Zhang Min (Orcid ID: 0000-0003-3331-3583)

Interaction Analysis under Misspecification of Main Effects: Some Common Mistakes and Simple Solutions

Authors:
Min Zhang[1], Youfei Yu[1], Shikun Wang[2], Maxwell Salvatore[1], Lars Fritsche[1] , Zihuai He[3], Bhramar Mukherjee[1]

Affiliations:
[1] Department of Biostatistics, University of Michigan School of Public Health
[2] MD Anderson Cancer Center Department of Biostatistics
[3] Department of Neurology and Neurological Sciences, Stanford University

*Corresponding Author
Bhramar Mukherjee, PhD
1415 Washington Heights, Ann Arbor, MI
Phone: (734) 764-6544
Email: bhramar@umich.edu

Abstract

The statistical practice of modeling interaction with two linear main effects and a product term is ubiquitous in the statistical and epidemiological literature. Most data modelers are aware that the misspecification of main effects can potentially cause severe type I error inflation in tests for interactions, leading to spurious detection of interactions. However, modeling practice has not changed. In this paper, we focus on the specific situation where the main effects in the model are misspecified as linear terms and characterize its impact on common tests for statistical interaction. We then propose some simple alternatives that fix the issue of potential type I error inflation in testing interaction due to main effect misspecification. We show that when using the sandwich variance estimator for a linear regression model with a quantitative outcome and two independent factors, both the Wald and score tests asymptotically maintain the correct type I error rate. However, if the independence assumption does not hold or the outcome is binary, using the sandwich estimator does not fix the problem. We further demonstrate that flexibly modeling the main effect under a generalized additive model can largely reduce or often remove bias in the estimates and

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/sim.8505

This article is protected by copyright. All rights reserved.

Author Manuscript

Author Manuscript

maintain the correct type I error rate for both quantitative and binary outcomes regardless of the independence assumption. We show, under the independence assumption and for a continuous outcome, overfitting and flexibly modeling the main effects does not lead to power loss asymptotically relative to a correctly specified main effect model. Our simulation study further demonstrates the empirical fact that using flexible models for the main effects does not result in a significant loss of power for testing interaction in general. Our results provide an improved understanding of the strengths and limitations for tests of interaction in the presence of main effect misspecification. Using data from a large biobank study "*The Michigan Genomics Initiative*", we present two examples of interaction analysis in support of our results.

Keywords: Generalized Additive Model (GAM), Gene-Environment Interaction, Independence, Joint Tests, Power, Robust Tests, Sandwich Variance Estimator, Type I error.

Introduction

The scientific notion of interaction between two factors tries to capture the phenomenon that the effect of one factor is different in the presence or absence of another factor.[1] This could be of the nature that one factor is activated/silenced only in the presence of another factor, thus exhibiting a complete synergistic or antagonistic effect. It could also be more subtle in terms of modification of the strength of association of one factor with the outcome when the other factor is set at two different levels. This definition does not assume any particular structure of the joint response surface determined by the two factors, except that under the hypotheses of no-interaction, the implied marginal response surfaces of one factor are simple constant shifts when the other factor is fixed at two different levels. Interaction is often statistically assessed by fitting a regression model for a quantitative or binary outcome by including two linear main effects and products between the two factors. However, missing a quadratic term (say) in one variable that truly exists can lead to the detection of spurious interactions in a linear model as the cross-product term then tries to mimic/approximately capture the second order features of the model. There exists some

This article is protected by copyright. All rights reserved.

literature on this topic in statistics, genetics, and epidemiology.[2-8] For longitudinally measured quantitative outcomes main effect misspecification is discussed in He et al.[9]

In this paper, we consider a specific scenario related to the effect of misspecification of main effect structure on tests for statistical interaction: when the true underlying main effect is nonlinear but a linear model is specified for the main effects. When such main effect misspecification is present, then, in general, the standard statistical tests (e.g., the Wald or score test based on model-based standard error) will lead to an invalid test of interaction and potentially severe type I error rate inflation. Under certain conditions, the type I error inflation may be fixed by using robust inference (e.g., using sandwich variance estimator) and this phenomenon has been empirically observed by, for example, Voorman et al.[10] and Cornelis et al,[6] and formally studied by Tchetgen Tchetgen and Kraft,[5] He et al,[9] and Sun et al.[8] This problem has also been discussed recently in analyzing treatment and biomarker interaction as it is natural to assume independence of treatment with other covariates in a randomized clinical trials.[11-12]

We show that for quantitative outcomes when a linear regression model is applied, and the two factors are independent, both the usual Wald and score tests, when modified by the sandwich variance estimator asymptotically maintain correct type-1 error. However, if the independence assumption does not hold or the outcome is binary and analyzed by a logistic regression model, using the sandwich estimator does not fix the problem. We further demonstrate that flexibly modeling the main effect under a generalized additive model using a flexible nonparametric term can reduce bias in the estimates and maintain correct type-1 error for both quantitative and binary outcomes regardless of the independence between the two factors. We show, under the independence assumption and for a continuous outcome, overfitting and flexibly modeling the main effects does not lead to power loss asymptotically relative to a correctly specified main effect model. Our simulation studies indicate by flexibly modeling the main effect we do not lose power significantly for testing interaction in general. Using data from the Michigan Genomics Initiative,

Author Manuscript

This article is protected by copyright. All rights reserved.

a large ongoing biobank study at the University of Michigan, we illustrate our theoretical and simulation results as they pertain to two examples on interaction analysis.

