Biometrics 000, 1–8 DOI: 000
Submitted in March 2020

Two-Stage Sparse Regression Screening to Detect Biomarker-Treatment Interactions in Randomized Clinical Trials

Jixiong Wang1,* , Ashish Patel1,** , James M.S. Wason1,2,*** , and Paul J. Newcombe1,****

1MRC Biostatistics Unit, University of Cambridge, Cambridge CB2 0SR, U.K.
2Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne NE2 4BN, U.K.
*email: jixiong.wang@mrc-bsu.cam.ac.uk
**email: ashish.patel@mrc-bsu.cam.ac.uk
***email: james.wason@mrc-bsu.cam.ac.uk
****email: paul.newcombe@mrc-bsu.cam.ac.uk

SUMMARY: 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. We adapt recently proposed two-stage interaction detecting procedures in the setting of randomized clinical trials. We also propose a new stage 1 multivariate screening strategy using ridge regression to account for correlations among biomarkers. For this multivariate screening, we prove the asymptotic between-stage independence, required for the family-wise error rate control, under the biomarker-treatment independence. Simulation results show that in various scenarios, the ridge regression screening procedure can provide substantially greater power than the traditional one-biomarker-at-a-time screening procedure in highly correlated data. We also exemplify our approach in two real clinical trial data applications.

KEY WORDS: Biomarker; Clinical trial; Interaction; Randomization; Ridge regression; Two-stage.

1. Introduction
Recent developments in medicine have seen a shift toward targeted therapeutics (Collins and Varmus, 2015). It has been shown that individual variability can often contribute to differences in response to the same treatment (Khoury et al., 2012). For example, a strong association of carbamazepine-induced Stevens-Johnsons syndrome and the human leukocyte antigen-B (HLA-B)*1502 allele was reported (Lochner et al., 2011). Another example is the Class II allele HLA-DRB1*07:01 that has been associated with lapatinib-induced liver injury (Parham et al., 2016). Detecting such interactions between biomarkers and treatments in randomized clinical trials is of growing interest.

Discovering biomarker-treatment interactions helps identify baseline predictive biomarkers in clinical trials (Robert et al., 2014): biomarkers which influence treatment efficacy can be used to find the subgroup of patients who are most likely to benefit from the new treatment, as well as being used to predict subgroup treatment effects (Wang and Dai, 2016). Consequently, new adaptive design approaches can be used in settings where there are genetically-driven subgroups to improve efficiency (Wason et al., 2015). 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” (Collins and Varmus, 2015).

Detecting biomarker-treatment interactions in large-scale studies of human populations is a non-trivial task, which faces several challenging problems (McAllister et al., 2017). 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.

In the 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 (Piegorsch et al., 1994), empirical Bayes (Mukherjee and Chatterjee, 2008), Bayesian model averaging (Li and Conti, 2008), and two-stage tests with different screening procedures (Kooperberg and LeBlanc, 2008; Murray et al., 2008; Hsu et al., 2012; Gauderman et al., 2013; Wason and Dudbridge, 2012). 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 paper, we will examine how existing gene-environment interaction testing methods may be modified so that they are transferable to the biomarker-treatment setting. Another significant drawback of the tradi-
tional approach testing each biomarker one at a time is that it fails to model correlations between biomarkers. We also propose a novel screening test in two-stage approaches detecting biomarker-treatment interactions, which utilizes ridge regression to model correlated high-dimensional data. We prove this new two-stage method is 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 method can provide better performance than traditional one-biomarker-at-a-time approaches for correlated biomarkers. In the generic context of variable selection, screening strategies have been explored to focus algorithms on a reduced search space \cite{Fan and Lv 2008} \cite{Wang and Leng 2016}. In this work, we explore the use of variable pre-screening specifically to help identify interactions.

