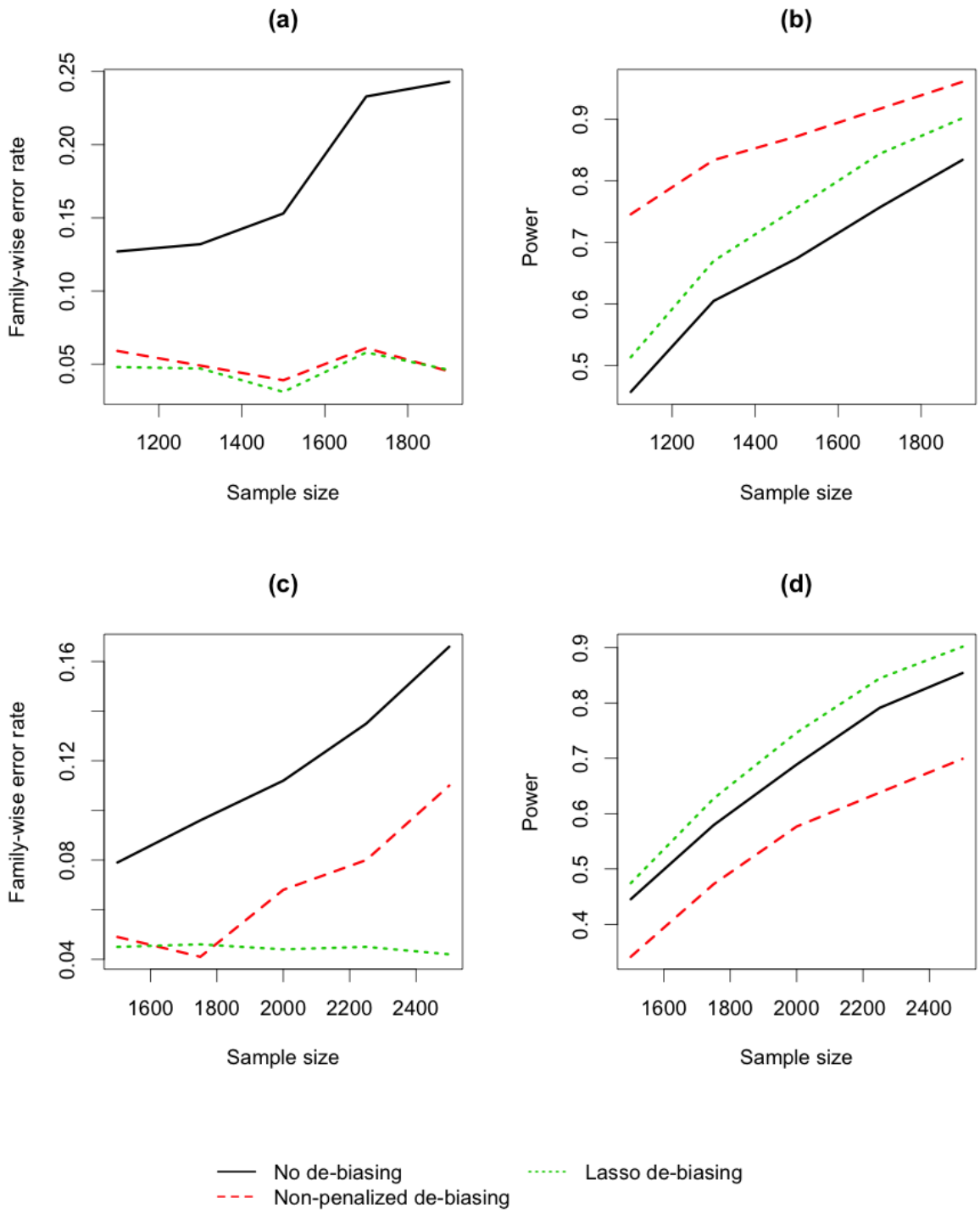

Figure 4.1: Comparison of interaction tests with different de-biasing strategies in simulated data. The four panels represent: (a) family-wise error rate, 100 independent biomarkers, (b) power, 100 independent biomarkers, (c) family-wise error rate, 1,000 independent biomarkers, (d) power, 1,000 independent biomarkers.

In Figure 4.1(a), I simulated a scenario with a moderate number of biomarkers $m = 100$

and sample sizes around $1,500$. The issue of asymptotic bias is evident for the standard interaction test as it suffered an increasing family-wise error rate with increasing sample size. Its family-wise error rate was inflated to above 0.1, while the two proposed de-biasing approaches demonstrated successful family-wise error rate control at the desired level of 0.05. Although the primary purpose of the de-biased interaction estimator is to reduce the number of type I errors, it is worth examining power. In Figure 4.1(b), both of the two proposed approaches showed higher power than the standard interaction test. This indicates, in this simulated scenario, the bias is in the opposite direction to the signal of the interacting biomarker, but has been corrected by the de-biasing approaches. The group lasso de-biasing procedure had lower power than the non-penalized de-biasing procedure in the scenario with $m \ll n$, possibly because its penalizing regression coefficients led to underestimated $var(\boldsymbol{X}_i^T\boldsymbol{\beta})$ and $var(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1)$ in equations (4.12) and (4.13).

In Figure 4.1(c), with a larger number of biomarkers $m = 1,000$, the proposed de-biasing procedures demonstrated improved family-wise error rate control compared to the standard interaction test. The standard interaction test still resulted in a severely inflated family-wise error rate. An inflated family-wise error rate was also evident for the non-penalized de-biasing procedure where the error rate is increased above the targeted level of 0.05 with increasing sample size. When the number of features is similar to the sample size, fitting a multivariate model like (4.3) can suffer from overfitting, thus leading to less accurate estimation of r and r_T expressed by equations (4.12) and (4.13). Notice that there are two trends for the de-biasing procedures as the sample size increases: 1) Asymptotic bias tends to result in more type I errors, and 2) the de-biasing procedure tends to better correct this bias, thus reducing the number of type I errors. Therefore, further increasing the sample size will likely improve the performance of the non-penalized de-biasing procedure. However, in this simulated scenario, the former trend dominated {Figure 4.1(c)} and the family-wise error rate increased with sample size for the non-penalized de-biasing procedure. Group lasso demonstrated better performance in this scenario, likely attributed to its use of penalization to avoid overfitting, which is expected to offer an advantage in the setting where the number of features is similar to the sample size. Figure 4.1(d) compares power between these approaches. The group lasso de-biasing procedure still resulted in better power than the standard interaction test, demonstrating the improvement it offers over standard logistic regression analysis with no de-biasing is robust across different dimension settings. The non-penalized de-biasing procedure, however, had lower power than the standard interaction test without de-biasing, owing to less accurate regression coefficient estimation due to overfitting.

In the above simulated scenarios, I assume all true predictors are available to the two de-biasing procedures for estimating r and r_T. Next, I examine how these methods perform when there exist unmeasured interacting biomarkers.

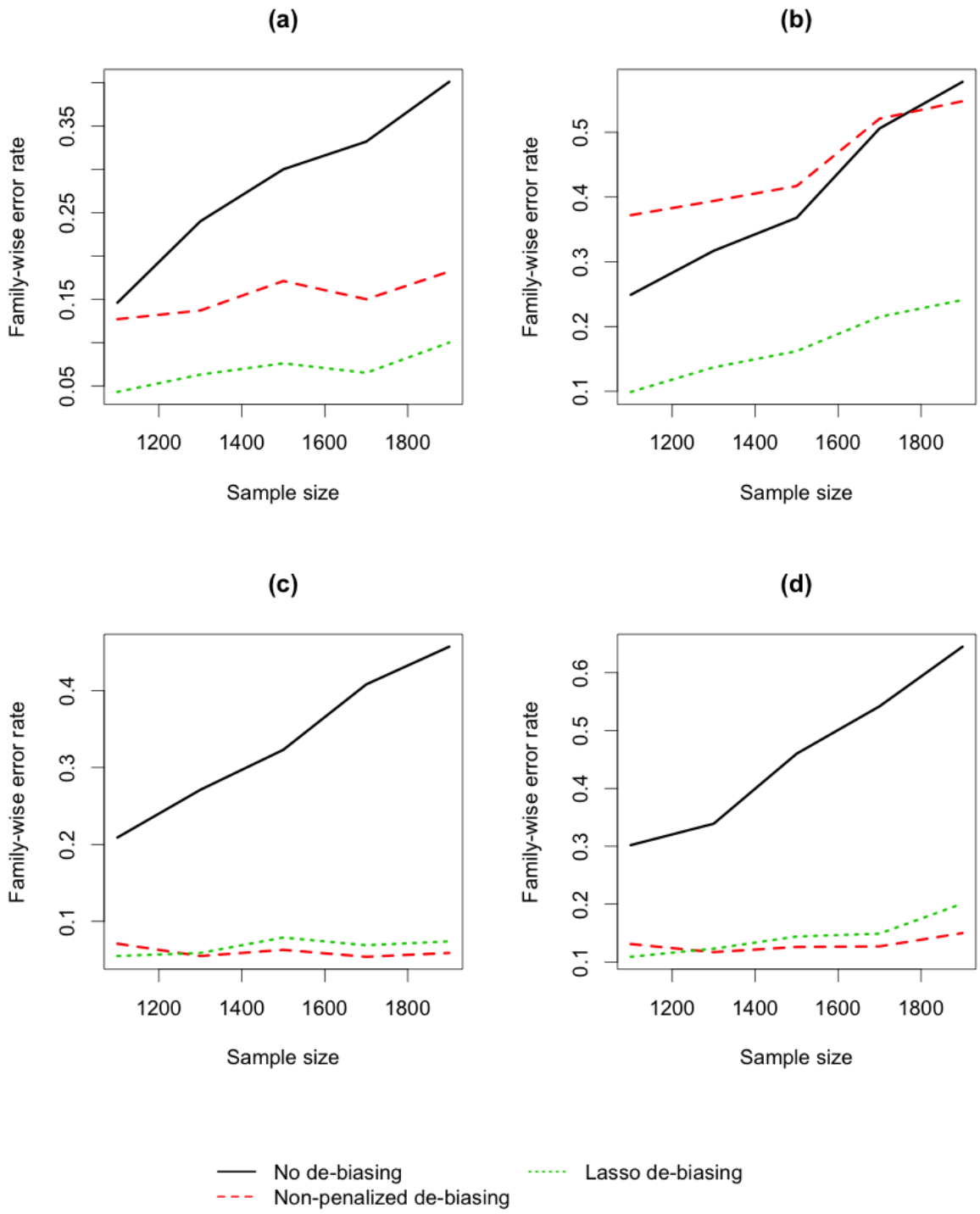

Figure 4.2: Comparison of interaction tests with different de-biasing strategies in simulated data. The four panels represent: (a) 100 independent biomarkers ($\rho = 0$), including 1 unmeasured interacting biomarker, (b) 100 independent biomarkers ($\rho = 0$), including 2 unmeasured interacting biomarkers, (c) 100 highly correlated biomarkers ($\rho = 0.6$), including 1 unmeasured interacting biomarker, (d) 100 highly correlated biomarkers ($\rho = 0.6$), including 2 unmeasured interacting biomarkers.

Using the base scenario with 100 independent biomarkers described above, I ascribed one biomarker with an interaction effect log(3) and a main effect log(1.5), in addition to the existing three interacting biomarkers and three biomarkers having only main effects. This additional interacting biomarker was "masked" from my interaction testing procedures. In practice, this corresponds to the case where there exists an unknown interacting confounder. Figure 4.2(a) shows that all the three interaction testing procedures resulted in inflated family-wise error rates above 0.05. Although the two de-biasing approaches demonstrated improved family-wise error rate control, the error rates increased with increasing sample size, which is evidence that the loss of information from the "masked" interacting biomarker is detrimental to the effectiveness of de-biasing. In Figure 4.2(b), I simulated a scenario with two masked interacting biomarkers in addition to the three interacting biomarkers. As more information was hidden, the non-penalized de-biasing procedure sometimes performed worse than the standard interacting test in this scenario. Although the group lasso de-biasing procedure also failed to control the family-wise error rate at 0.05, it demonstrated improved error rate control compared with the standard interaction test.

Next, I simulated scenarios with correlated biomarkers. All 100 biomarkers were partitioned into 10 clusters of correlated biomarkers, containing 10 biomarkers each. The 10 biomarkers within each cluster are correlated with each other ($\rho = 0.6$), but there are no correlations between biomarkers in different clusters. As in the scenario of Figure 4.2(a), four biomarkers were ascribed main effects and interaction effects, i.e. $\beta_{X_1} = \beta_{X_{11}} = \beta_{X_{21}} = \beta_{X_{31}} = \log(1.5)$ and $\beta_{X_1 \times T} = \beta_{X_{11} \times T} = \beta_{X_{21} \times T} = \beta_{X_{31} \times T} = \log(3)$. One of these four interacting biomarkers was masked. Three other biomarkers were ascribed main effects on the trait without interactions, i.e. $\beta_{X_{41}} = \beta_{X_{51}} = \beta_{X_{61}} = \log(4.5)$. All other biomarkers were assumed to not have effects on the outcome. Figure 4.2(c) shows family-wise error rates of all the interaction testing procedures under this scenario with highly correlated biomarkers. Although one interacting biomarker remained hidden, both of the two de-biasing approaches demonstrated substantially better error rate control than the standard interacting test. Intuitively, this is because information of this hidden biomarker is likely to be retained by its correlation with other "observed" biomarkers. In Figure 4.2(d), I introduced two hidden interacting biomarkers in this scenario with highly correlated biomarkers. The two de-biasing procedures did not control the family-wise error rate at 0.05, but both of them demonstrated improved error rate control compared with the standard approach without de-biasing.

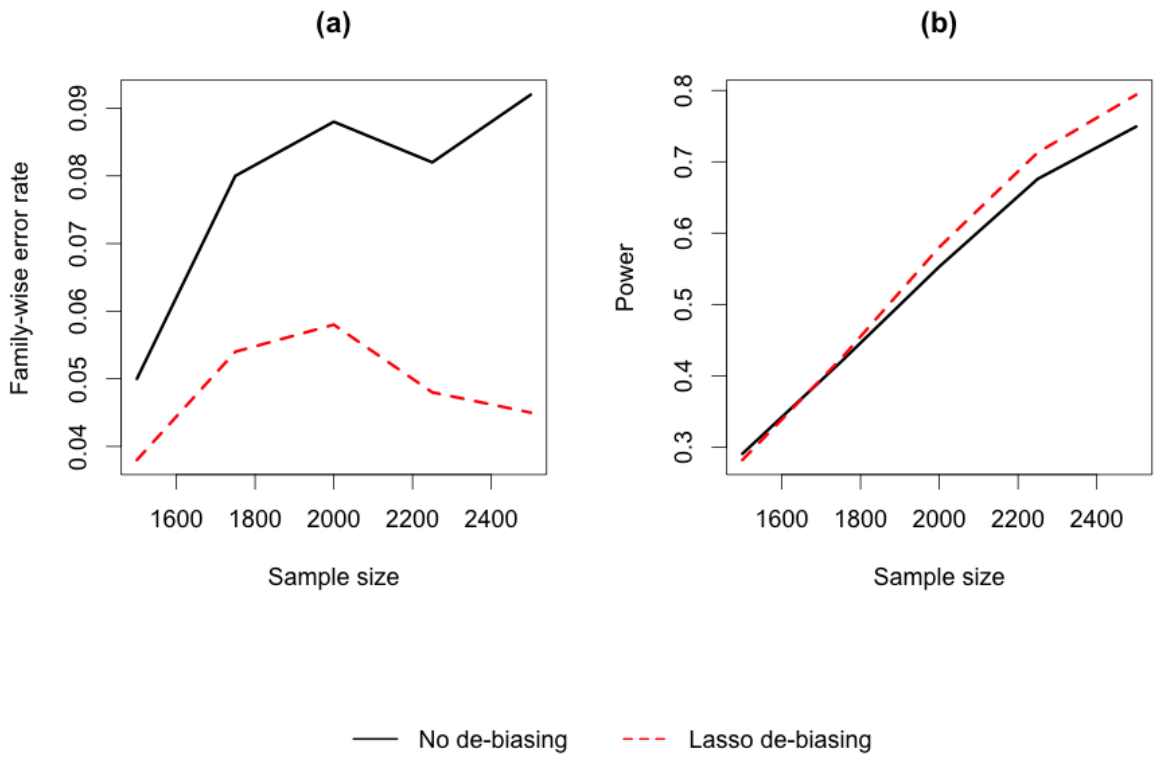

Figure 4.3: Comparison of interaction tests with different de-biasing strategies in simulated data. The two panels represent: (a) family-wise error rate, $5,000$ independent biomarkers, (b) power, $5,000$ independent biomarkers.

Lastly, I illustrated applicability of the group lasso de-biasing procedure in a high-dimensional setting where $n < m$. The parameterization remains the same with the base scenario, except that each generated data set now has $5,000$ biomarkers. Figure 4.3(a) and (b) show that the group lasso de-biasing procedure effectively controlled the family-wise error rate at around 0.05 without harming power compared with the standard interaction test.