This paper contributes to the current literature by considering both quantitative and binary outcomes, proposing and studying two general ways of handling main effect misspecification (i.e., robust inference and flexible modeling of main effects), and studying the advantages and disadvantages of each method in terms of both type I error control and power under different assumptions regarding independence. Our results provide an improved understanding of the strengths and limitations of each method, in both finite samples and large samples, for interaction tests in the presence of main effect misspecification.

Methods

Tests for statistical interaction

We are interested in evaluating the interaction effect between two variables X_1 and X_2 on the outcome Y, which can be quantitative or binary, based on a study with n individuals. The observed data are denoted by (X_{1i}, X_{2i}, Y_i) for $i = 1, \dots, n$. Denoting $\mu_i = E(Y_i|X_{1i}, X_{2i})$, we suppose the test of interaction is based on the following regression model,

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} \qquad (1)$$

where $\beta = [\beta_0, \beta_1, \beta_2, \beta_3]^T$ are unknown regression parameters, and $g(\mu)$ is a link function. Specifically, we assume a linear regression model is used for quantitative outcomes and a logistic regression model is used for binary outcomes, i.e., $g(\mu_i) = \mu_i$ for quantitative outcomes and $g(\mu_i) = \text{logit}(\mu_i) \equiv \log(\frac{\mu_i}{1-\mu_i})$ for binary outcomes. The parameter β_3 measures the magnitude of a linear statistical interaction between X_1 and X_2. Based on the regression model, to test the interaction between X_1 and X_2, one can test the hypothesis H_0: $\beta_3 = 0$ vs. H_1: $\beta_3 \neq 0$. We first describe two commonly used tests, namely, the Wald test and score test, and inferential procedures using the model-based standard error and the empirical sandwich standard error.

This article is protected by copyright. All rights reserved.

Author Manuscript

Author Manuscript

Wald test

The Wald test is one of the most commonly used methods for testing unknown parameters in a parametric regression model. It is constructed using the maximum likelihood estimate of the parameter of interest and its standard error. Considering model (1), let $\hat{\beta}$ denote the usual maximum likelihood estimate of β. For both linear and logistic regression models, it is the solution to the estimating equation

$$\sum_i X_i\{Y_i - g^{-1}(X_i^T\beta)\} = 0,$$

where $X_i = [1, X_{1i}, X_{2i}, X_{1i}X_{2i}]^T$. Two methods can be used to estimate the variance and covariance matrix of $\hat{\beta}$. In model-based inference, the variance estimate is obtained by assuming the specified linear/logistic regression model is correct. Alternatively, one can obtain the empirical estimate of variance without assuming the corresponding mean regression model is correctly specified using the so-called sandwich variance estimate. See Appendix for details. We denote the predictions and residuals as $\hat{\mu}_i = g^{-1}(X_i^T\hat{\beta})$ and $\hat{\epsilon}_i = Y_i - \hat{\mu}_i$ respectively. For a linear regression model with a quantitative outcome, the model-based and sandwich variance estimates of $\hat{\beta}$ are

$$\hat{V}_{model}(\hat{\beta}) = \frac{1}{n-p}\left(\sum_i \hat{\epsilon}_i^2\right)\left(\sum_i X_i X_i^T\right)^{-1},$$

$$\hat{V}_{sandwich}(\hat{\beta}) = \frac{n}{(n-p)}\left(\sum_i X_i X_i^T\right)^{-1}\left(\sum_i X_i X_i^T\,\hat{\epsilon}_i^2\right)\left(\sum_i X_i X_i^T\right)^{-1},$$

respectively, where p is the dimension of X_i. For a logistic regression model with a binary outcome, the model-based and sandwich variance estimates of $\hat{\beta}$ are

$$\hat{V}_{model}(\hat{\beta}) = \frac{n}{n-p}\left\{\sum_i X_i X_i^T \hat{\mu}_i(1-\hat{\mu}_i)\right\}^{-1},$$

This article is protected by copyright. All rights reserved.