2. Methods

2.1 Standard Single-Step 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

\[
G\{E(Y_i \mid X_{ij}, T_i)\} = \beta_0 + \beta_{X_j} X_{ij} + \beta_T T_i + \beta_{X_j \times T} X_{ij} \times T_i
\]

with $Y_i$ denoting the response outcome, $T_i$ the binary treatment-control indicator, and $X_{ij}, \ldots, X_{im}$ representing $m$ biomarkers, for the $i$th patient. $G$ is a canonical link function. The null hypothesis $\beta_{X_j \times T} = 0$ 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 number of biomarkers $m$ to be considered is potentially large. Given the desired overall family-wise error rate $\alpha$, a Bonferroni correction \cite{Dunn 1961} requires an adjusted significance level for each individual test to be $\alpha/m$. Although the Bonferroni correction is typically used for its simplicity and flexibility, with regard to our interest in high-dimensional interaction testing it is worth exploring whether other procedures are able to provide improved efficiency. In the supplementary material, we demonstrate theoretically some alternative family-wise error rate controlling methods can only provide a small improvement across the settings we consider in this paper: when $m$ is large and only a small subset of biomarkers have true interactions with treatment. (Sidak correction \cite{Sidak 1967}, Holm-Bonferroni procedure \cite{Holm 1979} and Hochberg procedure \cite{Hochberg 1988})

2.2 Two-Stage Interaction Tests with Some Existing Screening Methods

Two-stage approaches use a screening test as a filtering stage (stage 1) to select a subset of biomarkers, and then in stage 2, only test interactions within the reduced set of biomarkers (a smaller $m$ in the Bonferroni Correction results in a less stringent significance level used in each single test), thus increasing power. 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 is proved follows.

A common stage 1 screening test used in two-stage interaction testing is a marginal association test \cite{Kooperberg and LeBlanc 2008}. Considering this type of screening test in the clinical trial setting, the marginal effect of a biomarker on the outcome can be measured in a regression model of the form

\[
G\{E(Y_i \mid X_{ij})\} = \delta_{0j} + \delta_{X_j} X_{ij}
\]

The screening procedure is conducted by testing the null hypothesis $\delta_{X_j} = 0$ 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 within the set of biomarkers selected at stage 1. Another way to utilize stage 1 information is to test all $m$ biomarkers in stage 2 using weighted significance levels, that add up to the targeted error rate $\alpha$, based on ordered biomarkers from stage 1. One possible weighting scheme \cite{Ionita-Laza et al. 2007} is: the $B$ most significant biomarkers, i.e. with lowest $p$-values in stage 1, are compared with an adjusted significance level $(\alpha/2)/B$, the next $2B$ biomarkers are compared with $(\alpha/4)/(2B)$, ..., the next $2^k B$ biomarkers are compared with $(\alpha/2^{k+1})/(2^k B)$, and so on.

The motivation of conducting marginal association tests to screen for candidate interaction tests is that we expect a biomarker that has an interaction with the treatment for the disease will also show some level of marginal association with the response. 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.

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 \cite{Murray et al. 2008} and Dai et al. (2012) proved that: with stage 1 and 2 test statistics being asymptotically independent and $m^*$ defined as the number of stage 1 selected biomarkers, using a Bonferroni adjusted significance level $\alpha = \alpha_1/m^*$ 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 $\alpha$.

In the context of gene-environment interaction studies, an alternative type of screening is testing the correlation between a gene and the environmental factor \cite{Murray et al. 2008}. However, such a screening procedure is not generally applicable to finding biomarker-treatment interactions in randomized clinical trials. We make this argument and also discuss the applicability of other related proposals more formally in the supplementary material.

2.3 New Stage 1 Sparse Regression Screening Procedure Accounting for Biomarker-Biomarker Correlations

One significant drawback of existing two-stage interaction testing procedures is that biomarkers are only tested one at a time. This ignores correlations between the biomarkers. 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 we considered
modern sparse regression methods to model correlated high-
dimensional data. These techniques have improved the de-
velopment of risk predictors from high-dimensional genomic
information (Wu et al., 2009; Lunn et al., 2006; Newcombe
et al., 2017).