In summary, the group lasso de-biasing procedure demonstrated robustness across different scenarios; when there are substantial correlations between biomarkers, the de-biasing procedures are able to improve family-wise error rate control even if some of interacting biomarkers were unmeasured.

In Appendix C.3, I provide additional simulation results under $\{-0.5, 0.5\}$ treatment coding, which showed slightly improved power of the group lasso de-biasing procedure compared with $\{0, 1\}$ treatment coding used in the simulations conducted in this section.

4.6 Data applications

4.6.1 PREVAIL trial

I applied my de-biasing approaches to a phase II trial data set with high-dimensional gene expression biomarkers. The data set and any required pre-processing were described in Section 1.6.3. Of all the 61 patients, 32 were in the lactoferrin-treated group, and the remaining were in the placebo group. No significant difference in clinical outcomes was found between the lactoferrin vs placebo groups by the trial. I restricted the analysis to $10,000$ probes with the highest standard deviations. ICU (intensive care unit) mortality was used as the binary response endpoint. Both the standard interaction test without de-biasing and the group lasso de-biasing procedure described in Section 4.5 did not find any significant biomarker-treatment interactions targeting a family-wise error rate of 0.05.

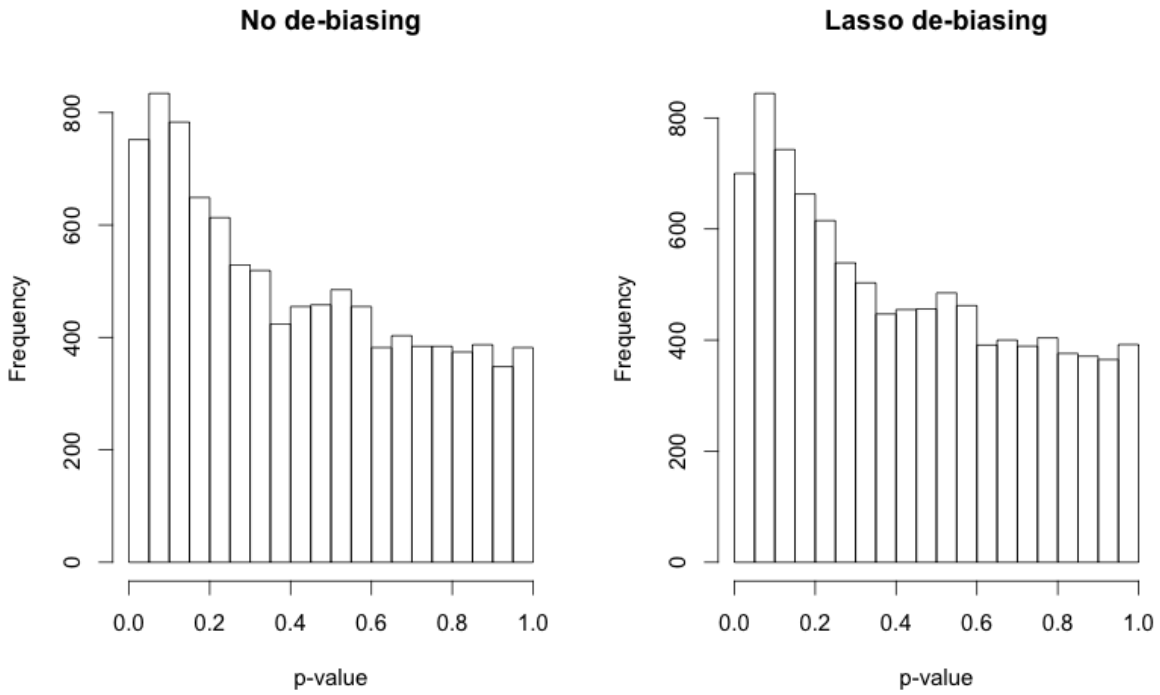

Figure 4.4: Histograms of p-values of interaction tests with different de-biasing strategies.

Figure 4.4 shows distributions of p-values of the standard interaction test without de-biasing and the group lasso de-biasing procedure. If the null hypotheses for all the biomarkers are correct, the distribution of p-values is expected to be uniform. However, both the distributions shown in Figure 4.4 were skewed towards 0. This could indicate that the p-values were skewed towards 0 due to model misspecification and the de-biasing procedure did not remove the bias here. Or this could indicate the existence of genes

interacting with the treatment but not sufficiently significant to be detected in this trial (not all the null hypotheses are true). The shapes of the distributions for the two procedures did not notably differ, which implies that the de-biasing procedure may not adjust the p-values much. Figure 4.5 shows the scatter plot of p-values on a logarithm scale before and after de-biasing. The diagonally aligned points of log-scale p-values before and after de-biasing mean that the de-biasing procedure did not make significant changes to the p-values.

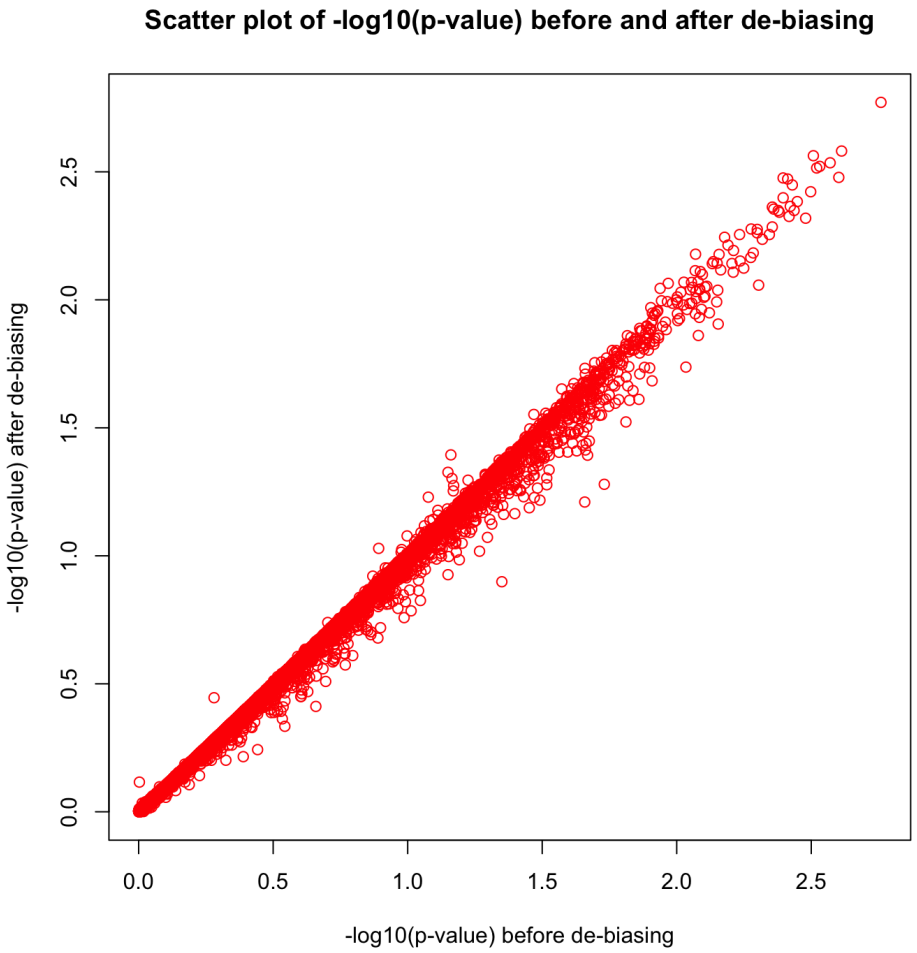

Figure 4.5: Scatter plot of -log10(p-value) before and after de-biasing: x-axis: -log10(p-value) of the no de-biasing procedure; y-axis: -log10(p-value) of the lasso de-biasing procedure.

Table 4.1: Empirical de-biasing parameters

	Estimate (averaged)
r	1.06
r_T	1.05
$var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)$ for the fitted model	0.305
$var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1)$ for the fitted model	0.250
$var(\boldsymbol{X}_i^T\boldsymbol{\beta})$ for the saturated model	0
$var(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1)$ for the saturated model	0

Table 4.1 lists the empirical de-biasing parameters estimated in the group lasso de-biasing procedure for adjusting the p-values of the standard interaction test. The fact that $r \approx r_T$ also indicates that the de-biasing procedure did not adjust the p-values very much. The two estimated variances of linear predictors $var(\boldsymbol{X}_i^T\boldsymbol{\beta})$ and $var(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1)$ for the saturated model are both zeros, because group lasso did not find any predictor with a non-zero coefficient, thus rendering the coefficient estimator $\hat{\boldsymbol{\beta}}$ for the saturated model a vector of all zeros.

4.7 Discussion

Biomarker-treatment interaction analysis plays an important role in explaining response heterogeneity to a therapy. However, interaction tests based on misspecified models can sometimes lead to asymptotic bias of interaction coefficient estimates. In this chapter, I studied the problem of interaction inference based on one-biomarker-at-a-time logistic regression models for binary outcomes under potential model misspecification. I found that when there exists at least one biomarker with a biomarker-treatment interaction, one-at-a-time interaction models for (almost) all candidate biomarkers are sensitive to model misspecification. The interaction effect estimate for a biomarker is generally only unbiased when the biomarker is not associated with the outcome. This issue can result in a lot of false positives when testing biomarker-treatment interaction based on one-at-a-time models. I proposed two de-biasing interaction testing procedures and demonstrated through simulations that my methods are able to substantially improve family-wise error rate control in various scenarios.

My proposed de-biased interaction estimators rely on fitting a saturated model including the treatment main effect term, the main effect and interaction terms of all biomarkers. In the high-dimensional setting, where the number of model parameters is larger than the sample size, it is not feasible to fit this model directly. Thus, I proposed use of group lasso to fit the saturated model. Other penalized regression methods, such as ridge regression and elastic net [87], can also be used. A performance comparison between group lasso and

these alternative methods for use in my framework is a topic I am interested to investigate in future work.

My work primarily focused on asymptotic bias of the estimated biomarker-treatment interaction coefficient for logistic regression. The results and approaches are also applicable to probit regression for binary outcomes, because the probit function is the inverse of the standard normal cumulative distribution function, which I used to approximate the sigmoid function in logistic regression. Generalized linear models with other types of link functions, e.g. the log function for Poisson regression, can also suffer from the issue of asymptotic bias when testing biomarker-treatment interactions [63]. I suggest extensions to these models as a topic of future work.

The work in this chapter is also applicable to gene-environment interaction testing, when the environmental factor is dichotomous and an assumption of gene-environment independence is valid. The de-biasing approaches are also potentially extendable for categorical environmental factors. How to alleviate the issue of asymptotic bias for continuous environmental factors or treatment variables (e.g. drug dose) may be an interesting topic to explore in the future.

Two-stage interaction tests (described in Chapter 2), incorporating a screening step to select a subset of biomarkers into stage 2 one-biomarker-at-a-time interaction tests, have been proposed to increase power. Unfortunately, my proposed de-biasing approaches for one-biomarker-at-a-time interaction tests are generally not applicable in this two-stage framework when a marginal association screening test is used in stage 1 (this type of screening was described in Section 1.2.2.2 and has been extended in Chapter 2). This is because the formula of my proposed de-biased interaction estimator involves both the main effect and interaction estimators for a biomarker, which introduces association between the de-biased interaction estimator (potentially used in stage 2) and the stage 1 marginal effect estimator for this biomarker (as a biomarker's main effect is associated with its marginal effect). Thus, my proposed de-biased estimator would break the between-stage independence required for family-wise error rate control by the methods described in Chapter 2. One possible solution is to utilize permutation tests to estimate the distribution of the stage 2 test statistic and further control the overall family-wise error rate [43]. Developing de-biased two-stage interaction tests for binary outcomes could be an interesting topic to study in the future.

In summary, I studied interaction testing based on one-biomarker-at-a-time models under model misspecification, and proposed two de-biasing interaction testing procedures. The simulation results show that my methods can provide improved family-wise error rate control for detecting biomarker-treatment interactions across various realistic randomized clinical trial scenarios.

Chapter 5

Conclusions and future directions

5.1 Conclusions

Throughout this thesis, I have been focusing on development of methods to analyze or utilize biomarker information in randomized clinical trials. Chapters 2 and 4 concentrated on the problem of detecting biomarker-treatment interactions. Chapter 3 had a focus on how to utilize biomarker information during the course of a randomized clinical trial.

In Chapter 2, I sought to overcome two limitations of traditional interaction analysis. First, the multiple testing burden of considering a large number of biomarkers may lead to low power for detecting biomarker-treatment interactions in clinical trials. Thus, only very large trials are sufficiently powered to detect an interaction effect. Developing statistical methods to overcome the low power problem in traditional interaction testing holds increasing importance, as modern medical research sees a growing amount of high-dimensional data. The same problem occurs in the closely related area of gene-environment interaction studies, when testing for a large number of genetic biomarkers (typically in thousands to millions). I examined how existing gene-environment interaction testing approaches can be modified for detecting biomarker-treatment interactions in the randomized clinical trial setting. Particularly, I found that the two-stage interaction testing approach is a promising framework to port from gene-environment to biomarker-treatment interaction analyses. Under a two-stage interaction testing framework, the analyst first performs a stage 1 screening procedure to select a reduced subset of biomarkers for the subsequent stage 2 interaction test, thus alleviating the multiple testing burden. I adapted this two-stage framework for application to detecting biomarker-treatment interactions in randomized clinical trials and found that it provides greater power than a traditional single-step interaction test.

The second limitation of traditional interaction analysis is that interaction testing is usually based on one-biomarker-at-a-time models. This ignores correlations between biomarkers. Thus, the number and locations of signals become ambiguous when there exist

substantial biomarker-biomarker correlations, which could lead to a lot of false positives. When the two-stage framework mentioned above uses a one-biomarker-at-a-time screening procedure, false positives from the screening test will lower power of the subsequent stage 2 interaction test, as now there are much more biomarkers that the multiple testing correction needs to account for. Chapter 2 proposed use of sparse regression methods accounting for biomarker-biomarker correlations at stage 1 to address this limitation. Two sparse regression screening procedures, utilizing lasso and ridge regression respectively, were proposed to be used in the two-stage framework. In order to preserve the overall family-wise error rate when applying a two-stage approach, stage 1 and 2 test statistics need to be (asymptotically) independent to each other. In novel theoretical developments, I proved the between-stage independence required for use of the new sparse regression screening procedures.

I demonstrated that, in a variety of simulated scenarios with highly correlated biomarkers, these sparse regression screening methods performed substantially better than the traditional single-step one-biomarker-at-time test and the two-step approach using a one-biomarker-at-a-time screening procedure. These results show that use of sparse regression techniques in the two-stage framework offers promising power increases for detecting biomarker-treatment interactions in randomized clinical trials. I further applied my methods in three real randomized clinical trials. Although I did not find any significant interaction effects using sparse regression for first stage screening, the empirical between-stage correlations were not significantly different from zero, verifying my theory in practice.

In Chapter 3, I changed my focus to the problem of how to utilize biomarker information during the course of a randomized clinical trial to improve trial efficiency. When a treatment is only or mostly beneficial to a subgroup of patients, an ordinary clinical trial may not be sufficiently powered to detect a significant treatment effect. In the absence of a predictive biomarker signature before the trial, the adaptive signature design (ASD) provides a useful framework to develop a signature and test the subgroup treatment effect all within the same trial. The approach works in a two-stage manner: stage 1 to develop a signature by selecting "sensitive" biomarkers which exhibit significant univariate biomarker-treatment interaction effects, and stage 2 to classify a patient into a sensitive subgroup based on predicted marginal odds ratios for each stage 1 selected biomarker. Upon examination of the existing literature in this area I found two key areas for improvement. First, the choice of classifying patients according to simple univariate odds ratios is sub-optimal. Secondly, the originally proposed ASD requires three tuning parameters to develop the classifier: one for selecting predictive biomarkers and two for thresholding the predicted odds ratios between the new and control arms. Selection of these tuning parameters based on cross-validation is computationally intensive.

To overcome these two limitations, I proposed two new classifiers to be used within the ASD framework for subgroup identification: the multivariate risk difference (MRD) classifier which thresholds the predicted risk difference, and the multivariate gradient-based (MGB) classifier which classifies patients according to the expected change of the treatment effect test statistic. The prediction of risk differences and expected test statistic changes are based on fitting a multivariate regression model using all stage 1 selected biomarkers. This multivariate model considers the joint effect of all sensitive biomarkers, thus potentially allowing a more sensitive summary measure for each patient than considering each sensitive biomarker separately. I argued theoretically and demonstrated through simulations how these two classification criteria provide greater power than the originally proposed univariate ASD classifier. The MGB classifier also leads to reduced computational times, because it only requires one tuning parameter for selecting sensitive biomarkers. I exemplified my methods using real data from two randomized trials. While I found no evidence of significant treatment signatures, I demonstrated that the adaptive signature design does little to compromise the trial's ability to detect treatment efficacy within the overall sample, even if a sensitive subgroup does not exist.