$$\hat{V}_{sandwich}(\hat{\beta}) = \frac{n}{(n-p)}\left\{\sum_i X_i X_i^T \hat{\mu}_i(1-\hat{\mu}_i)\right\}^{-1}\left(\sum_i X_i X_i^T \hat{\epsilon}_i^2\right)\left\{\sum_i X_i X_i^T \hat{\mu}_i(1-\hat{\mu}_i)\right\}^{-1},$$

respectively. Under H_0, if the model for main effects (i.e., effects of X_1 and X_2) is correct, then asymptotically the Wald test statistic $\hat{\beta}_3^2/\hat{V}_{model}(\hat{\beta}_3)$ with model-based variance estimate and its sandwich version $\hat{\beta}_3^2/\hat{V}_{sandwich}(\hat{\beta}_3)$ follow a Chi-square distribution with 1 degree of freedom. For a level α test, we reject $H_0: \beta_3 = 0$ when the test statistic is greater than $\chi^2_{1,\alpha}$, where $\chi^2_{1,\alpha}$ satisfies $P(\chi_1^2 > \chi^2_{1,\alpha}) = \alpha$.

Score test

Unlike the Wald test which is based on fitting a full model including both main effects of X_1 and X_2 and their interaction term, the score test is based on the score statistics of a model under the null hypothesis. Specifically, under the null hypothesis, model (1) reduces to the model with only main effects:

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}, \qquad (2)$$

where $\beta = [\beta_0, \beta_1, \beta_2]^T$ are unknown parameters in the null model. Let $\tilde{\beta}$ be the maximum likelihood estimate of β under this null model and $\tilde{\beta}$ is the solution to the estimating equation

$$\sum_i X_{o,i}\{Y_i - g^{-1}(X_{o,i}^T\beta)\} = 0,$$

where $X_{o,i} = [1, X_{1i}, X_{2i}]^T$. We denote the predictions and residuals from model (2) as $\tilde{\mu}_i = g^{-1}(X_{o,i}^T\tilde{\beta})$ and $\tilde{\epsilon}_i = Y_i - g^{-1}(X_{o,i}^T\tilde{\beta})$ respectively. The score statistic with respect to β_3 is

$$S = \frac{1}{n}\sum_i X_{1i}X_{2i}(Y_i - \tilde{\mu}_i).$$

For a linear regression model for a quantitative outcome, the model-based and sandwich variance estimate of S are

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

$$\hat{V}_{model}(S) = \frac{1}{n^2(n-p)}\left(\sum_i \tilde{\epsilon}_i{}^2\right)\tilde{A}\left(\sum_i X_i X_i^T\right)\tilde{A}^T,$$

$$\hat{V}_{sandwich}(S) = \frac{1}{n(n-p)}\tilde{A}\left(\sum_i X_i X_i^T \tilde{\epsilon}_i{}^2\right)\tilde{A}^T,$$

respectively, where $\tilde{A} = \left[-\left(\sum_{i=1}^n X_{1i}X_{2i}X_{o,i}^T\right)\left(\sum_{i=1}^n X_{o,i}\, X_{o,i}^T\right)^{-1}, 1\right]$ and p is the dimension of $X_{o,i}$. For a logistic regression model for binary outcomes, the model based and sandwich variance estimate of S are respectively,

$$\hat{V}_{model}(S) = \frac{1}{n(n-p)}\tilde{B}\left(\sum_i X_i X_i^T \tilde{\mu}_i(1-\tilde{\mu}_i)\right)\tilde{B}^T,$$

$$\hat{V}_{sandwich}(S) = \frac{1}{n(n-p)}\tilde{B}\left(\sum_i X_i X_i^T \tilde{\epsilon}_i{}^2\right)\tilde{B}^T,$$

where $\tilde{B} = \left[-\{\sum_{i=1}^n X_{1i}X_{2i}\tilde{\mu}_i(1-\tilde{\mu}_i)X_{o,i}^T\}\{\sum_{i=1}^n \tilde{\mu}_i(1-\tilde{\mu}_i)X_{o,i}\, X_{o,i}^T\}^{-1}, 1\right]$. Under H_0, if the model for main effects is correct, both model based score test statistic $S^2/\hat{V}_{model}(S)$ and its sandwich version $S^2/\hat{V}_{sandwich}(S)$ follows a Chi-square distribution with 1 degree of freedom. We reject $H_0: \beta_3 = 0$ when the test statistics are sufficiently large.

Misspecification of Main effects

So far, we have discussed four tests (Wald and score tests with a model-based variance estimate, Wald and score tests with a sandwich variance estimate). When the main effects for X_1 and X_2 are correctly specified, all four tests lead to correct type I error rates. However, the underlying model is often unknown, and X_1, X_2 or both likely have a non-linear effect. Misspecifying the main effects may lead to spurious findings.

To remedy the type I error inflation due to misspecification of main effects one solution is to

This article is protected by copyright. All rights reserved.