We propose a new stage 1 multivariate screening test of
the following form to account for biomarker-biomarker corre-
lations

\[ G(E(Y_i | X_{i1}, \ldots, X_{im}) = \delta_0 + \delta_T T_i + \sum_{j=1}^{m} \delta_X_j X_{ij} \]  

(3)

This multivariate version of the marginal association screen-
ing 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.

To fit this multivariate model, we use ridge regression,
which applies regularization to avoid overfitting in high-
dimensional low-sample-size problems. Typically, the ob-
tjective of ridge regression is to minimize a loss function \( L_n \) along
with an \( L_2 \) regularization term:

\[ L_n(\delta) + \lambda_n ||\delta||^2_2 \]

where \( ||\delta||^2_2 = \delta_T^2 + \sum_{j=1}^{m} \delta_X_j^2 \). 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, we pro-
pose 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 described in Section 2.2.
Fitting a ridge model can be efficiently done through the
pathwise coordinate descent method (Friedman et al., 2010),
and the optimal value of the regularization parameter \( \lambda_n \)
can be chosen using cross-validation.

2.4 Proof of Independence between Stage 1 Sparse Regression
Screening and Stage 2 Standard Interaction Tests

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

For the \( i \)th subject, let \( Y_i \) denote the outcome variable,
\( X_i = (T_i, X_{i1}, \ldots, X_{im})^T \) be a vector of the binary treat-
ment-control indicator and \( m \) biomarkers. Consider the proposed
stage 1 marginal association screening test based on the
multivariate model of the form

\[ G(E(Y_i | X_i)) = X_i^T \delta \]

where \( \delta = (\delta_T, \delta_X_1, \ldots, \delta_X_m)^T \). The model underlying the
stage 2 standard one-biomarker-at-a-time interaction test is of
the form

\[ G(E(Y_i | V_{ij})) = V_{ij}^T \beta_j \quad (j = 1, \ldots, m) \]

where \( V_{ij} = (X_{ij}, T_i, X_{ij}T_i)^T \) and \( \beta_j = (\beta_X_j, \beta_T, \beta_X_jT_j)^T \).
The above forms ignore intercepts without loss of generality.
We first show the between-stage asymptotic independence
for the stage 1 multivariate regression marginal association
estimator without regularization.

**Theorem 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 \( \beta_{X_jT} = 0 \),

\[ \text{cov}(n^{1/2}(\hat{\delta}_X_j - \delta_{X_j}), n^{1/2}(\beta_{X_jT} - \beta_{X_jT})) \to 0 \]

in probability, where \( \hat{\delta}_X_j \) and \( \beta_{X_jT} \) are the maximum like-
lihood estimators for unknown parameters \( \delta_{X_j} \) and \( \beta_{X_jT} \)
respectively without regularization (i.e. \( \lambda_n = 0 \)).

The proof is provided in the appendix. Previous works
(Murcray et al., 2008; Dai et al., 2012) have demonstrated
the stage 1 univariate marginal association screening tests are
independent with the stage 2 one-biomarker-at-a-time inter-
action tests. Theorem 1 extends this to show independence
still holds when stage 1 tests are extended to a multivariate
regression. Our proof relies on: 1) the inclusion of the treat-
ment main effect in the multivariate regression of the form
(3); 2) an assumption of independence between the treatment
assignment and biomarker values, which is valid in the context
of a randomized clinical trial.

Next we establish the asymptotically linear form of the
ridge estimator.

**Lemma 1:** Under standard regularity conditions (Van der
Vaard, 2006, p. 51-52) and if \( \lambda_n = O(n^{1/2}) \), i.e.
\( \lim_{n \rightarrow \infty} \lambda_n/n^{1/2} = \lambda_0 \geq 0 \), then

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

in distribution, where \( \hat{\delta}^\lambda \) is the ridge estimator, \( \mathcal{N} \) is a normal
distribution, \( \sigma \) and \( \Sigma \) are a constant and an invertible constant
matrix.