Chapter 4 returned to the topic of detecting biomarker-treatment interactions. I proposed methods to address a generic issue in interaction testing based on (misspecified) one-biomarker-at-a-time models for binary outcomes. The biomarker-treatment interaction coefficient estimate can be biased under model misspecification and lead to an increased number of false positives. This problem is closely related to the issue of invalid model-based inference in gene-environment interaction studies when the outcome-environment relationship is misspecified in the model. I translated the relevant theoretical results in the existing gene-environment interaction studies to the biomarker-treatment interaction setting, and derived conditions in which asymptotic bias of the interaction estimator will occur. I found when there exists at least one biomarker having biomarker-treatment interaction, one-at-a-time interaction models for all other biomarkers are sensitive to model misspecification. In this case, if a biomarker is associated with the response outcome directly or indirectly (through correlations with other confounding covariates), the interaction effect estimate for this biomarker is generally biased. This implies that prognostic biomarkers could sometimes be mistakenly identified as predictive biomarkers by using these (misspecified) one-biomarker-at-a-time interaction testing models.

Chapter 4 proposed two de-biasing procedures to adjust original one-biomarker-at-a-time test statistics. The de-biasing procedures rely on fitting a saturated model involving the main effect and interaction effect terms of all the available biomarkers, thus taking information from all interacting biomarkers to "correct" the bias. One de-biasing procedure fits this multivariate model directly without penalization, while the other de-biasing procedure utilizes the penalized regression technique, group lasso, to fit this model, which

is necessary in a high-dimensional setting. In simulated scenarios where all interacting biomarkers are observed, my proposed de-biasing interaction testing procedures are able to control the family-wise error rate at the desired level. The group lasso de-biasing procedure showed robustness across different dimension settings, different biomarker-biomarker correlation settings, and demonstrated improved family-wise error rate control even if some of interacting biomarkers are unmeasured. I illustrated applicability of the group lasso de-biasing interaction testing procedure in a trial data set with high-dimensional gene expression biomarkers.

5.2 Future directions

Limitations of the work presented in this thesis and some potential future topics to investigate have been discussed in each relevant chapter respectively. In this section, I go on to describe two possible extensions of the proposed methods, which could considerably increase the utility of my work, thus being the highest priority.

5.2.1 Extension of the ASD framework to account for biomarker-biomarker correlations

The classifiers proposed in Chapter 3 use standard one-biomarker-at-a-time interaction tests for selecting biomarkers with significant interaction with the treatment. These univariate models ignore correlations between biomarkers, which is sub-optimal. Although the MRD and MGB classifiers do consider joint effects of all the selected sensitive biomarkers, they do not account for correlations between all the biomarkers. Accounting for biomarker-biomarker correlations can reduce the number of false positives when selecting sensitive biomarkers, thus potentially increasing accuracy of predicting relevant quantities (e.g. the odds ratio, risk difference or expected test statistic change) for building the classifier to select sensitive patients. This could also potentially reduce computational times for the MRD and MGB classifiers, since they will consider fewer sensitive biomarkers when fitting the multivariate model.

Simulation studies in Chapter 3 assumed a simple setting where all the biomarkers are independent of one another. One might wonder how my proposed methods perform when there are substantial correlations between biomarkers and how we can further account for correlations among all the biomarkers other than correlations among only the selected sensitive subset of biomarkers. The two-stage interaction detecting framework incorporating sparse regression screening proposed in Chapter 2 is immediately available for use in the ASD framework to replace the standard one-biomarker-at-a-time interaction test. Using a two-stage interacting testing approach could increase power for selecting

sensitive biomarkers with biomarker-treatment interactions, as demonstrated in Chapter 2. It would be interesting to see how much the increased accuracy of interaction detection when using multivariate models will impact the overall power of the ASD framework for detecting the treatment effect.

The goal of testing interactions in the signature development phase within the ASD framework is to estimate relevant quantities (e.g. the odds ratio, risk difference or expected test statistic change) for building a classifier to select sensitive patients in the classification phase. Thus, another direction would be to extend the proposed framework in order to take account of biomarker-biomarker correlations. One approach may be to explore the applicability of sparse regression methods that are able to predict such quantities (e.g. the odds ratio, risk difference or expected test statistic change) for an individual without explicitly selecting sensitive biomarkers with biomarker-treatment interactions in the signature development phase. For example, using group lasso to fit a multivariate model considering all the biomarkers together has the potential of providing correlation-adjusted biomarker effect estimates which could be used to build a (newly proposed or existing) ASD classifier directly, without the need for explicitly selecting sensitive biomarkers. Therefore, as well as offering potentially more precise patient stratification by accounting for correlated biomarkers, the use of multivariate models in the ASD would avoid tuning the parameter μ for selecting sensitive biomarkers and can potentially further reduce computational times. An alternative approach to dimension reduction, while accounting for correlations between covariates, is to use Bayesian variable selection. In Appendices A.8 and C.2, I describe a framework for how to use Bayesian variable selection for detecting biomarker-treatment interactions in the ASD and demonstrate its performance in a limited set of simulations.

5.2.2 Extension of the two-stage interaction testing framework to binary outcomes with family-wise error rate control

The theoretical work in Chapter 2, proving asymptotic between-stage independence for stage 1 sparse regression screening, relies on homogeneity of variance, which is a standard assumption in linear regression. Therefore, the overall family-wise error rate control of the two-stage interaction testing procedures incorporating these sparse regression screening methods is only expected to work for continuous outcomes. There are two technical caveats of applying the two-stage interaction framework under logistic regression for binary outcomes. First, it is unknown whether the asymptotic between-stage independence for stage 1 sparse regression screening still holds or not in the absence of homoscedasticity. Under logistic regression for binary outcomes, the error variances differ for each value of the linear predictor, which is usually referred to as "heteroscedasticity", in contrast

to homoscedasticity. If the lack of homoscedasticity when modelling binary outcomes means that stage 1 and stage 2 test statistics are correlated with each other, the overall two-stage framework may result in an inflated family-wise error rate. The second issue which complicates application of the two-stage interaction framework for binary outcomes is the asymptotic bias of interaction coefficient estimates based on stage 2 one-biomarker-at-a-time interaction models, which has been discussed in Chapter 4. This issue, which is generic to interaction testing under logistic regression for binary outcomes, could result in an inflated family-wise error rate. Notice that the classifiers (existing and newly proposed) within the ASD framework described in Chapter 3 use one-biomarker-at-a-time models for binary outcomes, thus they are subject to asymptotic bias of interaction effect estimates. However, since prediction is the end goal of detecting biomarker-treatment interactions in the ASD, strict family-wise error rate control of interaction testing is not required (though a reduced number of false positives may increase the prediction accuracy). The de-biasing interaction testing procedures proposed in Chapter 4 are able to correct asymptotic bias induced by the (misspecified) one-biomarker-at-a-time interaction models in a variety of scenarios. However, application of these de-biasing procedures within the two-stage framework as stage 2 interaction tests will induce correlation between stage 1 and 2 test statistics, if (univariate or sparse regression) marginal association screening tests are used at stage 1. This is because these de-biasing procedures take account of both the main effect and interaction effect estimators for a biomarker, and the biomarker's main effect estimator (used in the stage 2 de-biased test) is usually associated with its marginal effect estimator (used in the stage 1 screening test). There are three possible routes to solutions for the two issues described above which currently face application of the two-stage interaction testing framework for binary outcomes:

1. Use stage 1 screening methods (existing or newly proposed in Chapter 2) and stage 2 de-biasing interaction testing procedures proposed in Chapter 4, and develop methods to control the overall family-wise error rate for the two-stage framework with correlated stage 1 and 2 tests: one possible solution to the between-stage correlation is to use permutation tests as suggested by Kooperberg and LeBlanc [43].

2. Use stage 1 screening methods (existing or newly proposed in Chapter 2) and stage 2 de-biasing interaction testing procedures proposed in Chapter 4, and derive conditions in which the between-stage correlation makes the overall two-stage procedure conservative in family-wise error rate control: one example of proving conservative family-wise error rate control for using a screening test correlated to stage 2 is shown in [38].

3. Develop new stage 1 screening and new stage 2 de-biasing interaction testing procedures, and prove asymptotic between-stage independence or demonstrate family-wise

error rate control building on the approaches (a permutation test or a proof of conservative error rate control) mentioned within the above two points. Ideally, the new methods are able to account for biomarker-biomarker correlations.

The first possible solution mentioned above can be attempted immediately. The second possible solution needs mathematical derivation for conditions demonstrating conservative family-wise error rate control and could lead to a negative result. The third solution needs substantial innovation and could be overoptimistic. It would be interesting to investigate into the possibility of extending the two-stage framework to interaction analysis for binary outcomes, as this would considerably increase its utility.

5.3 Concluding remark

In today's information-driven age, the analysis of randomized clinical trial data while utilizing biomarker information holds growing importance in modern medical research. In this thesis, I have developed several new methods to improve the existing approaches for detecting biomarker-treatment interactions and the existing adaptive signature design for using biomarker information in trials. In Chapter 2, I have shown that use of the two-stage interaction testing approach with sparse regression screening procedures is able to provide increased power for detecting biomarker-treatment interactions in randomized clinical trials. Chapter 3 proposed two new types of classifiers, for selecting patients who likely benefit from a new treatment, to be used in the adaptive signature design. These new classification strategies demonstrated improved power for determining the effectiveness of a new treatment. In Chapter 4, I proposed new procedures to tackle a generic interaction testing issue for binary outcomes under model misspecification and demonstrated my newly proposed procedures can provide improved family-wise error rate control. All of the methods proposed in this thesis have shown promising performance in various simulated scenarios and I have illustrated their applicability in the analysis of several real trial data sets. Existing literature of biomarker-treatment interaction studies and adaptive signature designs can hopefully benefit from this work and I am keen to see future extensions of my methods further enhance the utility of the work in this thesis.

Bibliography

[1] Lynn M Almli, Richard Duncan, Hao Feng, Debashis Ghosh, Elisabeth B Binder, Bekh Bradley, Kerry J Ressler, Karen N Conneely, and Michael P Epstein. Correcting systematic inflation in genetic association tests that consider interaction effects: application to a genome-wide association study of posttraumatic stress disorder. *JAMA Psychiatry*, 71(12):1392–1399, 2014.

[2] Olaf Beckonert, Hector C Keun, Timothy MD Ebbels, Jacob Bundy, Elaine Holmes, John C Lindon, et al. Metabolic profiling, metabolomic and metabonomic procedures for nmr spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, 2 (11):2692–2703, 2007.

[3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[4] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

[5] Leonard Bottolo, Sylvia Richardson, et al. Evolutionary stochastic search for bayesian model exploration. *Bayesian Analysis*, 5(3):583–618, 2010.

[6] Shannon R Bowling, Mohammad T Khasawneh, Sittichai Kaewkuekool, and Byung Rae Cho. A logistic approximation to the cumulative normal distribution. *Journal of Industrial Engineering and Management*, 2(1):114–127, 2009.

[7] Norman E Breslow and NE Day. Statistical methods in cancer research. Vol. 2. The design and analysis of cohort studies. Lyon, France: International agency for research on cancer, 1987, 1987.

[8] Damian Brzyski, Christine B Peterson, Piotr Sobczyk, Emmanuel J Candès, Malgorzata Bogdan, and Chiara Sabatti. Controlling the rate of GWAS false discoveries. *Genetics*, 205(1):61–75, 2017.

[9] Alexander C Cambon, Kathy B Baumgartner, Guy N Brock, Nigel GF Cooper, Dongfeng Wu, and Shesh N Rai. Classification of clinical outcomes using high-throughput informatics: Part 1–nonparametric method reviews. *Model Assisted Statistics and Applications*, 10(1):3–23, 2015.

[10] Alexander C Cambon, Kathy B Baumgartner, Guy N Brock, Nigel GF Cooper, Dongfeng Wu, and Shesh N Rai. Classification of clinical outcomes using high-throughput informatics: Part 2-parametric method reviews. *Model Assisted Statistics and Applications*, 10(2):89–107, 2015.

[11] Alexander C Cambon, Kathy B Baumgartner, Guy N Brock, Nigel GF Cooper, Dongfeng Wu, and Shesh N Rai. Properties of adaptive clinical trial signature design in the presence of gene and gene-treatment interaction. *Communications in Statistics-Simulation and Computation*, 46(10):8233–8250, 2017.

[12] Svetlana Cherlin and James MS Wason. Developing and testing high-efficacy patient subgroups within a clinical trial using risk scores. *Statistics in Medicine*, 2020.

[13] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795, 2015.

[14] Karen N Conneely and Michael Boehnke. So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *The American Journal of Human Genetics*, 81(6):1158–1168, 2007.

[15] Marilyn C Cornelis, Eric J Tchetgen Tchetgen, Liming Liang, Lu Qi, Nilanjan Chatterjee, Frank B Hu, and Peter Kraft. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *American Journal of Epidemiology*, 175(3):191–202, 2012.

[16] James Y Dai, Michael LeBlanc, and Charles Kooperberg. Semiparametric estimation exploiting covariate independence in two-phase randomized trials. *Biometrics*, 65(1): 178–187, 2009.

[17] James Y Dai, Charles Kooperberg, Michael Leblanc, and Ross L Prentice. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika*, 99(4):929–944, 2012.

[18] James Y Dai, Shuying S Li, and Peter B Gilbert. Case-only method for cause-specific hazards models with application to assessing differential vaccine efficacy by viral and host genetics. *Biostatistics*, 15(1):196–203, 2014.

[19] James Y Dai, Xinyi Cindy Zhang, Ching-Yun Wang, and Charles Kooperberg. Augmented case-only designs for randomized clinical trials with failure time endpoints. *Biometrics*, 72(1):30–38, 2016.

[20] Janet E Dancey and Boris Freidlin. Targeting epidermal growth factor receptor—are we missing the mark? *The Lancet*, 362(9377):62–64, 2003.

[21] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[22] Omar M Eidous and Rima Abu-Shareefa. New approximations for standard normal distribution function. *Communications in Statistics-Theory and Methods*, 49(6): 1357–1374, 2020.

[23] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[24] Peter Fonagy, Stephen Butler, David Cottrell, Stephen Scott, Stephen Pilling, Ivan Eisler, et al. Multisystemic therapy versus management as usual in the treatment of adolescent antisocial behaviour (START): a pragmatic, randomised controlled, superiority trial. *The Lancet Psychiatry*, 5(2):119–133, 2018.

[25] Boris Freidlin and Richard Simon. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research*, 11(21):7872–7878, 2005.

[26] Boris Freidlin, Wenyu Jiang, and Richard Simon. The cross-validated adaptive signature design. *Clinical Cancer Research*, 16(2):691–698, 2010.

[27] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33 (1):1, 2010.

[28] Wenjiang J Fu. Penalized estimating equations. *Biometrics*, 59(1):126–132, 2003.

[29] Xiaoyi Gao, Joshua Starmer, and Eden R Martin. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 32(4):361–369, 2008.

[30] W James Gauderman, Pingye Zhang, John L Morrison, and Juan Pablo Lewinger. Finding novel genes by testing G×E interactions in a genome-wide association study. *Genetic Epidemiology*, 37(6):603–613, 2013.

[31] W James Gauderman, Bhramar Mukherjee, Hugues Aschard, Li Hsu, Juan Pablo Lewinger, Chirag J Patel, John S Witte, Christopher Amos, Caroline G Tai, David Conti, et al. Update on the state of the science for analytical methods for gene-environment interactions. *American Journal of Epidemiology*, 186(7):762–770, 2017.

[32] Jelle J Goeman and Aldo Solari. Multiple hypothesis testing in genomics. *Statistics in Medicine*, 33(11):1946–1978, 2014.

[33] Piotr S Gromski, Howbeer Muhamadali, David I Ellis, Yun Xu, Elon Correa, Michael L Turner, et al. A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta*, 879:10–23, 2015.

[34] Viktor Grunwald and Manuel Hidalgo. Developing inhibitors of the epidermal growth factor receptor for cancer treatment. *Journal of the National Cancer Institute*, 95 (12):851–867, 2003.

[35] Yosef Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[36] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[37] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70, 1979.

[38] Li Hsu, Shuo Jiao, James Y Dai, Carolyn Hutter, Ulrike Peters, and Charles Kooperberg. Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genetic Epidemiology*, 36(3):183–194, 2012.

[39] Iuliana Ionita-Laza, Matthew B McQueen, Nan M Laird, and Christoph Lange. Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *The American Journal of Human Genetics*, 81(3): 607–614, 2007.

[40] James Jaccard. Interaction effects in logistic regression. Series: Quantitative applications in the social sciences, 2001.

[41] Muin J Khoury, Marta L Gwinn, Russell E Glasgow, and Barnett S Kramer. A population approach to precision medicine. *American Journal of Preventive Medicine*, 42(6):639–645, 2012.

[42] Keith Knight, Wenjiang Fu, et al. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378, 2000.