Author Manuscript

replace the usual model-based statistical inference by the robust inference based on sandwich variance estimation. An alternative solution is to use a Generalized Additive Model (GAM)[13] to model the main effect of X_1 more flexibly. GAM extends a generalized linear model to include smooth functions of explanatory variables with the smoothness determined by a parameter that either directly controls the smoothness of the curve or the estimated predictive accuracy. We consider two types of GAMs:

$$\text{GAM1: } g\big(E(Y_i|X_{1i}, X_{2i})\big) = \beta_0 + \beta_1 s_1(X_{1i}) + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i}$$

$$\text{GAM2: } g\big(E(Y_i|X_{1i}, X_{2i})\big) = \beta_0 + \beta_1 s_1(X_{1i}) + \beta_2 s_2(X_{2i}) + \beta_3 X_{1i} X_{2i}$$

where $s_j(x)$, $j = 1, 2$, are smooth functions using thin plate splines.[14] Although GAM is a common method to model non-linear effects, it has not been recognized and well discussed in interaction analysis to address main effect misspecification. The strategy here is to try to model the main effect of X_1 and X_2 correctly using nonparametric models where only a mild smoothness assumption is made to achieve type I error control. Modeling the main effect correctly and flexibly (or approximately so) can lead to an improvement in power relative to a robust sandwich inference based on an incorrectly specified main effect model, as demonstrated in our simulation studies. Moreover, a flexible main effect model, even unnecessary, does not result in power loss under the independence assumption for continuous outcomes, relative to a correctly specified main effect model as we discuss later. A similar phenomenon is discussed and proved in He et al.[9] in the setting of testing for gene-environment interaction for repeated measurements. However, note that we are still considering the true interaction term to be linear.

In this paper, we focus on testing interaction alone, i.e., testing for $\beta_3 = 0$. In Tchetgen Tchetgen and Kraft,[4] they considered the joint test of one factor (e.g., genetic factor) and its interaction with another factor (e.g., environmental factor), i.e., testing for $\beta_2 = 0$ and $\beta_3 = 0$ jointly. They showed that when assuming gene-environment independence for a binary outcome modeled using logistic regression, a joint test using a Wald or score test combined with the sandwich variance estimator leads to the correct type I error rate even when one of the main effects is misspecified.

This article is protected by copyright. All rights reserved.

Author Manuscript

As our results will show, for logistic regression, robustness against main effect misspecification using a sandwich variance estimator does not hold in general for testing for interaction alone. Such robustness will only hold under the additional assumption that the true β_2 is zero, as commented by Tchetgen Tchetgen and Kraft.[5]

Simulation Design

We conducted simulation studies under misspecification of main effects to evaluate the performance of the methods mentioned above based on 500 replicates: 1. Wald test with model-based variance estimate; 2. Wald and score tests with sandwich variance estimate; 3. Wald test with model-based variance estimate but using GAM to model the possibly non-linear main effect. Additionally, when the outcome is quantitative, we also compare these methods with the rule ensemble method of Friedman and Popescu[15] for testing interaction, where the form of the interaction is completely arbitrary. We refer to this method by *RuleFit* (Predictive Learning via Rule Ensemble) and we implemented it using the R-package pre.[16] The details on implementation of the *RuleFit* are given in the Supplementary material. We simulated four continuous and binary outcome models with a linear, quadratic, log or exponential main effect for X_1 as follows,

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$
$$g(\mu) = \beta_0 + \beta_1 (X_1 + 2X_1^2) + \beta_2 X_2 + \beta_3 X_1 X_2$$
$$g(\mu) = \beta_0 + \beta_1\, log(X_1) + \beta_2 X_2 + \beta_3 X_1 X_2$$
$$g(\mu) = \beta_0 + \beta_1\, exp(X_1) + \beta_2 X_2 + \beta_3 X_1 X_2$$

where $\mu = E(Y|X_1, X_2)$; $g(\mu) = \mu$ for continuous outcomes; $g(\mu) = \text{logit}(\mu)$ for binary outcomes. The two factors X_1 and X_2 are both continuous variables generated from normal/log-normal distributions, and we consider settings where they are independent or dependent, as detailed in **Supplementary Tables S1** and **S2**. For continuous outcomes, we consider sample size $n = 500$, $(\beta_0, \beta_1, \beta_2) = (1,2,3)$ and, for binary outcomes, we consider $n = 2{,}000$, $(\beta_1, \beta_2) = (1,2)$ and β_0 is chosen such that the marginal prevalence of Y is 0.2. We vary β_3 to evaluate type I error rate $(\beta_3 = 0)$ and power $(\beta_3 > 0)$. We present the results in **Figures 1** and **2**. Additionally,

This article is protected by copyright. All rights reserved.

we evaluated the type I error rate ($\beta_3 = 0$) under greater sample size up to 10,000 and present the results in **Figures 3** and **4**. The exact numerical values can be found in **Supplementary Tables S3-S6**.