Based on the asymptotic distributions derived in Lemma 1
and Theorem 1 we are able to prove the asymptotic in-
dependence between the stage 1 ridge marginal association
screening estimator and the stage 2 one-at-a-time interaction
estimator in the following corollary.

**Corollary 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 \( \beta_{X_jT} = 0 \),

\[ \text{cov}(n^{1/2}(\hat{\delta}_X_j - \delta_{X_j}), n^{1/2}(\beta_{X_jT} - \beta_{X_jT})) \to 0 \]

in probability, where \( \hat{\delta}_X_j \) is the maximum likelihood estimator
with the ridge penalty.

Proofs of Lemma 1 and Corollary 1 are given in the supple-
mentary material.

3. Results

3.1 Simulation Study

To evaluate performance of our proposed biomarker-
treatment interaction testing procedure described above, we
generated simulated data sets, each having \( m = 1,000 \)
biomarkers. Data were simulated under the model
\( Y_i = \beta_0 + \beta_T T_i + \sum_{j=1}^{m} (\beta_X_j X_{ij} + \beta_{X_jT} X_{ij} T_i) + \varepsilon_i \), where
the treatment main effect was set to \( \beta_T = 0.5 \) and the
intercept \( \beta_0 = 0 \). We partitioned the 1,000 biomarkers 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. $\beta_{X_{21}} = \beta_{X_{41}} = \beta_{X_{81}} = \beta_{X_{91}} = 1.5$. 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 $\varepsilon_i$ was generated from a normal distribution with standard deviation 5. We consider 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 is defined according to the idea of “cluster discoveries” in Brzyski et al. (2017) as $pr(\text{reject at least one } H_{0i}^k \text{ for any } X_j \in C_i)$, where at least one $H_{0i}^k$ is true for any $X_k \in C_i$, where $H_{0i}^k$ is the null hypothesis for $X_j$ and $H_{1i}^k$ is the alternative hypothesis for $X_i$.

Four different screening procedures are compared: 1) “Univariate screening (threshold)”: A selection of biomarkers to take forward to stage 2 is based on significance in a regression of response on the biomarkers one at a time, of the form $X_i \times T$. A significance level $\alpha_t = 0.05$ is used without adjustment for each stage 1 biomarker test. 2) “Univariate screening (rank)”: All biomarkers are taken forward to stage 2, and the stage 1 $p$-value ranking is used to conduct a stage 2 weighted hypothesis test described in Section 2.2 with $B = 5$ (a number recommended by Gauderman et al., 2013). 3) “Ridge screening (rank)”: Ridge regression is used to estimate marginal effects at stage 1. Then all biomarkers are ordered based on these stage 1 coefficients and the rank will be used by the stage 2 weighted hypothesis test with $B = 5$. The optimal $\lambda_0$ is chosen based on 5-fold cross-validation errors. The R package glmnet (Friedman et al., 2010) was used. 4) “No screening”: A standard single-stage interaction test of the form $\{1\}$, targeting an overall family-wise error rate $\alpha = 0.05$, is performed as a baseline comparator and also as the stage 2 test for all the three two-stage approaches described above.

In Fig. 1(a), with highly correlated biomarkers, the proposed ridge regression screening procedure demonstrated substantially higher power than the univariate screening procedures, showing a clear benefit of accounting for correlations between the biomarkers at stage 1. In Fig. 1(b), with independent biomarkers, where the multivariate regression is not required for unbiased effect estimation, the univariate screening and the ridge screening procedures using weighted hypothesis tests perform similarly. All three two-stage tests outperformed the single-step interaction test by providing better power at the same family-wise error rate level whether biomarkers are correlated or independent.