[43] Charles Kooperberg and Michael LeBlanc. Increasing the power of identifying gene×gene interactions in genome-wide association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 32 (3):255–263, 2008.

[44] Andrea Lamont, Michael D Lyons, Thomas Jaki, Elizabeth Stuart, Daniel J Feaster, Kukatharmini Tharmaratnam, Daniel Oberski, Hemant Ishwaran, Dawn K Wilson, and M Lee Van Horn. Identification of predicted individual treatment effects in randomized clinical trials. *Statistical methods in medical research*, 27(1):142–157, 2018.

[45] Dalin Li and David V Conti. Detecting gene-environment interactions using a combined case-only and case-control approach. *American Journal of Epidemiology*, 169(4):497–504, 2008.

[46] Junlong Li, Lihui Zhao, Lu Tian, Tianxi Cai, Brian Claggett, Andrea Callegaro, Benjamin Dizier, Bart Spiessens, Fernando Ulloa-Montoya, and Lee-Jen Wei. A predictive enrichment procedure to identify potential responders to a new therapy for randomized, comparative controlled clinical studies. *Biometrics*, 72(3):877–887, 2016.

[47] Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015.

[48] Chaichon Locharernkul, Vorasuk Shotelersuk, and Nattiya Hirankarn. Pharmacogenetic screening of carbamazepine-induced severe cutaneous allergic reactions. *Journal of Clinical Neuroscience*, 18(10):1289–1294, 2011.

[49] David J Lunn, John C Whittaker, and Nicky Best. A Bayesian toolkit for genetic association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 30(3):231–247, 2006.

[50] Shigeyuki Matsui. Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays. *BMC Bioinformatics*, 7(1):156, 2006.

[51] Kimberly McAllister, Leah E Mechanic, Christopher Amos, Hugues Aschard, Ian A Blair, Nilanjan Chatterjee, et al. Current challenges and new opportunities for gene-environment interaction studies of complex diseases. *American Journal of Epidemiology*, 186(7):753–761, 2017.

[52] Bhramar Mukherjee and Nilanjan Chatterjee. Exploiting gene-environment independence for analysis of case–control studies: an empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, 64(3):685–694, 2008.

[53] Cassandra E Murcray, Juan Pablo Lewinger, and W James Gauderman. Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology*, 169(2):219–226, 2008.

[54] John Muscedere, David M Maslove, J Gordon Boyd, Nicole O'Callaghan, Stephanie Sibley, Steven Reynolds, et al. Prevention of nosocomial infections in critically ill patients with lactoferrin: a randomized, double-blind, placebo-controlled study. *Critical Care Medicine*, 46(9):1450–1456, 2018.

[55] Paul James Newcombe, H Raza Ali, FM Blows, E Provenzano, Paul David Pharoah, C Caldas, et al. Weibull regression with Bayesian variable selection to identify prognostic tumour markers of breast cancer survival. *Statistical Methods in Medical Research*, 26(1):414–436, 2017.

[56] Dale R Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, 74(4):765–769, 2004.

[57] Robert B O'Hara, Mikko J Sillanpää, et al. A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1):85–117, 2009.

[58] LR Parham, LP Briley, L Li, J Shen, PJ Newcombe, KS King, et al. Comprehensive genome-wide evaluation of lapatinib-induced liver injury yields a single genetic signal centered on known risk allele HLA-DRB1* 07: 01. *The Pharmacogenomics Journal*, 16(2):180, 2016.

[59] Walter W Piegorsch, Clarice R Weinberg, and Jack A Taylor. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*, 13(2):153–162, 1994.

[60] Eric Polley and M van der Laan. Selecting optimal treatments based on predictive factors. *Design and Analysis of Clinical Trials with Time-to-Event Endpoints*, pages 441–454, 2009.

[61] Aaditya Ramdas, Rina Foygel Barber, Martin J Wainwright, and Michael I Jordan. A unified treatment of multiple testing with prior knowledge using the p-filter. *arXiv preprint arXiv:1703.06222*, 2017.

[62] Jacques Robert, Valérie Le Morvan, Elisa Giovannetti, Godefridus J Peters, et al. On the use of pharmacogenetics in cancer treatment and clinical trials. *European Journal of Cancer*, 50(15):2532–2543, 2014.

[63] Michael Rosenblum and Mark J Van Der Laan. Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics*, 65(3):937–945, 2009.

[64] Charles L Sawyers. The cancer biomarker problem. *Nature*, 452(7187):548–552, 2008.

[65] Zach Shahn, David Madigan, et al. Latent class mixture models of treatment effect heterogeneity. *Bayesian Analysis*, 12(3):831–854, 2017.

[66] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[67] Alessandro Soranzo and Emanuela Epure. Very simply explicitly invertible approximations of normal cumulative and normal quantile function. *Applied Mathematical Sciences*, 8(87):4323–4341, 2014.

[68] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[69] Ryan Sun, Raymond J Carroll, David C Christiani, and Xihong Lin. Testing for gene–environment interaction under exposure misspecification. *Biometrics*, 74(2): 653–662, 2018.

[70] Eric J Tchetgen Tchetgen and Peter Kraft. On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is mis-specified. *Epidemiology (Cambridge, Mass.)*, 22(2):257, 2011.

[71] Nils Ternes, Federico Rotolo, Georg Heinze, and Stefan Michiels. Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. *Biometrical Journal*, 59(4):685–701, 2017.

[72] Mark R Thursz, Paul Richardson, Michael Allison, Andrew Austin, Megan Bowers, Christopher P Day, Nichola Downs, Dermot Gleeson, Alastair MacGilchrist, Allister Grant, et al. Prednisolone or pentoxifylline for alcoholic hepatitis. *New England Journal of Medicine*, 372(17):1619–1628, 2015.

[73] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[74] Marynka M Ulaszewska, Christoph H Weinert, Alessia Trimigno, Reto Portmann, Cristina Andres Lacueva, René Badertscher, Lorraine Brennan, Carl Brunius, Achim Bub, Francesco Capozzi, et al. Nutrimetabolomics: an integrative action for metabolomic analyses in human nutritional studies. *Molecular Nutrition & Food Research*, 63(1):1800384, 2019.

[75] Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000.

[76] Jixiong Wang, Ashish Patel, James MS Wason, and Paul J Newcombe. Two-stage penalized regression screening to detect biomarker-treatment interactions in randomized clinical trials. *Biometrics*, 2021.

[77] Xiangyu Wang and Chenlei Leng. High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):589–611, 2016.

[78] Xiaoyu Wang and James Y Dai. TwoPhaseInd: an R package for estimating gene–treatment interactions and discovering predictive markers in randomized clinical trials. *Bioinformatics*, 32(21):3348–3350, 2016.

[79] James MS Wason and Frank Dudbridge. A general framework for two-stage analysis of genome-wide association studies and its application to case-control studies. *The American Journal of Human Genetics*, 90(5):760–773, 2012.

[80] James MS Wason, Jean E Abraham, Richard D Baird, Ioannis Gournaris, Anne-Laure Vallier, James D Brenton, et al. A Bayesian adaptive design for biomarker trials with linked treatments. *British Journal of Cancer*, 113(5):699, 2015.

[81] Barnet Woolf et al. On estimating the relation between blood group and disease. *Ann Hum Genet*, 19(4):251–253, 1955.

[82] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6): 714–721, 2009.

[83] Timothy A Yap, Shahneen K Sandhu, Paul Workman, and Johann S De Bono. Envisioning the future of early anticancer drug development. *Nature Reviews Cancer*, 10(7):514–523, 2010.

[84] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[85] Pingye Zhang, Juan Pablo Lewinger, David Conti, John L Morrison, and W James Gauderman. Detecting gene–environment interactions for a quantitative trait in a genome-wide association study. *Genetic Epidemiology*, 40(5):394–403, 2016.

[86] Lihui Zhao, Lu Tian, Tianxi Cai, Brian Claggett, and Lee-Jen Wei. Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association*, 108(502):527–539, 2013.

[87] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320, 2005.

Appendix A

Derivations

A.1 Power of case-only interaction tests

In this section, I first review the original validity proof of case-only tests by Piegorsch et al. [59], and then illustrate why it is not beneficial using this approach when cases are not oversampled.

I start by writing the interaction coefficient $\beta_{X_j \times T}$ in model (1.1) as (Y_i is assumed to be a binary response endpoint here, denoted as R_i)

$$\begin{aligned}\exp(\beta_{X_j \times T}) &= \frac{pr(R_i = 1 \mid X_{ij} = 1, T_i = 1) pr(R_i = 1 \mid X_{ij} = 0, T_i = 0)}{pr(R_i = 0 \mid X_{ij} = 1, T_i = 1) pr(R_i = 0 \mid X_{ij} = 0, T_i = 0)} \\ &\quad / \frac{pr(R_i = 1 \mid X_{ij} = 1, T_i = 0) pr(R_i = 1 \mid X_{ij} = 0, T_i = 1)}{pr(R_i = 0 \mid X_{ij} = 1, T_i = 0) pr(R_i = 0 \mid X_{ij} = 0, T_i = 1)}\end{aligned}$$

where X_{ij} is assumed to be binary.

Applying Bayes' rule, i.e. $pr(R_i = r \mid X_{ij} = x, T_i = t) = pr(X_{ij} = x, T_i = t \mid R_i = r) pr(R_i = r) / pr(X_{ij} = x, T_i = t)$, and next making substitutions with the definition of the conditional probability, i.e. $pr(X_{ij} = x, T_i = t \mid R_i = r) = pr(T_i = t \mid R_i = r, X_{ij} = x) pr(X_{ij} = x \mid R_i = r)$), we have

$$\begin{aligned}\exp(\beta_{X_j \times T}) &= \frac{pr(T_i = 1 \mid R_i = 1, X_{ij} = 1) pr(T_i = 0 \mid R_i = 1, X_{ij} = 0)}{pr(T_i = 1 \mid R_i = 0, X_{ij} = 1) pr(T_i = 0 \mid R_i = 0, X_{ij} = 0)} \\ &\quad / \frac{pr(T_i = 0 \mid R_i = 1, X_{ij} = 1) pr(T_i = 1 \mid R_i = 1, X_{ij} = 0)}{pr(T_i = 0 \mid R_i = 0, X_{ij} = 1) pr(T_i = 1 \mid R_i = 0, X_{ij} = 0)} \\ &= \frac{pr(T_i = 1 \mid R_i = 1, X_{ij} = 1) pr(T_i = 0 \mid R_i = 1, X_{ij} = 0)}{pr(T_i = 0 \mid R_i = 1, X_{ij} = 1) pr(T_i = 1 \mid R_i = 1, X_{ij} = 0)} \\ &\quad \times \frac{pr(T_i = 0 \mid R_i = 0, X_{ij} = 1) pr(T_i = 1 \mid R_i = 0, X_{ij} = 0)}{pr(T_i = 1 \mid R_i = 0, X_{ij} = 1) pr(T_i = 0 \mid R_i = 0, X_{ij} = 0)} \quad \text{(A.1)}\end{aligned}$$

Notice the first multiplicative term involves only $R_i = 1$ and the second term only $R_i = 0$. Next we show the second term is close to 1 under certain conditions. First, the

biomarker-treatment independence assumption implies

$$\frac{pr(T_i = 1 \mid X_{ij} = 1)}{pr(T_i = 0 \mid X_{ij} = 1)} = \frac{pr(T_i = 1 \mid X_{ij} = 0)}{pr(T_i = 0 \mid X_{ij} = 0)} \tag{A.2}$$

Recall using the law of total probability

$$\begin{aligned} pr(T_i = t \mid X_{ij} = x) = pr(T_i = t \mid R_i = 0, X_{ij} = x) pr(R_i = 0 \mid X_{ij} = x) \\ + pr(T_i = t \mid R_i = 1, X_{ij} = x) pr(R_i = 1 \mid X_{ij} = x) \end{aligned}$$

$pr(R_i = 1 \mid X_{ij} = x)$ is negligible under the rare response assumption, so we can write $pr(T_i = t \mid X_{ij} = x) \approx pr(T_i = t \mid R_i = 0, X_{ij} = x)$ and substitute them into (A.2)

$$\frac{pr(T_i = 0 \mid R_i = 0, X_{ij} = 1) pr(T_i = 1 \mid R_i = 0, X_{ij} = 0)}{pr(T_i = 1 \mid R_i = 0, X_{ij} = 1) pr(T_i = 0 \mid R_i = 0, X_{ij} = 0)} \approx 1 \tag{A.3}$$

Now we have shown that the second term of (A.1) is close to 1. Then we have

$$\exp(\beta_{X_j \times T}) \approx \frac{pr(T_i = 1 \mid R_i = 1, X_{ij} = 1) pr(T_i = 0 \mid R_i = 1, X_{ij} = 0)}{pr(T_i = 0 \mid R_i = 1, X_{ij} = 1) pr(T_i = 1 \mid R_i = 1, X_{ij} = 0)} \tag{A.4}$$

where no $R_i = 0$ is involved. We are able to estimate the interaction coefficient $\beta_{X_j \times T}$ with the marginal effect γ_{X_j} in (2.1) using cases only (responders where $R_i = 1$).

Notice $var_{standard}$, the variance of the maximum likelihood estimator of the logarithm of (A.1) is the sum of the variances of the maximum likelihood estimators of the logarithm of (A.4) and the logarithm of (A.3), var_{res} and var_{nres}. Thus when using (A.4) to estimate the interaction, we essentially regard there are infinitely many controls (non-responders where $R_i = 0$) and var_{nres} is 0. This is a gain compared with estimating the interaction using (A.1) when cases (responders) are "oversampled" and var_{nres} is (relatively) far from 0.

However, in randomized clinical trials, even if the response event is rare, we already have all the responders and non-responders in the data set. var_{nres} is effectively close to 0 relative to var_{res}, no matter which model to use, (A.1) or (A.4).

A.2 Between-stage independence proof: Murcray et al.

In this section, we show how Murcray et al. proved the asymptotic independence between the stage 1 gene-environment association test statistic and the stage 2 standard interaction test statistic.

Assuming a binary value of a biomarker X, a binary treatment assignment variable T and a binary response outcome R, the $2 \times 2 \times 2$ table for the study is:

Table A.1: $2 \times 2 \times 2$ contingency table in a randomized clinical trial

$R=0$	$T=0$	$T=1$	$R=1$	$T=0$	$T=1$
$X=0$	n_{000}	n_{001}	$X=0$	n_{100}	n_{101}
$X=1$	n_{010}	n_{011}	$X=1$	n_{110}	n_{111}

Assume that the random variables $\boldsymbol{Y} = (n_{000}, n_{001}, n_{010}, n_{011})^T$ and $\boldsymbol{Z} = (n_{100}, n_{101}, n_{110}, n_{111})^T$ both yield the multinomial distribution, i.e. $\boldsymbol{Y} \sim Mult\{(p_{000}, p_{001}, p_{010}, p_{011})^T, N_Y\}$ and $\boldsymbol{Z} \sim Mult\{(q_{100}, q_{101}, q_{110}, q_{111})^T, N_Z\}$ where N_Y represents the number of controls (non-responders) and N_Z the number of cases (responders).

Applying the normal approximation to the multinomial distribution, we know $\boldsymbol{Y} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$ and $\boldsymbol{Z} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$, where $\xrightarrow{d}$ means "convergence in distribution", the corresponding means and covariances are

$$\begin{aligned} \boldsymbol{\mu}_Y &= (N_Y p_{000}, N_Y p_{001}, N_Y p_{010}, N_Y p_{011})^T \\ \boldsymbol{\mu}_Z &= (N_Z q_{100}, N_Z q_{101}, N_Z q_{110}, N_Z q_{111})^T \end{aligned}$$

$$\boldsymbol{\Sigma}_Y = \begin{pmatrix} N_Y p_{000}(1-p_{000}) & -N_Y p_{000} p_{001} & -N_Y p_{000} p_{010} & -N_Y p_{000} p_{011} \\ -N_Y p_{000} p_{001} & N_Y p_{001}(1-p_{001}) & -N_Y p_{001} p_{010} & -N_Y p_{001} p_{011} \\ -N_Y p_{000} p_{010} & -N_Y p_{001} p_{010} & N_Y p_{010}(1-p_{010}) & -N_Y p_{010} p_{011} \\ -N_Y p_{000} p_{011} & -N_Y p_{001} p_{011} & -N_Y p_{010} p_{011} & N_Y p_{011}(1-p_{011}) \end{pmatrix}$$

$$\boldsymbol{\Sigma}_Z = \begin{pmatrix} N_Z q_{100}(1-q_{100}) & -N_Z q_{100} q_{101} & -N_Z q_{100} q_{110} & -N_Z q_{100} q_{111} \\ -N_Z q_{100} q_{101} & N_Z q_{101}(1-q_{101}) & -N_Z q_{101} q_{110} & -N_Z q_{101} q_{111} \\ -N_Z q_{100} q_{110} & -N_Z q_{101} q_{110} & N_Z q_{110}(1-q_{110}) & -N_Z q_{110} q_{111} \\ -N_Z q_{100} q_{111} & -N_Z q_{101} q_{111} & -N_Z q_{110} q_{111} & N_Z q_{111}(1-q_{111}) \end{pmatrix}$$

Define $\boldsymbol{X} = (\boldsymbol{Y}^T, \boldsymbol{Z}^T)^T$, thus we have $\boldsymbol{X} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\boldsymbol{\mu}_Y^T, \boldsymbol{\mu}_Z^T)^T$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_Y & 0 \\ 0 & \boldsymbol{\Sigma}_Z \end{pmatrix}$$

Define the transformation function $f(\boldsymbol{X}) = \{f_1(\boldsymbol{X}), f_2(\boldsymbol{X})\}^T$, where f_1 is the numerator of the stage 1 test statistic and f_2 the numerator of the stage 2 test statistic. When using a standard interaction test, we have $f_2 = \log(OR_2) = \log(n_{111} n_{100} n_{001} n_{010} / n_{101} n_{110} n_{011} n_{000})$, which is the numerator of a Wald test statistic.