Results

Analytical results: main effect misspecification and independence assumption

Result 1 (Wald test): For quantitative outcomes, under the null hypothesis (i.e., there is no interaction between X_1 and X_2 in the true model) and under the assumption of independence of X_1 and X_2, if a linear regression model is used, then regardless of whether the main effects for $X_1$and X_2 are correctly specified or not, $\hat{\beta}_3$ converges in probability to 0, and $\sqrt{n}\,\hat{\beta}_3$ converges in distribution to a normal distribution. The asymptotic variance can be consistently estimated by the empirical sandwich variance estimator.

Result 2 (score test): For quantitative outcomes, under the null hypothesis (i.e., there is no interaction between X_1 and X_2 in the true model) and under the assumption of independence of X_1 and X_2, if a linear regression model is used and both X_1 and X_2 are centered, then regardless of whether the main effects for $X_1$and X_2 are correctly specified or not, the score for testing the interaction of X_1 and X_2 , i.e., $S = \frac{1}{n}\sum_i S_i(\tilde{\beta}) = \frac{1}{n}\sum_i\{X_{i1}X_{i2}(Y_i - \tilde{\beta}X_{i1} - \tilde{\beta}_2X_{i2})\}$, is unbiased for zero and $\frac{1}{\sqrt{n}}\sum_i S_i(\tilde{\beta})$ converges in distribution to a normal distribution. The asymptotic variance can be consistently estimated by the empirical sandwich variance estimator.

The detailed proofs for results 1 and 2 are in **Appendix (A) and (B).** We refer to the assumption that X_1 and X_2 are independent as the independence assumption. The results show that in the interaction analysis of a quantitative trait based on a linear regression model, under the independence assumption, the type I error inflation caused by main effect misspecification can be corrected by replacing the model-based variance estimator with the empirical sandwich variance

Author Manuscript

This article is protected by copyright. All rights reserved.

estimator.

However, for binary traits modeled using logistic regression with $g(\mu) = \text{logit}(\mu)$, this robustness property against main effect misspecification does not hold for testing $\beta_3 = 0$ unless, additionally, one of X_1 or X_2 has no main effect, say $\beta_2 = 0$. An explanation of why robustness does not hold for logistic regression models is given in **Appendix (B)**. The lack of robustness for logistic regression follows from a general result studied by Tchetgen Tchetgen.[4] As a result, the Wald test and score test cannot be corrected by only changing the variance estimation. In general, for binary outcomes modeled using logistic regression, the simple correction using the empirical sandwich variance estimation only works for jointly testing $\beta_2 = \beta_3 = 0$. We have provided codes for implementing the tests mentioned above at https://github.com/youfeiyu/GbyEtests.

In summary, with respect to type I error control, inference based on the empirical sandwich variance estimation offers a simple solution to main effect misspecification in the setting where the outcome is quantitative, a linear regression model is used, and the independence assumption holds. In other settings (e.g., binary outcomes, independence assumption is violated), a correct specification of the main effect is often required to guarantee correct type I error at the nominal level. In addition to type I error control, another consideration of importance in testing for interaction is power. Correct specification of the main effect offers an advantage in terms of power by reducing the residual variance even when robustness against main effect misspecification in terms of type I error control holds. In general, overfitting the main effects but not the interaction term using models will not reduce power asymptotically relative to a correct specification of the main effect. In particular, flexibly modeling the main effects using GAM will not lead to power loss asymptotically under the independence assumption. This result is shown in **Appendix (C)**.

Simulation results

Because model-based score tests behave similarly to the model-based Wald test, we omit results

Author Manuscript

This article is protected by copyright. All rights reserved.

on modeled-based score tests in our **Figures** and **Tables**. **Figure 1** presents empirical power curves of various methods for testing $\beta_3 = 0$ when the outcome is continuous and the sample size is 500. Note that the point in each power curve corresponding to $\beta_3 = 0$ is the empirical type I error rate. We observe that when there is no misspecification of main effects, model-based and sandwich Wald and score tests all maintain the type I error rate at nominal levels regardless of whether X_1 and X_2 are independent (**Figure 1, panels A** and **E**) and have similar power. When the true main effect of X_1 is nonlinear but is mistakenly modeled using a linear form, model-based Wald test leads to inflated type I error rates, regardless of whether X_1 and X_2 are independent (**Figure 1, panels B-D, F-H**). When X_1 and X_2 are independent, **Figure 1, panels B** and **D** show that both the sandwich Wald test and the sandwich score test can fix the type I error inflation and maintain type I error rate at the nominal level of 0.05 when the main effect of X_1 is quadratic or exponential, while, for example, the corresponding model-based Wald test leads to a type I error rate of 0.37 when the main effect of X_1 is quadratic. When the main effect of X_1 is a logarithmic function (**Figure 1C**), sandwich Wald and score tests still exhibit type I error inflation (0.11 and 0.07, respectively) even when X_1 and X_2 are independent. However, this inflation decreases as sample size increases (**Figure 3**). When sample size >2000, sandwich score test achieves type I error rate at the nominal level of 0.05, while sandwich Wald test requires even larger sample size ($> 10^5$) to achieve the type I error rate at the nominal level of 0.05 (**Supplementary Table S3**). When X_1 and X_2 are dependent and the true main effect of X_1 is nonlinear, all model-based and sandwich tests assuming a linear main effect exhibit severe type I error inflation when the true main effect of X_1 is nonlinear (**Figure 1, panels F- H**). For example, the level 0.05 sandwich score test leads to a type I error rate ranging from 0.12-0.83 in **Figure 3 F-H**. Wald tests using GAM to flexibly model the main effect (GAM1 and GAM2) lead to a well-controlled type I error rate in all scenarios considered here regardless of whether X_1 and X_2 are independent.