In Fig. 1(c), we simulated scenarios with one biomarker having an interaction, no correlations among the biomarkers, and changed only the main effect of the interacting biomarker $\beta_{X_1}$, i.e. main effects of the other four biomarkers were the same as the previous scenario. The sample size was fixed at 1,500. Fig. 1(c) reveals that there will be some special cases, in which the main and interaction effect parameters are in opposite directions such that they cancel out, where all two-stage approaches give lower power than a standard single-step interaction test.

In Fig. 1(d), we compared power of the different screening strategies while varying the proportion of explained variation by the true model. Specifically, we changed the standard deviance of the normal distribution from which $\varepsilon_i$ was drawn from. For this exploration, biomarkers were set to be correlated at 0.6. Fig. 1(d) shows that when the true model explains either a low or high proportion of the variance, all the methods tend to perform similarly to each other. In the wide spectrum between the two extremes, the comparison is rather consistent: the sparse regression screening strategy performs best, followed by the two univariate screening procedures, with the single-step interaction test always resulting in the lowest power.

In the supplementary material, we summarize family-wise error rates in different scenarios, which shows no inflation.
Two-Stage Sparse Regression Screening

5

for all the screening procedures. Additionally, we provide simulation results of scenarios in which we changed the interaction effect of the interacting biomarker and the correlation between biomarkers. Relative patterns of performance among the screening strategies were consistent with the results described above, demonstrating the robustness of our method and findings.

3.2 Data Applications

In addition to validating our methods through simulations, we exemplified our approaches in two real data applications.

Our first application was to data from the randomized controlled trial START [Fonagy et al., 2018], 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). We used a secondary outcome of this trial, the 18 months’ follow-up outcome from Inventory of Callous and Unemotional Traits, as the continuous outcome and applied our interaction testing procedures to detect covariates having interactions with the treatment. We 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. Correlation among these covariates is generally low (a correlation plot is provided in the supplementary material).

We performed all the four screening procedures described in the previous section with a significance level of \( \alpha = 0.05 \) and did not find any significant interactions. Table 1 lists the top covariates output by the univariate screening and ridge screening procedures. It is shown that the selected covariates from these two procedures are similar in this data set where covariates have low correlation.

In the second application with binary outcomes, we applied our approaches to detect biomarkers of steroid response in the treatment of alcoholic hepatitis (STOPAH trial) [Thurz et al., 2015]. The dataset 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). 28-day mortality was used as the binary response endpoint. We 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 for detecting interaction with treatment. There exists substantial correlation among these biomarkers (a correlation plot is provided in the supplementary material).

All the four methods described in the previous section with a significance level of \( \alpha = 0.05 \) did not find any significant biomarker-treatment interactions. Table 2 summarizes the top biomarkers from different marginal screening procedures: The results are quite different between the sparse regression screening and the univariate screening, likely owing to the high correlation among the biomarkers.

In addition, we examined the empirical correlation between stage 1 ridge screening and stage 2 interaction test statistics applied in the above two real data sets. Table 3 summarizes results from Pearson correlation tests, which shows that the empirical correlation between stages is close to zero and in all cases the 95% confidence interval contains zero as expected.

4. Discussion

We propose, for the first time with formal justification, the use of sparse ridge regression in a two-stage interaction testing framework for identifying biomarker signatures of treatment efficacy in randomized clinical trials. Interaction testing frameworks which are designed to scale to large numbers of covariates will become ever more important as -omics technologies continue to drop in price and become routinely measured in clinical trials.

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. Thus, lasso as an alternative sparse regression model which does not exhibit this effect averaging behavior may be expected to perform better. However, as lasso uses a \( L_1 \) penalty which is not a smooth function, it is challenging to prove it meets the between-stage independence requirement to preserve the overall family-wise error rate in two-stage approaches using current asymptotic theory. Other formulations of sparse regression methods, e.g. group lasso [Yuan and Lin, 2006], elastic net [Zou and Hastie, 2005], and Bayesian variable selection [O’Hara et al., 2009], that could be used at stage 1 or 2 are also worth exploring in future research.