We denote the first-order partial derivatives (the Jacobian matrix) as

$$Df(\boldsymbol{\mu}) = \begin{pmatrix} \frac{\partial f_1}{\partial n_{000}} & \frac{\partial f_2}{\partial n_{000}} \\ \frac{\partial f_1}{\partial n_{001}} & \frac{\partial f_2}{\partial n_{001}} \\ \frac{\partial f_1}{\partial n_{010}} & \frac{\partial f_2}{\partial n_{010}} \\ \frac{\partial f_1}{\partial n_{011}} & \frac{\partial f_2}{\partial n_{011}} \\ \frac{\partial f_1}{\partial n_{100}} & \frac{\partial f_2}{\partial n_{100}} \\ \frac{\partial f_1}{\partial n_{101}} & \frac{\partial f_2}{\partial n_{101}} \\ \frac{\partial f_1}{\partial n_{110}} & \frac{\partial f_2}{\partial n_{110}} \\ \frac{\partial f_1}{\partial n_{111}} & \frac{\partial f_2}{\partial n_{111}} \end{pmatrix} = \begin{pmatrix} df_{11} & df_{12} \\ df_{21} & df_{22} \\ df_{31} & df_{32} \\ df_{41} & df_{42} \\ df_{51} & df_{52} \\ df_{61} & df_{62} \\ df_{71} & df_{72} \\ df_{81} & df_{82} \end{pmatrix}$$

and also

$$Df(\boldsymbol{\mu})^T \boldsymbol{\Sigma} Df(\boldsymbol{\mu}) = \begin{pmatrix} \sigma_{11}^2 & cov_{12} \\ cov_{21} & \sigma_{22}^2 \end{pmatrix}$$

The delta method implies that

$$f(\boldsymbol{X}) = \begin{Bmatrix} f_1(\boldsymbol{X}) \\ f_2(\boldsymbol{X}) \end{Bmatrix} \xrightarrow{d} \mathcal{N}\{f(\boldsymbol{\mu}), Df(\boldsymbol{\mu})^T \boldsymbol{\Sigma} Df(\boldsymbol{\mu})\}$$

The off-diagonal entries of the covariance matrix $Df(\boldsymbol{\mu})^T \boldsymbol{\Sigma} Df(\boldsymbol{\mu})$ can be evaluated by

$$\begin{aligned} cov_{12} \quad = \quad cov_{21} = \quad & (df_{11}a_{11} + df_{21}a_{12} + df_{31}a_{13} + df_{41}a_{14})df_{12} \\ & +(df_{11}a_{12} + df_{21}a_{22} + df_{31}a_{23} + df_{41}a_{24})df_{22} \\ & +(df_{11}a_{13} + df_{21}a_{23} + df_{31}a_{33} + df_{41}a_{34})df_{32} \\ & +(df_{11}a_{14} + df_{21}a_{24} + df_{31}a_{34} + df_{41}a_{44})df_{42} \\ & +(df_{51}b_{11} + df_{61}b_{12} + df_{71}b_{13} + df_{81}b_{14})df_{52} \\ & +(df_{51}b_{12} + df_{61}b_{22} + df_{71}b_{23} + df_{81}b_{24})df_{62} \\ & +(df_{51}b_{13} + df_{61}b_{23} + df_{71}b_{33} + df_{81}b_{34})df_{72} \\ & +(df_{51}b_{14} + df_{61}b_{24} + df_{71}b_{34} + df_{81}b_{44})df_{82} \\ = \quad & (df_{12}a_{11} + df_{22}a_{12} + df_{32}a_{13} + df_{42}a_{14})df_{11} \\ & +(df_{12}a_{12} + df_{22}a_{22} + df_{32}a_{23} + df_{42}a_{24})df_{21} \\ & +(df_{12}a_{13} + df_{22}a_{23} + df_{32}a_{33} + df_{42}a_{34})df_{31} \\ & +(df_{12}a_{14} + df_{22}a_{24} + df_{32}a_{34} + df_{42}a_{44})df_{41} \\ & +(df_{52}b_{11} + df_{62}b_{12} + df_{72}b_{13} + df_{82}b_{14})df_{51} \\ & +(df_{52}b_{12} + df_{62}b_{22} + df_{72}b_{23} + df_{82}b_{24})df_{61} \\ & +(df_{52}b_{13} + df_{62}b_{23} + df_{72}b_{33} + df_{82}b_{34})df_{71} \\ & +(df_{52}b_{14} + df_{62}b_{24} + df_{72}b_{34} + df_{82}b_{44})df_{81} \end{aligned}$$

which is a measure of dependence between the stage 1 and stage 2 test statistics.

$f_1 = \log(OR_1) = \log[(n_{011} + n_{111})(n_{000} + n_{100})/\{(n_{001} + n_{101})(n_{010} + n_{110})\}]$, where the odds ratio OR_1 is the numerator of the Wald test statistic for the gene-environment correlation test against the full data set of cases and controls combined.

The full Jacobian matrix is then

$$Df(\boldsymbol{\mu}) = \begin{pmatrix} df_{11} & df_{12} \\ df_{21} & df_{22} \\ df_{31} & df_{32} \\ df_{41} & df_{42} \\ df_{51} & df_{52} \\ df_{61} & df_{62} \\ df_{71} & df_{72} \\ df_{81} & df_{82} \end{pmatrix} = \begin{pmatrix} \frac{1}{N_Y p_{000} + N_Z q_{100}} & \frac{-1}{N_Y p_{000}} \\ \frac{-1}{N_Y p_{001} + N_Z q_{101}} & \frac{1}{N_Y p_{001}} \\ \frac{-1}{N_Y p_{010} + N_Z q_{110}} & \frac{1}{N_Y p_{010}} \\ \frac{1}{N_Y p_{011} + N_Z q_{111}} & \frac{-1}{N_Y p_{011}} \\ \frac{1}{N_Y p_{000} + N_Z q_{100}} & \frac{1}{N_Z q_{100}} \\ \frac{-1}{N_Y p_{001} + N_Z q_{101}} & \frac{-1}{N_Z q_{101}} \\ \frac{-1}{N_Y p_{010} + N_Z q_{110}} & \frac{-1}{N_Z q_{110}} \\ \frac{1}{N_Y p_{011} + N_Z q_{111}} & \frac{1}{N_Z q_{111}} \end{pmatrix}$$

Following $df_{11} = df_{51}$, $df_{21} = df_{61}$, $df_{31} = df_{71}$, $df_{41} = df_{81}$, we can derive

$$cov_{12} = cov_{21} = -df_{11} + df_{21} + df_{31} - df_{41} + df_{51} - df_{61} - df_{71} + df_{81} = 0$$

These two tests are independent of each other.

A.3 Between-stage independence proof: Dai et al.

The stage 1 marginal association screening model (1.4) and the stage 2 interaction testing model (1.1) are nested, since the latter contains all the terms of the former. Dai et al. proved, in this case, the two stages are independent with each other. Consider the stage 1 marginal association screening test based on the model of the form

$$G\{E(Y_i \mid \boldsymbol{X}_i)\} = \boldsymbol{X}_i^T \boldsymbol{\delta}$$

where $\boldsymbol{\delta}$ is a q-vector. The model underlying the stage 2 standard one-biomarker-at-a-time interaction test is of the form

$$G\{E(Y_i \mid \boldsymbol{V}_i)\} = \boldsymbol{V}_i^T \boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is a p-vector ($p > q$). The above forms ignore intercepts without loss of generality. Notice that $\boldsymbol{V}_i$ includes $\boldsymbol{X}_i$.

Dai et al. [17] showed that the covariance matrix between stage 1 and stage 2 test statistics is $\boldsymbol{A}_1^{-1}\boldsymbol{B}\boldsymbol{A}_2^{-1}$, where

$$\begin{aligned}
\boldsymbol{A}_1 &= E[(\boldsymbol{X}_i\boldsymbol{X}_i^T)\{Y_i - E(Y_i \mid \boldsymbol{X}_i)\}^2] \\
\boldsymbol{B} &= E[(\boldsymbol{X}_i\boldsymbol{V}_i^T)\{Y_i - E(Y_i \mid \boldsymbol{X}_i)\}\{Y_i - E(Y_i \mid \boldsymbol{V}_i)\}] \\
\boldsymbol{A}_2 &= E[(\boldsymbol{V}_i\boldsymbol{V}_j^T)\{Y_i - E(Y_i \mid \boldsymbol{V}_i)\}^2]
\end{aligned}$$

We simplify the expression of $\boldsymbol{B}$ as

$$\begin{aligned}
\boldsymbol{B} &= E[(\boldsymbol{X}_i\boldsymbol{V}_i^T)\{Y_i^2 - Y_iE(Y_i \mid \boldsymbol{X}_i) - Y_iE(Y_i \mid \boldsymbol{V}_i) + E(Y_i \mid \boldsymbol{X}_i)E(Y_i \mid \boldsymbol{V}_i)\}] \\
&= E[(\boldsymbol{X}_i\boldsymbol{V}_i^T)E\{Y_i^2 - Y_iE(Y_i \mid \boldsymbol{X}_i) - Y_iE(Y_i \mid \boldsymbol{V}_i) + E(Y_i \mid \boldsymbol{X}_i)E(Y_i \mid \boldsymbol{V}_i) \mid \boldsymbol{V}_i\}] \\
&= E\{\boldsymbol{X}_i\boldsymbol{V}_i^T var(Y_i \mid \boldsymbol{V}_i)\}
\end{aligned}$$

which uses the law of iterated expectations and the fact that $\boldsymbol{V}_i$ includes $\boldsymbol{X}_i$.

Similarly, we have $\boldsymbol{A}_1 = E\{\boldsymbol{X}_i\boldsymbol{X}_i^T var(Y_i \mid \boldsymbol{X}_i)\}$ and $\boldsymbol{A}_2 = E\{\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T var(Y_i \mid \boldsymbol{V}_i)\}$. Thus,

$$\boldsymbol{A}_1^{-1}\boldsymbol{B}\boldsymbol{A}_2^{-1} = E\{\boldsymbol{X}_i\boldsymbol{X}_i^T var(Y_i \mid \boldsymbol{X}_i)\}^{-1}E\{\boldsymbol{X}_i\boldsymbol{V}_i^T var(Y_i \mid \boldsymbol{V}_i)\}E\{\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T var(Y_i \mid \boldsymbol{V}_i)\}^{-1}$$

Now,

$$\begin{aligned}
\boldsymbol{B}\boldsymbol{A}_2^{-1} &= E\{\boldsymbol{X}_i\boldsymbol{V}_i^T var(Y_i \mid \boldsymbol{V}_i)\}E\{\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T var(Y_i \mid \boldsymbol{V}_i)\}^{-1} \\
&= \begin{Bmatrix}
1 & & & 0 & \cdots & 0 \\
& \cdots & & \vdots & \cdots & \vdots \\
& & 1 & 0 & \cdots & 0
\end{Bmatrix}
\end{aligned}$$

of which the left hand side $q \times q$ block equals to an identity matrix and the right hand side $q \times (p-q)$ block is a matrix of all zeros.

Premultiplying $E\{\boldsymbol{X}_i\boldsymbol{V}_i^T var(Y_i \mid \boldsymbol{V}_i)\}E\{\boldsymbol{V}_{ij}\boldsymbol{V}_{ij}^T var(Y_i \mid \boldsymbol{V}_i)\}^{-1}$ by $E\{\boldsymbol{X}_i\boldsymbol{X}_i^T var(Y_i \mid \boldsymbol{X}_i)\}^{-1}$ completes the covariance matrix. We are interested in the right hand side $q \times (p-q)$ block which corresponds to the covariance between $(\hat{\delta}_1, \ldots, \hat{\delta}_q)$ and $(\hat{\beta}_{q+1}, \ldots, \hat{\beta}_p)$. By the above, this is a block of zeros.

A.4 Discussion of alternative family-wise error rate controlling methods

Detail of alternative family-wise error rate controlling methods, including Šidák correction, Holm-Bonferroni procedure and Hochberg procedure, have been introduced in Section 1.3.

The Šidák correction is only slightly less stringent than the Bonferroni correction as demonstrated in the exampled presented in Section 1.3.

Now we examine the Holm-Bonferroni procedure. We consider the scenario when the false null hypotheses are sparse, e.g. m_0, the number of biomarkers with true biomarker-treatment interactions, is small compared with m. Let m_1 be the number of rejected true null hypotheses, then the total number of rejections $k-1$ yields

$$k-1 \leq m_0 + m_1 \approx m_0 + (m-m_0)\overline{\alpha} = m_0(1-\overline{\alpha}) + m\overline{\alpha} \approx m\overline{\alpha}$$

The first approximate equality sign $\approx$ follows from the fact that m_1 is close to its expectation $E(m_1) = (m-m_0)\overline{\alpha}$ with high possibility, because when the majority of biomarkers are in low linkage disequilibrium with each other, $var(m_1) \approx m_1\overline{\alpha}(1-\overline{\alpha})$ is small relative to m_1. The second $\approx$ follows from the $m_0 \ll m$ assumption. Thus, the hypotheses are essentially compared with the significance levels $\overline{\alpha}/m, \overline{\alpha}/(m-1), \ldots, \overline{\alpha}/(m+1-k) \approx \overline{\alpha}/\{m(1-\overline{\alpha})\}$. The Taylor series of $\overline{\alpha}/\{m(1-\overline{\alpha})\}$ around $\overline{\alpha}=0$ is

$$\frac{\overline{\alpha}}{m(1-\overline{\alpha})} = \frac{\overline{\alpha}}{m} + \frac{\overline{\alpha}^2}{m} + \cdots$$

This again differs from the Bonferroni adjusted significance level $\overline{\alpha}/m$ with $O(\overline{\alpha}^2/m)$.

In a very similar method, the Hochberg procedure, rejection of $H_1, ..., H_k$ is made after finding the maximal index k such that $p_k \leq \overline{\alpha}/(m+1-k)$. This method is more powerful than the Holm-Bonferroni procedure, but requires the hypotheses are independent or under certain forms of positive dependence. In a similar manner to the argument for Holm-Bonferroni, one can demonstrate that the improvement by applying this method is subtle when biomarker-treatment interactions are sparse.