We have shown that when X_1 and X_2 are independent, then overfitting the main effect in a linear model will not lead to power loss asymptotically. Based on our empirical results, Wald tests using

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

GAM for main effects have good performance in terms of power even when the independence assumption is not met. They are almost as powerful as tests based on a correctly specified main effect model (**Figure 1 A** and **E**). Additionally, they are significantly more powerful than sandwich Wald and score tests based on a misspecified main effect model when the corresponding type I error rate is also well controlled (**Figure 1, panels B-D**), i.e., when X_1 and X_2 are independent. For example, as shown in **Figure 1B**, when $\beta_3 = 0.2$, both GAM1 and GAM2 have power 0.99 whereas sandwich Wald and Score tests have power 0.11 and 0.08, respectively. This result is observed because the nonparametric modeling can correctly approximate the main effect therefore reducing the residual variance and improving power. Because the true effect of X_2 is linear in this setting, modeling the main effect of X_2 using a nonparametric function as in GAM2 is not necessary. However, we see that power curves for GAM1 and GAM2 are almost indistinguishable, indicating there is little or no loss of efficiency empirically for testing interaction by using a flexible model, even when unnecessary, to model the main effect in linear regression. Finally, we note the very flexible *RuleFit* method leads to severe inflated type I error and undesirable power in almost all scenarios considered here. The type I error inflation is likely due to the method not being able to evaluate the null distribution of the test statistics well since no analytic null distribution is available. One explanation for the power loss is the unnecessary flexible modeling of the interaction term. Based on our experience, overfitting the interaction often leads to severe power loss as it changes the null distribution and degrees of freedom used for evaluating significance, which is in contrary to overfitting the main effects.

Figure 2 presents empirical power curves of tests for interaction when the outcome is binary and the sample size is 2,000. As before, all model-based and sandwich Wald and score tests can control the type I error rate at the nominal level and have similar power when main effects are correctly modeled (**Figure 2, panels A** and **E**). However, we observe that, when the main effect is misspecified, the sandwich Wald and score tests are not able to maintain the type I error rate at the nominal level even when X_1 and X_2 are independent and the type I error inflation persists even as

This article is protected by copyright. All rights reserved.

sample size increases (**Figure 4**). For example, the sandwich Wald and score tests have a type I error rate of 0.83 when the main effect of X_1 is quadratic. The tests using GAM for main effects considerably improve type I error control and the type I error rates achieve the nominal level except for the scenario where the main effect of X_1 is exponential (e.g., **Figure 2D**, 0.19 and 0.18 for GAM1 and GAM2, respectively). We comment that this is a rather extreme case, and in this case, the type I error rates of other methods are almost 1.00. The type I error inflation decreases as sample size increases, which allows GAM to approximate the exponential function better (**Figure 4**). Compared with a parametric model for a binary outcome with correctly modeled main effects, we note that flexibly modeling the main effects using GAM when unnecessary leads to some loss of efficiency as shown in **Figure 2 panels A** and **E** and that GAM2 leads to slightly more loss of power compared to GAM1.

In summary, these results show that for continuous outcomes in a linear model, when X_1 and X_2 are independent, replacing the model-based variance estimate with the sandwich estimate in Wald and score tests can reduce or remove type I error inflation. However, this does not hold for binary outcomes in a logistic regression model. Using GAM to flexibly model main effects appears to be a simple and appropriate solution for main effect misspecification in terms of both type I error rate and power.