We also showed that there exist special cases where our proposed two-stage screening strategy offers no benefit, e.g. the case 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. We suggest exploring the weighting scheme thus changing how much stage 1 information to be used in the following stage 2 tests as a future topic for investigation. Another technical caveat was shown by Sun et al. [2018] that, for logistic regression, the interaction estimator under treatment mis-specification 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 our framework as an elevated family-wise error rate and potential corrections, will be the topic of future work. We note, however, that since no strong evidence of interactions was found in our STOPAH case study, this bias is not a concern there - specifically the bias will generally manifest as false positives as opposed to false negatives.
Biometrics, Submitted in March 2020

Table 1
Top covariates from different stage 1 marginal screening procedures

|        | START trial | 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 | Parental reported Strengths & Difficulties Conduct Problems Score |

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

|        | START (pentoxifylline) | STOPAH (pentoxifylline) | STOPAH (prednisolone) |
|--------|------------------------|-------------------------|-----------------------|
| Estimate | 0.044                  | 0.104                   | 0.008                 |
| P value  | 0.711                  | 0.523                   | 0.960                 |
| 95% confidence interval | (-0.188, 0.271) | (-0.214, 0.402) | (-0.304, 0.319) |

ACKNOWLEDGEMENTS

This work was funded by the UK Medical Research Council (grant number MR/R502303/1 to J.W., grant number MC_UU_00002/9 to A.P. and P.J.N., grant number MC_UU_00002/6 to J.M.S.W.). P.J.N. acknowledges support from the NIHR Cambridge Biomedical Research Centre. The authors thank the START trial investigators for use of their data. The authors also acknowledge Stephen Atkinson and Mark Thursz for providing the STOPAH trial data.

REFERENCES

Brzyski, D., Peterson, C. B., Sobczyk, P., Candèes, E. J., Bogdan, M., and Sabatti, C. (2017). Controlling the rate of GWAS false discoveries. Genetics 205, 61–75.

Collins, F. S. and Varmus, H. (2015). A new initiative on precision medicine. New England Journal of Medicine 372, 793–795.

Dai, J. Y., Kooperberg, C., Leblanc, M., and Prentice, R. L. (2012). Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. Biometrika 99, 929–944.

Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American statistical association 56, 52–64.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70, 849–911.

Fonagy, P., Butler, S., Cottrell, D., Scott, S., Pilling, S., Eisler, L., Fuggle, P., Kraam, A., Byford, S., Wason, J., et al. (2018). Multisystemic therapy versus management as usual in the treatment of adolescent antisocial behaviour (START): a pragmatic, randomised controlled, superiority trial. The Lancet Psychiatry 5, 119–133.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33, 1.

Gauderman, W. J., Zhang, P., Morrison, J. L., and Lewinger, J. P. (2013). Finding novel genes by testing G×E interactions in a genome-wide association study. Genetic epidemiology 37, 603–613.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75, 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics pages 65–70.

Hsu, L., Jiao, S., Dai, J. Y., Hutter, C., Peters, U., and
Kooperberg, C. (2012). Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genetic epidemiology* **36**, 183–194.

Ionita-Laza, I., McQueen, M. B., Laird, N. M., and Lange, C. (2007). Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *The American Journal of Human Genetics* **81**, 607–614.

Khoury, M. J., Gwinn, M. L., Glasgow, R. E., and Kramer, B. S. (2012). A population approach to precision medicine. *American journal of preventive medicine* **42**, 639–645.

Kooperberg, C. and LeBlanc, M. (2008). 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**, 255–263.

Li, D. and Conti, D. V. (2008). Detecting gene-environment interactions using a combined case-only and case-control approach. *American journal of epidemiology* **169**, 497–504.

Locharernkul, C., Shotelersuk, V., and Hirankarn, N. (2011). Pharmacogenetic screening of carbamazepine-induced severe cutaneous allergic reactions. *Journal of Clinical Neuroscience* **18**, 1289–1294.

Lunn, D. J., Whittaker, J. C., and Best, N. (2006). A Bayesian toolkit for genetic association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* **30**, 231–247.