A.5 Proof of independence between stage 1 sparse regression screening and stage 2 standard interaction tests

A.5.1 Proof of Lemma 2.5.2

The proof of Theorem 3 in Fu [28] gave

$$n^{1/2}(\hat{\boldsymbol{\delta}}^{\lambda} - \boldsymbol{\delta}) = -\{\frac{1}{n}\nabla_{\boldsymbol{\delta}\boldsymbol{\delta}^T} L_n(\boldsymbol{\delta}) + \frac{2\lambda_n}{n}\}^{-1}\{\frac{1}{n^{1/2}}\nabla_{\boldsymbol{\delta}} L_n(\boldsymbol{\delta}) + \frac{2\lambda_n \boldsymbol{\delta}}{n^{1/2}}\}$$

for the ridge estimator. Under regularity conditions described in Van der Vaart [75, p. 51-52]: $\nabla_{\boldsymbol{\delta}} L_n(\boldsymbol{\delta})/n^{1/2}$ is asymptotically normal with a mean 0 and a finite variance σ^2 by

the central limit theorem; $\nabla_{\boldsymbol{\delta}\boldsymbol{\delta}^T} L_n(\boldsymbol{\delta})/n$ converges in probability to $\boldsymbol{\Sigma}$ by the law of large numbers. When $\lambda_n = O(n^{1/2})$, $2\lambda_n/n$ vanishes and $2\lambda_n\boldsymbol{\delta}/n^{1/2}$ goes to $2\lambda_0\boldsymbol{\delta}$. Thus,

$$n^{1/2}(\hat{\boldsymbol{\delta}}^{\lambda} - \boldsymbol{\delta}) \xrightarrow{d} \mathcal{N}(-2\lambda_0\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}, \sigma^2\boldsymbol{\Sigma}^{-1})$$

A similar result for ridge regression can be immediately derived by Theorem 2 in Knight et al. [42]. When $\lambda_0 = 0$, this reduces to a well known result for the multivariate regression estimator without regularization

$$n^{1/2}(\hat{\boldsymbol{\delta}}^{0} - \boldsymbol{\delta}) \xrightarrow{d} \mathcal{N}(0, \sigma^2\boldsymbol{\Sigma}^{-1})$$

A.5.2 Proof of Corollary 2.5.2.1

Based on Lemma 2.5.2, we know the distribution of $n^{1/2}(\hat{\boldsymbol{\delta}}^{\lambda} - \boldsymbol{\delta})$ differs from that of $n^{1/2}(\hat{\boldsymbol{\delta}}^{0} - \boldsymbol{\delta})$ asymptotically only with a constant. Along with Theorem 2.5.1, the following holds immediately

$$cov\{n^{1/2}(\hat{\delta}^{\lambda}_{X_j} - \delta_{X_j}), n^{1/2}(\hat{\beta}_{X_j \times T} - \beta_{X_j \times T})\} \xrightarrow{p} 0$$

A.5.3 Proof for the lasso screening test

Theorem A.5.1. *For any $j = 1, \ldots, m$, under standard regularity conditions [75, p. 51-52], if $\lambda_n = \Theta(n^{1/2})$, i.e. $\lim_{n\to\infty} \lambda_n/n^{1/2} = \lambda_0 > 0$, and the true marginal association coefficient is zero, i.e. $\delta_{X_j} = 0$, then the stage 1 lasso variable selection indicator $I(\hat{\delta}^{\lambda}_{X_j} \neq 0)$ is independent of the stage 2 interaction estimator $\hat{\beta}_{X_j \times T}$.*

Proof. By Theorem 2 in [42], if $\lambda_n = \Theta(n^{1/2})$, then

$$\sqrt{n}(\boldsymbol{\delta}^{\lambda} - \boldsymbol{\delta}) \xrightarrow{d} argmin(V)$$

where

$$V(\boldsymbol{u}) = -2\boldsymbol{u}^T\boldsymbol{w} + \boldsymbol{u}^T\boldsymbol{C}\boldsymbol{u} + \lambda_0 \sum_{j=1}^{m} \{u_j sign(\delta_{X_j}) I(\delta_{X_j} \neq 0) + |u_j| I(\delta_{X_j} = 0)\}$$

in which $\boldsymbol{w}$ has a $\mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{C})$ distribution and $\boldsymbol{C} = E(\boldsymbol{X}_i\boldsymbol{X}_i^T)$.

Without loss of generality, suppose that the true marginal effects $\delta_{X_1}, \ldots, \delta_{X_r}$ are all non-zero and $\delta_{X_{r+1}} = \cdots = \delta_{X_m} = 0$. Let

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{pmatrix}$$

where $\boldsymbol{C}_{11}$ is a $r \times r$ matrix, and

$$\boldsymbol{w} = \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix}$$

$$\boldsymbol{u} = \begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{pmatrix}$$

$$\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix}$$

where $\boldsymbol{w}_1$, $\boldsymbol{u}_1$ and $\boldsymbol{\delta}_1$ are all r-vectors.

To assess when the lasso estimator $\hat{\delta}^{\lambda}_{X_j}$ is non-zero if $\delta_{X_j} = 0$, it is easier to consider when $\hat{\delta}^{\lambda}_{X_j}$ is zero, which is equivalent to considering when $\boldsymbol{u}_2 = \boldsymbol{0}$. Knight et al. [42] (the discussion after Theorem 3) showed that $\boldsymbol{u}_2 = \boldsymbol{0}$ {where $V(\boldsymbol{u})$ is minimized} if

$$-\frac{\lambda_0}{2}\boldsymbol{1} \leq \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\{\boldsymbol{w}_1 - \lambda_0 sign(\boldsymbol{\delta}_1)/2\} - \boldsymbol{w}_2 \leq \frac{\lambda_0}{2}\boldsymbol{1} \tag{A.5}$$

Next we will show that $\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\{\boldsymbol{w}_1 - \lambda_0 sign(\boldsymbol{\delta}_1)/2\} - \boldsymbol{w}_2$ is fully determined by the unpenalized estimator $\hat{\boldsymbol{\delta}}^0 = (\hat{\delta}^0_{X_1}, \ldots, \hat{\delta}^0_{X_m})^T$ when $n \to \infty$. Theorem 1 in [42] indicates that $\sqrt{n}(\hat{\boldsymbol{\delta}}^0 - \boldsymbol{\delta}) \xrightarrow{d} \boldsymbol{C}^{-1}\boldsymbol{w}$. Using blockwise matrix inversion, we derive

$$\sqrt{n}(\hat{\boldsymbol{\delta}}_2^0 - \boldsymbol{\delta}_2) \xrightarrow{d} -(\boldsymbol{C}_{22} - \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12})^{-1}(\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{w}_1 - \boldsymbol{w}_2)$$

After reorganizing, we have

$$\begin{aligned}\sqrt{n}(\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12} - \boldsymbol{C}_{22})(\hat{\boldsymbol{\delta}}_2^0 - \boldsymbol{\delta}_2) - \lambda_0\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}sign(\boldsymbol{\delta}_1)/2 \\ \xrightarrow{d} \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\{\boldsymbol{w}_1 - \lambda_0 sign(\boldsymbol{\delta}_1)/2\} - \boldsymbol{w}_2\end{aligned}$$

Substituting $\boldsymbol{\delta}_2 = \boldsymbol{0}$ and $\lim_{n\to\infty} sign(\hat{\boldsymbol{\delta}}_1^0) = sign(\boldsymbol{\delta}_1)$ into the above expression, we derive

$$\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\{\boldsymbol{w}_1 - \lambda_0 sign(\boldsymbol{\delta}_1)/2\} - \boldsymbol{w}_2 = \lim_{n\to\infty} f_n(\hat{\boldsymbol{\delta}}^0)$$

where $f_n(\hat{\boldsymbol{\delta}}^0) = \sqrt{n}(\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12} - \boldsymbol{C}_{22})\hat{\boldsymbol{\delta}}_2^0 - \lambda_0\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}sign(\hat{\boldsymbol{\delta}}_1^0)/2$. Along with condition (A.5), asymptotically (when $n \to \infty$), $\hat{\boldsymbol{\delta}}_2^{\lambda} = 0$ if

$$-\frac{\lambda_0}{2}\boldsymbol{1} \leq f_n(\hat{\boldsymbol{\delta}}^0) \leq \frac{\lambda_0}{2}\boldsymbol{1}$$

This means that $I(\hat{\delta}^{\lambda}_{X_j} = 0)$ is a function of $\hat{\boldsymbol{\delta}}^0$ when $\delta_{X_j} = 0$. Thus $I(\hat{\delta}^{\lambda}_{G_j} \neq 0)$ is also asymptotically a function of $\hat{\boldsymbol{\delta}}^0$. Since the unpenalized estimator $\hat{\boldsymbol{\delta}}^0 = (\hat{\delta}^0_{X_1}, \cdots, \hat{\delta}^0_{X_m})^T$ is independent of the interaction estimator $\hat{\beta}_{X_j \times T}$ by Theorem 2.5.1, $I(\hat{\delta}^{\lambda}_{X_j} \neq 0)$ is also asymptotically independent of $\hat{\beta}_{X_j \times T}$ when $\delta_{X_j} = 0$. Notice that when $\lambda_0 > 0$, there is a positive probability that $\hat{\delta}^{\lambda}_{X_j} = 0$ when $\delta_{X_j} = 0$, which is a desirable property using lasso in practice. □

A.6 Calculating the test statistic gradient for the MGB classifier

The gradient expressed by equation (3.3) can be calculated with cell values $n_{11}, n_{10}, n_{01}, n_{00}$ in Table 3.1 by

$$\begin{aligned}
d_{11} &= \frac{\partial Z}{\partial n_{11}} = \frac{1}{n_{11} SE}\left(1 + \frac{Z}{2 n_{11} SE}\right) \\
d_{10} &= \frac{\partial Z}{\partial n_{10}} = \frac{1}{n_{10} SE}\left(-1 + \frac{Z}{2 n_{10} SE}\right) \\
d_{01} &= \frac{\partial Z}{\partial n_{01}} = \frac{1}{n_{01} SE}\left(-1 + \frac{Z}{2 n_{01} SE}\right) \\
d_{00} &= \frac{\partial Z}{\partial n_{00}} = \frac{1}{n_{00} SE}\left(1 + \frac{Z}{2 n_{00} SE}\right)
\end{aligned}$$

where $SE = \sqrt{n_{11}^{-1} + n_{10}^{-1} + n_{01}^{-1} + n_{00}^{-1}}$.

Let $n = n_{11} + n_{10} + n_{01} + n_{00}$ and assume that $(n_{11}, n_{10}, n_{01}, n_{00})^T$ yields a multinomial distribution $Mult\{(p_{11}, p_{10}, p_{01}, p_{00})^T, n\}$. For any $t, r \in \{0, 1\}$, by law of large numbers, $\lim_{n\to\infty} n_{tr}/n = p_{tr}$. Thus,

$$\lim_{n\to\infty} \frac{d_{tr}}{SE} = \frac{1}{p_{tr}(\frac{1}{p_{11}} + \frac{1}{p_{10}} + \frac{1}{p_{01}} + \frac{1}{p_{00}})} \left\{ 1 + \frac{\log(\frac{p_{11} p_{00}}{p_{10} p_{01}})}{2 p_{tr}(\frac{1}{p_{11}} + \frac{1}{p_{10}} + \frac{1}{p_{01}} + \frac{1}{p_{00}})} \right\}$$

Substituting this result into A_1 and A_2 in Section 3.2.3 shows that both A_1 and A_2 converge to functions of cell probabilities, i.e. they are asymptotically constants, relative to individual probabilities $\hat{p}_{i11}$ and $\hat{p}_{i01}$.

A.7 Extending the MGB classifier

In practice, an alternative way to estimate the test statistic gradient is to add one to each cell of the contingency table and calculate the difference between the current test statistic

and the previous one:

$$\begin{aligned} d_{11} &\approx Z(n_{11}+1, n_{10}, n_{01}, n_{00}) - Z(n_{11}, n_{10}, n_{01}, n_{00}) \\ d_{10} &\approx Z(n_{11}, n_{10}+1, n_{01}, n_{00}) - Z(n_{11}, n_{10}, n_{01}, n_{00}) \\ d_{01} &\approx Z(n_{11}, n_{10}, n_{01}+1, n_{00}) - Z(n_{11}, n_{10}, n_{01}, n_{00}) \\ d_{00} &\approx Z(n_{11}, n_{10}, n_{01}, n_{00}+1) - Z(n_{11}, n_{10}, n_{01}, n_{00}) \end{aligned}$$

This idea can be generalized to any type of outcome wherever a test statistic is calculable. For a test where a statistic is not available, e.g. a Fisher's exact test, minimizing the p-value is an equivalent approach.

A.8 A proposed framework for detecting biomarker-treatment interactions using Bayesian variable selection

From a Bayesian perspective, a natural way to pose the variable selection problem is to define an indicator random variable I_j for each covariate X_{ij}, where $I_j = 1$ indicates the jth covariate is included, otherwise excluded. Thus, the statistical challenge is to estimate the marginal posterior probability of I_j. Variable selection is then achieved by selecting covariates or combinations of covariates with strong selection probabilities. We can regard the number of covariates to be selected into the model as a random variable N_v. Reversible jump MCMC (Markov chain Monte Carlo) is a technique that extends a traditional MCMC algorithm in order to draw posterior samples of potential models along with samples of parameter values. The Metropolis-Hastings acceptance rule is adjusted to allow updating the chain with a change of the model dimension N_v as well as the parameter values at an iteration. The algorithm requires choosing a prior for I_j: The prior distribution on each I_j is $Bernoulli(\pi_j)$ conditional on π_j and a recommended choice of the prior on π_j [55] is

$$\pi_j \sim Beta(1, m)$$

where m is the total number of covariates; this imposes some level of sparseness. When assuming all models of the same dimension are equally likely, this results in the so-called "Beta-Binomial" model space prior on the total number of selected covariates N_v. In contrast to frequentist analyses, Bayesian variable selection provides readily interpretable uncertainty around the selected covariates and combinations of covariates. However, evaluating posterior distributions in Bayesian analyses with the MCMC algorithm is generally more time consuming.

In this section, I propose a framework for detecting biomarker-treatment interactions using Bayesian variable selection. The proposal was inspired by the case-only interaction tests, which have been introduced in Section 1.2.2.1 and Appendix A.1.

I start my derivation by assuming models in responders and non-responders as

$$logit\{E(T_i \mid X_{ij}, R_i = 1)\} = \gamma_{0re} + \sum_{j=1}^{m} \gamma_{Xre_j} X_{ij} \tag{A.6}$$

$$logit\{E(T_i \mid X_{ij}, R_i = 0)\} = \gamma_{0nre} + \sum_{j=1}^{m} \gamma_{Xnre_j} X_{ij} \tag{A.7}$$

Following a similar discussion in Appendix A.1 {equation (A.1)}, I have

$$\exp(\beta_{X_j \times T}) = \frac{\exp(\gamma_{Xre_j})}{\exp(\gamma_{Xnre_j})}$$

where $\beta_{X_j \times T}$ is the interaction effect coefficient in model (1.1). Thus,

$$\beta_{X_j \times T} = \gamma_{Xre_j} - \gamma_{Xnre_j}$$

By applying Bayesian variable selection to models (A.6) and (A.7), I can obtain the estimated posterior probabilities $pr(\gamma_{Xre_j} \neq 0 \mid Data)$ and $pr(\gamma_{Xnre_j} \neq 0 \mid Data)$. Next, I calculate $pr(\beta_{X_j \times T} \neq 0 \mid Data)$ as

$$\begin{aligned}
& pr(\beta_{X_j \times T} \neq 0 \mid Data) \\
= \;& 1 - pr(\beta_{X_j \times T} = 0 \mid Data) \\
= \;& 1 - pr(\gamma_{Xre_j} = \gamma_{Xnre_j} \mid Data) \\
= \;& 1 - \{pr(\gamma_{Xre_j} = \gamma_{Xnre_j} = 0 \mid Data) + pr(\gamma_{Xre_j} = \gamma_{Xnre_j} \neq 0 \mid Data)\} \\
= \;& 1 - pr(\gamma_{Xre_j} = \gamma_{Xnre_j} = 0 \mid Data) \\
= \;& pr(\gamma_{Xre_j} \neq 0 \text{ or } \gamma_{Xnre_j} \neq 0 \mid Data) \\
= \;& pr(\gamma_{Xre_j} \neq 0 \mid Data) + pr(\gamma_{Xnre_j} \neq 0 \mid Data) \\
& - pr(\gamma_{Xre_j} \neq 0 \mid Data) pr(\gamma_{Xnre_j} \neq 0 \mid Data)
\end{aligned}$$

which uses $pr(\gamma_{Xre_j} = \gamma_{Xnre_j} \neq 0 \mid Data) \approx 0$ under biomarker-treatment independence dictated by randomization ($\gamma_{Xre_j} = \gamma_{Xnre_j} \neq 0$ implies that there exists non-zero biomarker-treatment association within the whole sample). This provides a way to use pre-existing main-effect Bayesian variable selection algorithms for detecting interactions in randomized clinical trials, without having to adapt the algorithms to search over the much wider model space of main effects and interaction effect terms. This framework is applicable under arbitrary response rates, and it only assumes the biomarker-treatment independence

condition in the context of randomized clinical trials.

There are three options to use the above procedure in the ASD framework:

1. Use it in stage 1 to select biomarkers with posterior selection probabilities larger than a threshold μ. The inner cross-validation then needs to tune μ.

2. Use it in stage 1 to select biomarkers with the combination of biomarkers having the largest posterior selection probability. In this way, there is not any parameter for stage 1 to be tuned by cross-validation.

3. Use it in stage 1 to obtain samples of coefficient values and then use these samples to estimate relevant stage 2 quantities (e.g. the odds ratio, risk difference or expected test statistic change) for building the classifier. For example, we need to evaluate the following probabilities for the kth stage 2 patient:

$$pr(R_k = 1 \mid \boldsymbol{X}_k, Data) = \int pr(R_k = 1 \mid \boldsymbol{X}_k, \boldsymbol{\beta}) pr(\boldsymbol{\beta} \mid Data) d\boldsymbol{\beta}$$

where $\boldsymbol{X}_k = (T_k, X_{k1}, \ldots, X_{km})^T$, $\boldsymbol{\beta} = (\beta_0, \beta_T, \beta_{X_1}, \ldots, \beta_{X_m}, \beta_{X_1 \times T}, \ldots, \beta_{X_m \times T})^T$ and $Data$ representing stage 1 data. In practice, main-effect coefficients, $\beta_{X_1}, \ldots, \beta_{X_m}$, are omitted as recommended by [11]. In Appendix C.2, I demonstrate this approach in a limited set of simulations.