Data Application: Interaction analysis in the Michigan Genomics Initiative

We illustrate our observations regarding the type I error inflation due to main effect misspecification and power enhancement by flexibly modeling the main effect respectively using two data examples. The first example is a genome-wide gene-environment interaction study that investigated the effect of interaction between body mass index (BMI) and single nucleotide polymorphisms (SNP) on chronic ulcer of skin across the genome. A non-linear relationship between the log-odds of having skin ulcer and BMI is noted here. The second example examined a series of models for BMI as the outcome of interest, modeled as a function of age and sex, and

Author Manuscript

This article is protected by copyright. All rights reserved.

interaction between age and sex. In the second example, a quadratic relationship between age and BMI is observed. The data corresponding to both examples came from the Michigan Genomics Initiative (MGI), an electronic health record (EHR)-linked biobank at the University of Michigan that started in 2012. More detailed descriptions regarding the recruiting criteria, description of the study cohort, and the enrollment procedure in MGI can be found in Fritsche et al.[17]

Example 1: Type I error inflation due to misspecified main effects

This example included 38,162 unrelated individuals of recent European ancestry with genotyped data, 2,186 (5.5%) of whom had a "chronic ulcer of the skin" in their records. The analytic dataset is 47.5% male and has a mean age of 54.5 (range = [18.0, 102.3]) and a mean BMI of 29.8 (range = [12.3, 91.1]). Age and BMI data came from the subjects' EHR and age at the time of BMI measurement was used. We first inspected the functional form of the relationship between the chronic ulcer of skin (*D*, say) and BMI by fitting the following generalized additive model

$$\text{logit}\{P(D = 1|\text{BMI}, X)\} = \alpha_0 + s(\text{BMI}) + \alpha_X X,$$

where D denotes the disease status (1 being a case) and X contains age, sex, genotyping array, and the first four principal components obtained from the principal component analysis of the genotyped markers. Both BMI and age were centered before analysis. The results from the model described above revealed a nonlinear relationship between chronic ulcer of skin and (centered) BMI (**Supplementary Figure S1A**).

We then investigated the SNP-BMI interactions as risk factors for chronic ulcer of skin. We tested the interaction effects between BMI and 272,672 genotyped variants with minor allele frequency $\geq$ 1% using PLINK 1.9. For each SNP considered in this analysis we fitted the model

$$\text{logit}\{P(D = 1|\text{SNP}, \text{BMI}, X)\} = \beta_0 + \beta_{SNP}\text{SNP} + s(\text{BMI}) + \beta_X X + \beta_{SNP\times BMI}\text{SNP} \times \text{BMI}$$

where the notations are defined in the same way as in Model (3) and the nonlinear relationship as observed in Figure S1 was modeled using GAM through the smooth function $s(\text{BMI})$. We also fitted a model with a linear main effect term of BMI to explore the impact of incorrectly specifying

This article is protected by copyright. All rights reserved.

Author Manuscript

the main effect on testing for the SNP × BMI interaction and then tested the interaction using both model-based and sandwich Wald tests.

Models were fitted using the full cohort (2,186 cases and 35,976 controls) as well as in a more balanced cohort with a 1:3 case-control ratio (2,186 cases and 6,558 randomly selected controls). For both cohorts, model-based Wald tests show an inflation of type I error (**Figure 5**), as the observed distribution of interaction p-values deviates from the expected distribution under the null hypothesis. The deviation was much more pronounced in the unbalanced full cohort than in the 1:3 case-control cohort, showing that the problem with misspecification is further amplified when coupled with unbalanced case-control ratios. The sandwich variance-based Wald tests also show some degree of type I error inflation, especially in the full cohort. The inflation was remedied after we modeled the main effect of BMI flexibly using GAM. This example shows that main effect misspecification can lead to inflated type I error.

Example 2: Power gain due to more accurate modeling of main effects

We looked at the relationship between two continuous variables, age (independent variable) and BMI (outcome), and whether there is an interaction of age with sex on BMI. We used all 38,162 individuals from the same cohort described in the previous example.

A generalized additive model for BMI as a function of age revealed a nonlinear relationship (**Supplementary Figure S2A**). We then constructed a series of generalized linear models (described in **Table 2**) for BMI using age and sex to explore the impact of accounting and not accounting for the nonlinearity of the main effect on the test of interaction. **Table 2** reports estimates of coefficients and p-values associated with the terms included in each model. **Supplementary Figure S3** plots BMI by age groups, stratified by sex to visually depict the interaction structure. Figure S3 shows an apparent sex and age interaction as the effect of age on BMI was larger for males than for females for individuals with age less than 65. The model-based

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

Wald test with a linear main effect for age leads to a p-value of 8.53×10^{-4} and the sandwich variance-based Wald test leads to a p-value of 5.14×10^{-4}. Both tests are statistically significant. The Wald test based on a model where the main effect of age is modeled using GAM leads to a much smaller p-value (5.52×10^{-6}). It is not possible to know the "truth" in any given data analysis, thus, our explanation cannot be proven and alternative explanations cannot be ruled out. If interaction truly does not exist, it is still possible to see a significant p-value from the model-based Wald test with a linear main effect due to type I error inflation. However, if this were the case, it will be unlikely to observe a highly significant p-value from the GAM-based method as this method does not have inflated type I error. Therefore, the considerably smaller p-value from GAM-based method is most likely due to increased power by modeling the main effect flexibly and reducing the residual error. This example demonstrates that when the interaction effect is non-null, flexible specification of main effect can offer enhanced power in detecting interaction effect, though there are more parameters in the model to estimate.