McAllister, K., Mechanic, L. E., Amos, C., Ashhard, H., Blair, I. A., Chatterjee, N., Conti, D., Gauderman, W. J., Hsu, L., Hutter, C. M., et al. (2017). Current challenges and new opportunities for gene-environment interaction studies of complex diseases. *American journal of epidemiology* **186**, 753–761.

Mukherjee, B. and Chatterjee, N. (2008). 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**, 685–694.

Murcray, C. E., Lewinger, J. P., and Gauderman, W. J. (2008). Gene-environment interaction in genome-wide association studies. *American journal of epidemiology* **169**, 219–226.

Newcombe, P. J., Raza Ali, H., Blows, F., Provenzano, E., Pharoah, P. D., Caldas, C., and Richardson, S. (2017). Weibull regression with Bayesian variable selection to identify prognostic tumour markers of breast cancer survival. *Statistical methods in medical research* **26**, 414–436.

O’Hara, R. B., Sillanpää, M. J., et al. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis* **4**, 85–117.

Parham, L., Briley, L., Li, L., Shen, J., Newcombe, P., King, K., Slater, A., Dilthey, A., Iqbal, Z., McVean, G., et al. (2016). 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**, 180.

Piegorsch, W. W., Weinberg, C. R., and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in medicine* **13**, 153–162.

Robert, J., Le Morvan, V., Giovannetti, E., Peters, G. J., et al. (2014). On the use of pharmacogenetics in cancer treatment and clinical trials. *European journal of cancer* **50**, 2532–2543.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* **62**, 626–633.

Sun, R., Carroll, R. J., Christiani, D. C., and Lin, X. (2018). Testing for gene–environment interaction under exposure misspecification. *Biometrics* **74**, 653–662.

Thursz, M. R., Richardson, P., Allison, M., Austin, A., Bowers, M., Day, C. P., Downs, N., Gleeson, D., MacGilchrist, A., Grant, A., et al. (2015). Prednisolone or pentoxifylline for alcoholic hepatitis. *New England Journal of Medicine* **372**, 1619–1628.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

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

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

Wason, J. M., Abraham, J. E., Baird, R. D., Gournaris, I., Vallier, A.-L., Brenton, J. D., Earl, H. M., and Mander, A. P. (2015). A Bayesian adaptive design for biomarker trials with linked treatments. *British journal of cancer* **113**, 699.

Wason, J. M. and Dudbridge, F. (2012). 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**, 760–773.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714–721.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49–67.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320.

**Supplementary Material**

Supplementary material available at the Biometrics website on Wiley Online Library includes proofs of Lemma 1 and Corollary 1 and theoretical discussions of the applicability/inafeasibility of alternative family-wise error rate controlling methods and case-only style interaction tests in the context of randomized clinical trials. We summarize family-wise error rates and provide additional simulation results to demonstrate the robustness of our method and findings.
The correlation plots of covariates from the two real data applications are also provided.

Received March 2020. Revised March 2020. Accepted March 2020.

APPENDIX

Proof of Theorem 1

Based on the unified approach to prove between-stage asymptotic independence by Dai et al. (2012), we need to evaluate the covariance matrix \( \mathbf{A}_1 - \mathbf{B} \mathbf{A}_2^{-1} \), where

\[
\begin{align*}
\mathbf{A}_1 &= E[(\mathbf{X}_i \mathbf{X}_j^T)(Y_i - E(Y_i | \mathbf{X}_i))^2] \\
\mathbf{B} &= E[(\mathbf{X}_i \mathbf{V}_ij^T)(Y_i - E(Y_i | \mathbf{X}_i))(Y_i - E(Y_i | \mathbf{V}_ij))] \\
\mathbf{A}_2 &= E[(\mathbf{V}_ij \mathbf{V}_ij^T)(Y_i - E(Y_i | \mathbf{V}_ij))^2]
\end{align*}
\]