A.9 Asymptotic score functions for logistic regression

A.9.1 Approximating the sigmoid function

Lemma A.9.1. *If $x \sim \mathcal{N}(\mu, \sigma^2)$, then*

$$E\{S(x)\} \approx S(\frac{\mu}{\sqrt{1 + \xi^2 \sigma^2}}) = S\{\frac{E(x)}{\sqrt{1 + \xi^2 \sigma^2}}\}$$

where $\xi^2 = \pi/8$.

Proof. By definition, we have

$$E\{S(x)\} = \int_{-\infty}^{\infty} S(x) \mathcal{N}(x \mid \mu, \sigma^2) dx = \int_{-\infty}^{\infty} \frac{1}{1 + \exp(-x)} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x - \mu)^2}{2\sigma^2}\} dx$$

which does not have an analytic expression. However, we can approximate the sigmoid function S with a probit link function Φ, which is the cumulative distribution function of

a standard normal distribution [6].

$$S(x) \approx \Phi(\xi x)$$

where $\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(y \mid 0, 1)dy$. ξ is a parameter used in the approximating function of the sigmoid function. I use $\xi^2 = \pi/8$ in the simulation studies. Therefore,

$$\begin{aligned} E\{S(x)\} &= \int_{-\infty}^{\infty} S(x)\mathcal{N}(x \mid \mu, \sigma^2)dx \\ &\approx \int_{-\infty}^{\infty} \Phi(\xi x)\mathcal{N}(x \mid \mu, \sigma^2)dx \\ &= \Phi(\frac{\xi\mu}{\sqrt{1+\xi^2\sigma^2}}) \\ &\approx S(\frac{\mu}{\sqrt{1+\xi^2\sigma^2}}) \end{aligned}$$

which uses the fact that $\int_{-\infty}^{\infty} \Phi(x)\mathcal{N}(\mu, \sigma^2)dx = \Phi(\mu/\sqrt{1+\sigma^2})$. □

Lemma A.9.2. *If* $x \sim \mathcal{N}(0, 1)$, *then*

$$E\{xS(\mu+\sigma x)\} \approx \frac{\sigma}{\sqrt{1+\xi^2\sigma)}} S'(\frac{\mu}{\sqrt{1+\xi^2\sigma^2}}) = \frac{E\{x(\mu+\sigma x)\}}{\sqrt{1+\xi^2\sigma^2}} S'(\frac{\mu}{\sqrt{1+\xi^2\sigma^2}})$$

where $\xi^2 = \pi/8$.

Proof. By definition and Lemma A.9.1, we have

$$\begin{aligned} & E\{xS(\mu+\sigma x)\} \\ = & \int_{-\infty}^{\infty} xS(\mu+\sigma x)\mathcal{N}(x \mid 0, 1)dx \\ \approx & \int_{-\infty}^{\infty} x\Phi\{\xi(\mu+\sigma x)\}\mathcal{N}(x \mid 0, 1)dx \\ = & \int_{-\infty}^{\infty} x\Phi(a+bx)\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \end{aligned}$$

where $\xi^2 = \pi/8$, $a = \xi\mu$ and $b = \xi\sigma$. Next, we use integration by substitution

$$\begin{aligned} & E\{xS(\mu+\sigma x)\} \\ \approx & \int_{-\infty}^{\infty} x\Phi(a+bx)\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \\ = & -\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \Phi(a+bx)d\{\exp(-\frac{x^2}{2})\} \\ = & -\frac{1}{\sqrt{2\pi}}\{\exp(-\frac{x^2}{2})\Phi(a+bx)|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \exp(-\frac{x^2}{2})d\Phi(a+bx)\} \end{aligned}$$

where the first additive term is 0. Thus, we have

$$\begin{aligned}
& E\{xS(\mu+\sigma x)\} \\
\approx\ & \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp(-\frac{x^2}{2})d\Phi(a+bx) \\
=\ & \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp(-\frac{x^2}{2})d\int_{-\infty}^{a+bx}\frac{1}{\sqrt{2\pi}}\exp(-\frac{y^2}{2})dy \\
=\ & \frac{b}{2\pi}\int_{-\infty}^{\infty}\exp(-\frac{x^2}{2})\exp\{-\frac{(a+bx)^2}{2}\}dx
\end{aligned}$$

We complete the square in this Gaussian integral as follows

$$\begin{aligned}
& E\{xS(\mu+\sigma x)\} \\
\approx\ & \frac{b}{2\pi}\int_{-\infty}^{\infty}\exp\{-\frac{(1+b^2)x^2+2abx+(\frac{ab}{\sqrt{1+b^2}})^2-(\frac{ab}{\sqrt{1+b^2}})^2+a^2}{2}\}dx \\
=\ & \frac{b}{2\pi}\int_{-\infty}^{\infty}\exp\{-\frac{(\sqrt{1+b^2}x+\frac{ab}{\sqrt{1+b^2}})^2}{2}\}\exp\{-\frac{a^2}{2(1+b^2)}\}dx \\
=\ & \frac{b}{\sqrt{2\pi(1+b^2)}}\exp\{-\frac{a^2}{2(1+b^2)}\}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi(1+b^2)^{-1}}}\exp\{-\frac{(x+\frac{ab}{1+b^2})^2}{2(1+b^2)^{-1}}\}dx
\end{aligned}$$

The function within the integral in the above formula is in fact a probability density function of a normal distribution $\mathcal{N}\{x \mid -ab/(1+b^2),(1+b^2)^{-1}\}$, the integral of which evaluates to 1. Thus, we have

$$\begin{aligned}
& E\{xS(\mu+\sigma x)\} \\
\approx\ & \frac{b}{\sqrt{2\pi(1+b^2)}}\exp\{-\frac{a^2}{2(1+b^2)}\} \\
=\ & \frac{b}{\sqrt{1+b^2}}\Phi'(\frac{a}{\sqrt{1+b^2}}) \\
=\ & \frac{\xi\sigma}{\sqrt{1+\xi^2\sigma^2}}\Phi'(\frac{\xi\mu}{\sqrt{1+\xi^2\sigma^2}}) \\
\approx\ & \frac{\sigma}{\sqrt{1+\xi^2\sigma^2}}S'(\frac{\mu}{\sqrt{1+\xi^2\sigma^2}}) \\
=\ & \frac{\sigma}{\sqrt{1+\pi\sigma^2/8}}S'(\frac{\mu}{\sqrt{1+\pi\sigma^2/8}}) \\
=\ & \frac{E\{x(\mu+\sigma x)\}}{\sqrt{1+\pi\sigma^2/8}}S'(\frac{\mu}{\sqrt{1+\pi\sigma^2/8}})
\end{aligned}$$

which uses $\Phi'(x)=d\Phi(x)/dx=\exp(-x^2/2)/\sqrt{2\pi}$ and $S'(x)=dS(x)/dx=S(x)\{1-S(x)\}$. This completes my proof. □

A.9.2 Asymptotic score functions

Let us consider the asymptotic limit of the four score equations (4.7). The first equation of this system is

$$E\{S(\boldsymbol{X}_i^T\boldsymbol{\beta}) - S(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)\} = 0$$

By Lemma A.9.1, we have

$$S\{\frac{E(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\sqrt{1+\xi^2 var(\boldsymbol{X}_i^T\boldsymbol{\beta})}}\} - S\{\frac{E(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)}{\sqrt{1+\xi^2 var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)}}\} = 0$$

which is equivalent to

$$\frac{E(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\sqrt{1+\xi^2 var(\boldsymbol{X}_i^T\boldsymbol{\beta})}} - \frac{E(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)}{\sqrt{1+\xi^2 var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)}} = 0$$

The second equation of (4.7) is

$$E\{X_{ij}S(\boldsymbol{X}_i^T\boldsymbol{\beta}) - X_{ij}S(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)\} = 0$$

Assuming $X_{ij} \sim \mathcal{N}(0,1)$, by Lemma A.9.2, we have

$$\frac{E(X_{ij}\boldsymbol{X}_i^T\boldsymbol{\beta})}{\sqrt{1+\xi^2 var(\boldsymbol{X}_i^T\boldsymbol{\beta})}} S'\{\frac{E(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\sqrt{1+\xi^2 var(\boldsymbol{X}_i^T\boldsymbol{\beta})}}\} -$$
$$\frac{E(X_{ij}\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)}{\sqrt{1+\xi^2 var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)}} S'\{\frac{E(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)}{\sqrt{1+\xi^2 var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)}}\} = 0$$

Applying the result of the first equation, we have

$$\frac{E(X_{ij}\boldsymbol{X}_i^T\boldsymbol{\beta})}{\sqrt{1+\xi^2 var(\boldsymbol{X}_i^T\boldsymbol{\beta})}} - \frac{E(X_{ij}\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)}{\sqrt{1+\xi^2 var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)}} = 0$$

The third equation of the system is

$$E\{T_i S(\boldsymbol{X}_i^T\boldsymbol{\beta}) - T_i S(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)\} = 0$$

Assuming $T_i \in \{0,1\}$, by Lemma A.9.1, we have

$$\frac{E(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1)}{\sqrt{1+\xi^2 var(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1)}} - \frac{E(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1)}{\sqrt{1+\xi^2 var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1)}} = 0$$

The fourth equation is

$$E\{X_{ij}T_iS(\boldsymbol{X}_i^T\boldsymbol{\beta}) - X_{ij}T_iS(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)\} = 0$$

With $T_i \in \{0, 1\}$, by Lemma A.9.2, we have

$$\frac{E(X_{ij}\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1)}{\sqrt{1 + \xi^2 var(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1)}} S'\{\frac{E(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1)}{\sqrt{1 + \xi^2 var(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1)}}\} -$$
$$\frac{E(X_{ij}\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1)}{\sqrt{1 + \xi^2 var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1)}} S'\{\frac{E(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1)}{\sqrt{1 + \xi^2 var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1)}}\} = 0$$

Applying the result of the third equation, we have

$$\frac{E(X_{ij}\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1)}{\sqrt{1 + \xi^2 var(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1)}} - \frac{E(X_{ij}\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1)}{\sqrt{1 + \xi^2 var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1)}} = 0$$

To summarize the results of all the four equations, we have

$$\begin{aligned}
&rE(\boldsymbol{X}_i^T\boldsymbol{\beta}) - E(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j) = 0 \\
&rE(X_{ij}\boldsymbol{X}_i^T\boldsymbol{\beta}) - E(X_{ij}\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j) = 0 \\
&r_TE(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1) - E(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1) = 0 \\
&r_TE(X_{ij}\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1) - E(X_{ij}\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1) = 0
\end{aligned}$$

of which r and r_T are defined as

$$\begin{aligned}
r &= \sqrt{\frac{1 + \xi^2 var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j)}{1 + \xi^2 var(\boldsymbol{X}_i^T\boldsymbol{\beta})}} \\
r_T &= \sqrt{\frac{1 + \xi^2 var(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1)}{1 + \xi^2 var(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1)}}
\end{aligned}$$

where $\xi^2 = \pi/8$ is a parameter used in the approximating function of the sigmoid function. This is the result shown in Section 4.4.

Notice the asymptotic limit of score equations for a linear regression model is

$$\begin{aligned}
&E(\boldsymbol{X}_i^T\boldsymbol{\beta}) - E(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j) = 0 \\
&E(X_{ij}\boldsymbol{X}_i^T\boldsymbol{\beta}) - E(X_{ij}\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j) = 0 \\
&E(\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1) - E(\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1) = 0 \\
&E(X_{ij}\boldsymbol{X}_i^T\boldsymbol{\beta} \mid T_i = 1) - E(X_{ij}\boldsymbol{V}_{ij}^T\boldsymbol{\delta}_j \mid T_i = 1) = 0
\end{aligned}$$

which differ from the asymptotic score equation system for a logistic regression model in the ratios r and r_T.

A.10 De-biased biomarker-treatment interaction estimator under alternative treatment coding

In models (4.5) and (4.6), I assume treatment $T_i \in \{0, 1\}$. Alternatively, I may encode treatment as $T_i \in \{-0.5, 0.5\}$, i.e. -0.5 for the control arm and 0.5 for the experiment arm. Then, solving equation (4.7) gives the following form of the de-biased interaction estimator

$$\widetilde{\beta}_{X_j \times T} = \frac{(\hat{r}_T^{-1} - \hat{r}^{-1})\hat{\delta}_{X_j} + \{0.5\hat{r}_T^{-1} - (\hat{p}_T - 0.5)\hat{r}^{-1}\}\hat{\delta}_{X_j \times T}}{1 - \hat{p}_T}$$

where

$$\begin{aligned} \hat{r} &= \sqrt{\frac{1 + \xi^2 \hat{var}(\boldsymbol{V}_{ij}^T \boldsymbol{\delta}_j)}{1 + \xi^2 \hat{var}(\boldsymbol{X}_i^T \boldsymbol{\beta})}} \\ \hat{r}_T &= \sqrt{\frac{1 + \xi^2 \hat{var}(\boldsymbol{V}_{ij}^T \boldsymbol{\delta}_j \mid T_i = 0.5)}{1 + \xi^2 \hat{var}(\boldsymbol{X}_i^T \boldsymbol{\beta} \mid T_i = 0.5)}} \end{aligned}$$

The variance of the estimator $\widetilde{\beta}_{X_j \times T}$ can be obtained as

$$\begin{aligned} & var(\widetilde{\beta}_{X_j \times T}) \\ = \; & [(\hat{r}_T^{-1} - \hat{r}^{-1})^2 var(\hat{\delta}_{X_j}) + \{0.5\hat{r}_T^{-1} - (\hat{p}_T - 0.5)\hat{r}^{-1}\}^2 var(\hat{\delta}_{X_j \times T}) \\ & + 2(\hat{r}_T^{-1} - \hat{r}^{-1})\{0.5\hat{r}_T - (\hat{p}_T - 0.5)\hat{r}^{-1}\} cov(\hat{\delta}_{X_j}, \hat{\delta}_{X_j \times T})]/(1 - \hat{p}_T)^2 \end{aligned}$$

Appendix B

Numerical analysis

B.1 Relationship between the MRD and MGB classifiers

We conduct a numerical study of the value of $A_1 = (\hat{d}_{11} - \hat{d}_{10})/(\hat{d}_{00} - \hat{d}_{01})$ defined in Section 3.2.3. Considering the contingency Table 3.1, we assume $n_{11}+n_{10} = n_{01}+n_{00} = 500$, which corresponds to a sample size of $1,000$ and a new treatment assignment probability of 0.5. We varied n_{11} and n_{01} from 1 to 499, which corresponds to 499×499 different contingency tables of cell values. A_1 was calculated for each table. The numerical result is shown in Figure B.1.

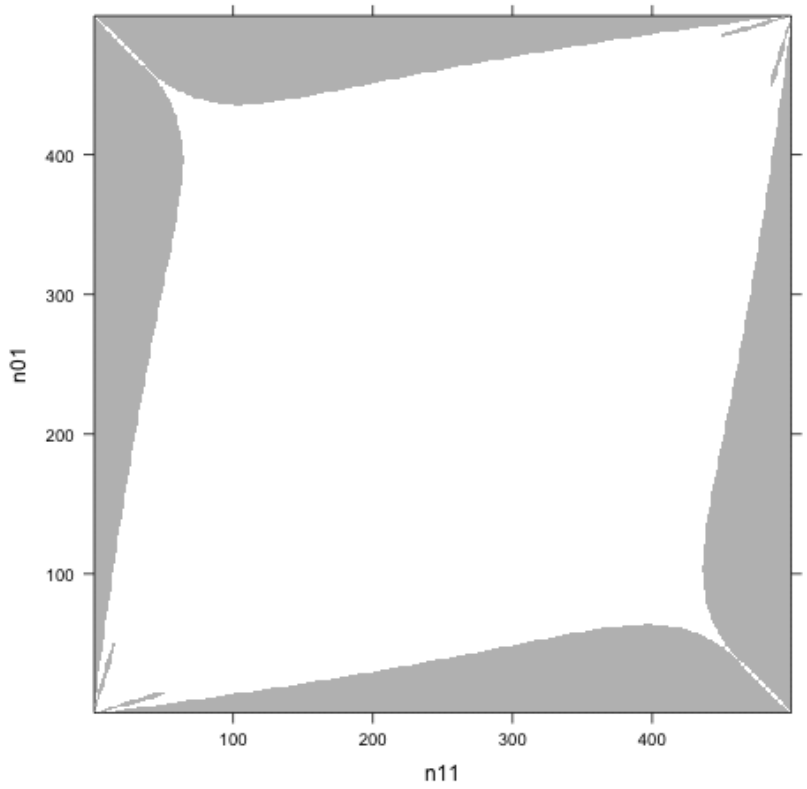

Figure B.1: Color map of A_1 under n_{11} and n_{01} of different values from 1 to 499: grey if $A_1 > 3/2$ or $A_1 < 2/3$; white otherwise

Figure B.1 shows that A_1 is away from 1 only for extreme cell values. Let us examine what happens in the two extreme areas in Figure B.1: the bottom-right corner and the top-left corner. In the bottom-right area, n_{11} is large and n_{01} is small, which means the new treatment is much more effective than the control. The overall test will be significant regardless of the subgroup test result. In the top-left area, the control treatment is much more effective than the new, which is unlikely to happen if the new treatment has entered a phase III trial. In sum, for various plausible scenarios, the value of A_1 is around 1. In addition, $A_4 = (1 - \hat{p}_t)/\hat{p}_t$ is asymptotically 1 when $p_t = 0.5$. Thus, we draw our conclusion $A_1 \approx A_4$, which indicates the risk difference classifier can perform similarly to the gradient-based classifier across various scenarios in practice.