We simplify the expression of \( \mathbf{B} \) as

\[
\mathbf{B} = E[(\mathbf{X}_i \mathbf{V}_ij^T)(Y_i^2 - Y_iE(Y_i | \mathbf{X}_i) - Y_iE(Y_i | \mathbf{V}_ij)) + E(Y_i | \mathbf{X}_i)E(Y_i | \mathbf{V}_ij)]
\]

\[
= E[(\mathbf{X}_i \mathbf{V}_ij^T)E(Y_i^2 - Y_iE(Y_i | \mathbf{X}_i) - Y_iE(Y_i | \mathbf{V}_ij)) + E(Y_i | \mathbf{X}_i)E(Y_i | \mathbf{V}_ij)]
\]

\[
= E(\mathbf{X}_i \mathbf{V}_ij^T)\text{var}(Y_i | \mathbf{X}_i)
\]

which uses the law of iterated expectations, the fact that \( \mathbf{X}_i \) includes \( \mathbf{V}_ij \) under the null hypothesis \( \beta_{X_i} = 0 \), and assumes the homogeneity of variance, i.e. \( \text{var}(Y_i | \mathbf{X}_i) \) is a constant.

Similarly, we have \( \mathbf{A}_1 = E(\mathbf{X}_i \mathbf{X}_i^T)\text{var}(Y_i | \mathbf{X}_i) \) and \( \mathbf{A}_2 = E(\mathbf{V}_ij \mathbf{V}_ij^T)\text{var}(Y_i | \mathbf{V}_ij) \). Thus,

\( \mathbf{A}_1^{-1} \mathbf{B} \mathbf{A}_2^{-1} \propto E(\mathbf{X}_i \mathbf{X}_i^T)^{-1}E(\mathbf{X}_i \mathbf{V}_ij^T)E(\mathbf{V}_ij \mathbf{V}_ij^T)^{-1} \)

We consider the second and the third terms

\[
\begin{align*}
E(\mathbf{X}_i \mathbf{V}_ij^T)_{(m+1) \times 3} &= \begin{pmatrix}
E(T_iX_{ij}) & E(T_i^2X_{ij}) & E(T_i^2X_{ij}) \\
E(X_{ik}X_{ij}) & E(T_iX_{ik}) & E(T_iX_{ik}X_{ij}) \\
\vdots & \vdots & \vdots \\
E(X_{im}X_{ij}) & E(T_iX_{im}) & E(T_iX_{im}X_{ij})
\end{pmatrix} \\
E(\mathbf{V}_ij \mathbf{V}_ij^T)^{-1} &= \begin{pmatrix}
E(X_{ij}^2) & E(T_iX_{ij}) & E(T_i^2X_{ij}) \\
E(T_iX_{ij}) & E(T_i^2X_{ij}) & E(T_i^2X_{ij}) \\
E(T_iX_{ij}) & E(T_i^2X_{ij}) & E(T_i^2X_{ij})
\end{pmatrix}^{-1} \\
&= \frac{1}{\det(E(\mathbf{V}_ij \mathbf{V}_ij^T))}
\end{align*}
\]

Thus, for the \((m + 1) \times 3\) matrix \( E(\mathbf{X}_i \mathbf{V}_ij^T)E(\mathbf{V}_ij \mathbf{V}_ij^T)^{-1} \), the \((k + 1)\)th element \((k = 1, \ldots, m)\) of the last column can be computed by

\[
\frac{1}{\det(E(\mathbf{V}_ij \mathbf{V}_ij^T))} \begin{pmatrix}
E(X_{ik}X_{ij}) & E(T_iX_{ik}) & E(T_iX_{ik}X_{ij}) \\
E(T_iX_{ij})E(T_i^2X_{ij}) - E(T_i^2)E(T_iX_{ij}) \\
\cdot & \cdot & \cdot \\
E(X_{ij}^2)E(T_i^2) - E(T_iX_{ij})^2
\end{pmatrix}
\]