B.2 Relationship between the ASD and MRD classifiers

We examine the odds ratio classifier used by the original ASD. Assuming $pr(T_i) = 0.5$, for a fixed risk difference $pr(R_i = 1 \mid T_i = 1) - pr(R_i = 1 \mid T_i = 0) = 0.2$, the odds ratio can

vary, e.g. when $p_{i11} = pr(T_i = 1, R_i = 1)$ is varying.

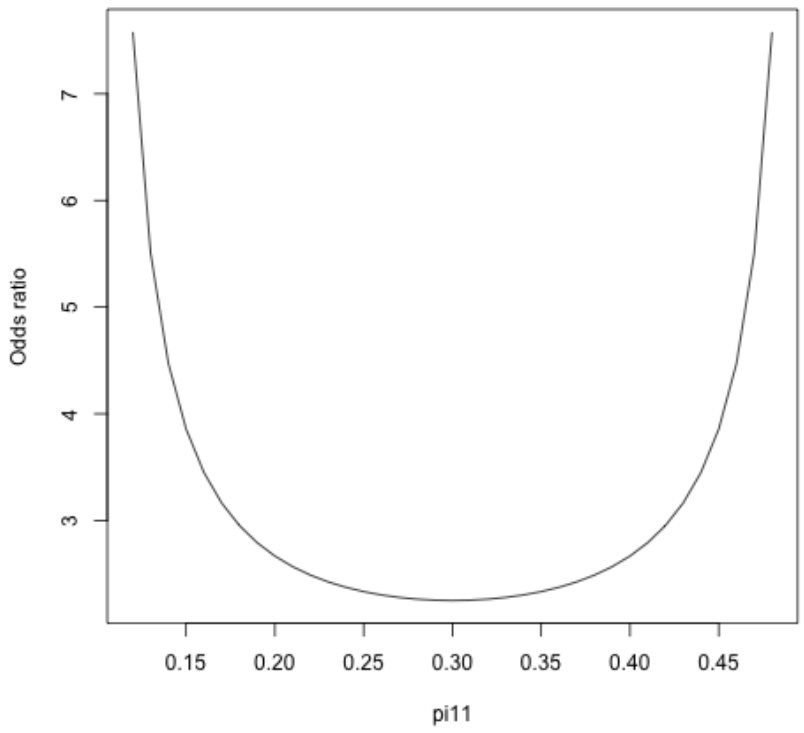

Figure B.2: The odd ratio varies with $p_{i11} = pr(T_i = 1, R_i = 1)$ for a fixed risk difference of 0.2.

The result is shown in Figure B.2. This demonstrates that if there is an optimal risk difference threshold (e.g. 0.2 in the above case), it is usually impossible to find an odds ratio threshold identifying the same subgroup.

Appendix C

Additional simulation results

C.1 Sparse regression screening procedures

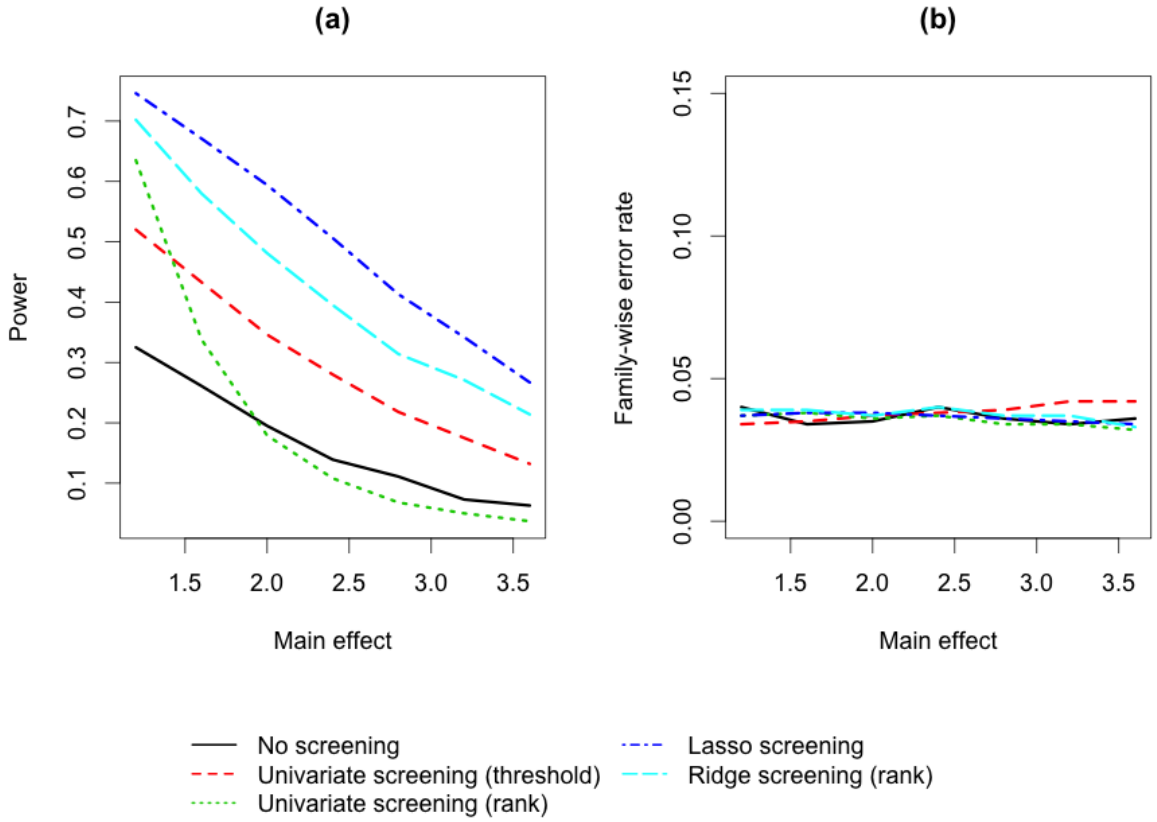

Figure C.1: Comparison of two-stage interaction tests with different screening testing procedures. The two panels represent: (a) power, changing the main effects of the four biomarkers $\beta_{X_{21}}, \beta_{X_{41}}, \beta_{X_{61}}, \beta_{X_{81}}$, (b) family-wise error rate, changing the main effects of the four biomarkers $\beta_{X_{21}}, \beta_{X_{41}}, \beta_{X_{61}}, \beta_{X_{81}}$.

In Figure C.1, we used the scenario with one biomarker having an interaction (biomarker-biomarker correlation $\rho = 0.6$) described in Section 2.7 as the base, and changed only the main effects of the four biomarkers with main effects alone $\beta_{X_{21}}, \beta_{X_{41}}, \beta_{X_{61}}, \beta_{X_{81}}$. Figure C.1 shows that the performance of the univariate screening using weighted hypothesis testing downgrades when the main effects of these four biomarkers become too large. This is because more noise biomarkers with marginal signals tend to fall into the top buckets (indeed, in the simulated scenario, 100 biomarkers have true "univariate" marginal signals). The ridge regression screening strategy does not suffer this issue as much fewer biomarkers (five including $X_1, X_{21}, X_{41}, X_{61}, X_{81}$) have true "multivariate" marginal signals.

Next, we applied Nyholt's method [56] and Gao's method [29] to the base scenario we described in Section 2.7 Simulation Study with a reduced number of repetitions (100). For the single-stage approach ("no screening"), we used the obtained "effective" number of independent tests, m_{eff}, in the Bonferroni correction to adjust for multiple testing. For the two-stage approaches using weighted hypothesis testing (two "rank" approaches), we first adjusted the significance level $\overline{\alpha}$ with $\overline{\alpha}m/m_{eff}$, then applied the weighting scheme as described in Section 1.2.2.2. Results are shown in Figure C.2.

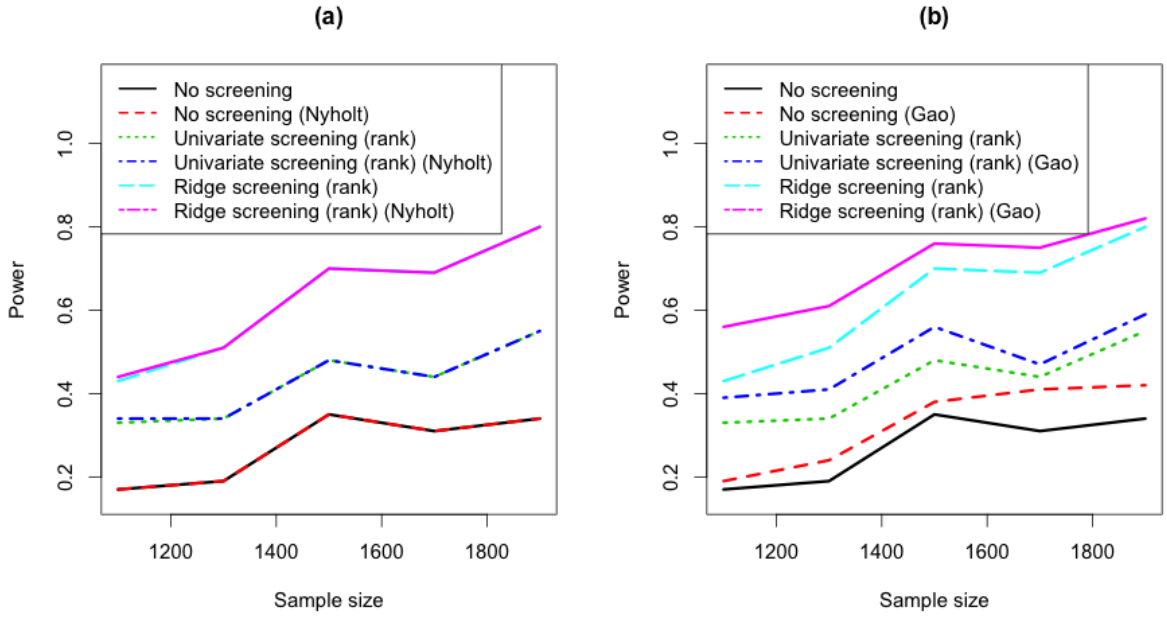

Figure C.2: Comparison of two-stage interaction tests with different screening testing procedures: The two panels represent: (a) power, comparing with Nyholt's multiple correction method, (b) power, comparing with Gao's multiple correction method.

In Figure C.2(a), Nyholt's method appeared to be too conservative to improve power for interaction discovery in the simulated scenario. With a closer look at several repetitions,

we found the obtained m_{eff} is around 992, which is not far away from $m = 1,000$. Some further experiments showed that Nyholt's method could only show moderate improvement for extremely highly correlated data ($\rho \geq 0.9$). In Figure C.2(b), we applied Gao's method with a PCA (principal component analysis) percentage cutoff $C = 95\%$ instead of the recommended 99.5%, as the latter also appeared to be too conservative. Gao's method performed better than Nyholt's method and showed moderate improvement in the simulated scenario (the improvement is still much smaller than incorporating the ridge screening to account for biomarker-biomarker correlations). However, we found the family-wise error rates exceeded above the targeted 0.05 (inflated to around 0.06 to 0.10) for the three methods shown in Figure C.2(b) after applying Gao's adjustment, which indicated $C = 95\%$ could be liberal.

C.2 The MGB classifier using Bayesian variable selection

Additional simulations were done to demonstrate my proposed framework in Appendix A.8 using Bayesian variable selection with the ASD. The R package **R2BGLiMS** [55] was used to run reversible jump MCMC. Simulation settings are the same as those described in Section 3.3 except that there exist non-zero correlations ($\rho = 0.6$) among biomarkers and the number of replicated data sets is 100. Figure C.3 showed that, in the scenarios with highly correlated biomarkers, the MGB design using Bayesian variable selection provided better power than the traditional ASD. However, running the reversible jump MCMC algorithm is time-consuming, especially within the cross-validated ASD framework. I may be interested in how to leverage parallel resources to speed up by using "population" based MCMC methods [5].

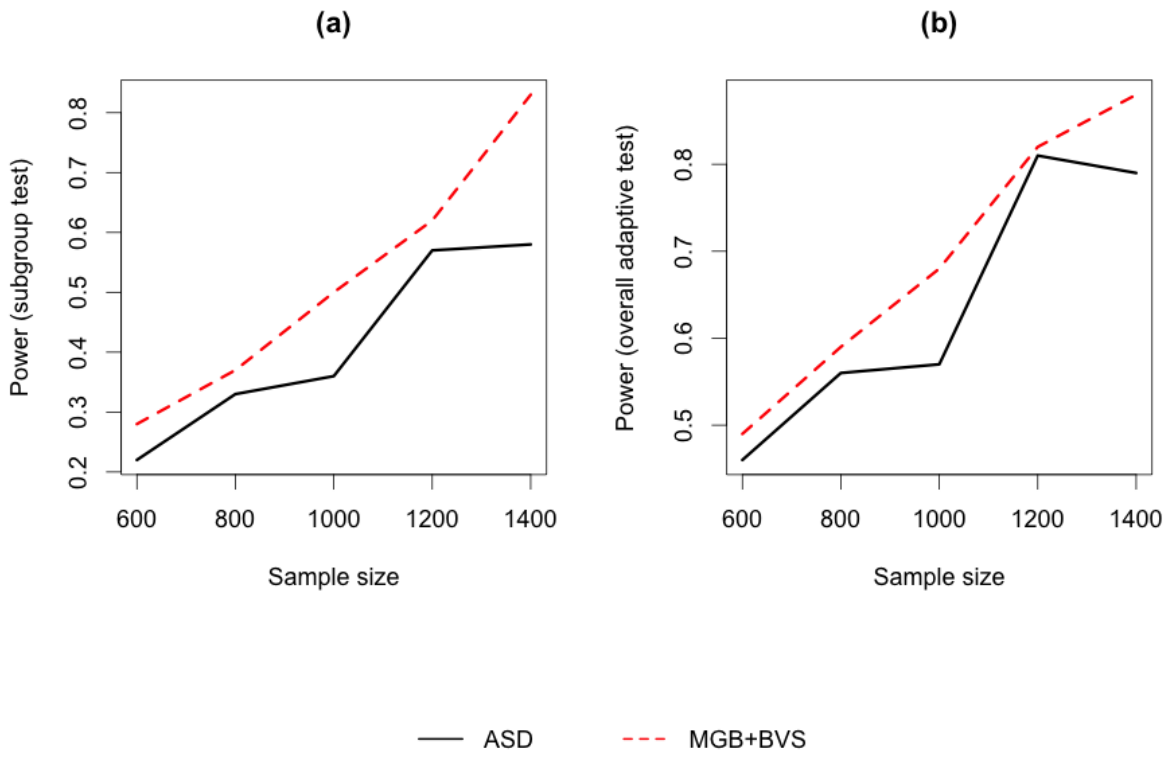

Figure C.3: Comparison of adaptive signature designs in simulated data. The two panels represent: (a) the subgroup 0.01-level test with increasing sample size, (b) the overall adaptive test (the subgroup 0.01-level test and the overall 0.04-level test) with increasing sample size.

C.3 De-biasing procedures under alternative treatment coding

Throughout Chapter 4, we have assumed treatment $T_i \in \{0, 1\}$. In this section, we encoded treatment as $T_i \in \{-0.5, 0.5\}$ and repeated simulations shown in Figure 4.1(a) and (b). When a $\{0, 1\}$ treatment coding is applied, in comparison with a $\{-0.5, 0.5\}$ coding, the variance in control arms decreases and the variance in experimental arms in the model increases. This will have an impact on fitting the saturated model (4.3) using penalized regression.

Table C.1: Comparison of different treatment coding schemes when using the lasso de-biasing procedure in simulated data.

Treatment coded as $T_i \in \{0, 1\}$					
Sample size	$1,100$	$1,300$	$1,500$	$1,700$	$1,900$
Power	0.508	0.667	0.753	0.841	0.900
Family-wise error rate	0.049	0.047	0.031	0.056	0.046
Treatment coded as $T_i \in \{-0.5, 0.5\}$					
Sample size	$1,100$	$1,300$	$1,500$	$1,700$	$1,900$
Power	0.514	0.670	0.756	0.844	0.902
Family-wise error rate	0.048	0.047	0.031	0.058	0.046

Table C.1 showed that the $\{-0.5, 0.5\}$ treatment coding does increase power of the group lasso de-biasing procedure compared with the $\{0, 1\}$ treatment coding, although the benefit is small (< 0.01) in our simulations.