Discussion

We consider the specific problem of main effect misspecification as linear terms when they are truly non-linear and its potential to lead to possibly severe type I error inflation in testing the interaction between two factors. We evaluated two simple strategies for addressing the problem with main effect misspecification. Namely, robust inference based on sandwich variance estimates and flexibly modeling the main effect using nonparametric methods such as GAM, using asymptotic theory and simulation studies. Our results show that for a linear regression model with a continuous outcome and two independent factors, replacing the model-based variance estimate with the sandwich variance estimate can lead to a valid test for interaction asymptotically. This result holds regardless of whether the main effects are correctly specified. However, this type of robustness using sandwich variance estimate does not hold in general for binary outcomes modeled using a logistic regression model, even under the assumption of independence of the two factors. Results from simulation studies are consistent with our asymptotic results. Further, based on our

This article is protected by copyright. All rights reserved.

simulation results, the sandwich score test converges faster than the sandwich Wald test as sample size increases and has better finite sample performance. The two examples from the Michigan Genomics Initiative further substantiate our points with actual data.

Using the sandwich variance estimate in a Wald or score test offers a simple solution for robust inference against main effect misspecification under the independence assumption for a continuous outcome. However, when the independence assumption does not hold or when the outcome is binary, this strategy will not be able to control the type I error rate. Moreover, even when these conditions are met and the sandwich method can control the type I error rate, it is still advantageous to try to model the main effects correctly or flexibly. We see that a Wald test combined with GAM for main effects can control the type I error rate in all settings considered here except one extreme case. In the case it does not completely control the type I error rate, it still considerably reduces type I error inflation and the performance improves as sample size increases. We note that the GAM method requires less sample size to control the type I error rate relative to the sandwich method when it works (**Figure 1C**). The strategy of flexibly modeling main effects using GAM is also appealing in terms of power, especially when the outcome is continuous. When the outcome is continuous, our simulation studies show that the GAM method leads to almost no power loss compared to a parametric model with correctly specified main effects in the settings considered here. Additionally, the GAM method is considerably more efficient than the sandwich method when type I error rate is controlled. When the outcome is binary, there is not a lot of loss of power relative to a correctly specified parametric main effect model. We comment that although we focused on Wald tests combined with GAM in our simulation studies, the strategy of using GAM or other nonparametric methods to model main effects flexibly can also be used with score test. Overall, the strategy to use GAM to model main effects flexibly offers an attractive and straightforward solution to robust and efficient testing of interaction under potential main effect misspecification. We have summarized our findings in a summary table (**Table 1**) as a useful guide for practitioners pursuing interaction analysis.

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

Our study complements previous work on main effect misspecification and tests of interaction. Among those, the most recent and closely related work is Sun et al.[7] Sun et al.[7] focus on theoretically identifying conditions under which valid tests can be obtained by using the sandwich estimator and further proposes to use a bootstrap inference with a corrected sandwich estimator to improve finite sample performances. Their simulation studies focus on Wald tests and scenarios where the robust inference can lead to valid inference asymptotically. Moreover, Sun et al.[7] only focus on type I error rate without considering power. However, a robust inference procedure can only solve the issue of main effect misspecification under somewhat restrictive conditions. Not all type I error inflation due to main effect misspecification can be fixed this way (e.g., generally, if independence does not hold for linear outcomes or if outcomes are binary). Our study considers both situations where the usual tests can and cannot be fixed by using a robust statistical inference. Further, it provides a solution that performs well in terms of both type I error rate and power for situations where valid tests cannot be obtained by using a robust inference. We consider the finite sample performance and the large sample properties of both Wald test and Score tests. In addition to the type I error rate, we focus on the power of various solutions under various situations as well. We provide an overall picture and improved understanding of various methods for tests of interaction when main effects may be possibly misspecified and provide practical guidance for data analysts. We also comment that the robustness property of the usual tests as shown in our results 1 and 2 can be viewed as a special case of the general results studied by Vansteelandt et al.[2] and Tchetgen Tchetgen[4] on multiply robust inference from the perspective of semiparametric theory. For if the test of interaction is robust to misspecification of the main effects, it must asymptotically be equivalent to the class of test statistics that are multiply robust.

Several limitations and possible extensions of this study exist. First, we focus on the setting where one does not adjust for other covariates in the model. Similar results and insights from our study can apply to the case when covariates adjustment is needed under additional assumptions. For

This article is protected by copyright. All rights reserved.

example, He et al.[8] show a similar robustness property as our results 1 and 2 under the assumption that other covariates can be divided into two parts and each part is correlated with either X_1 or X_2 but not both. In Sun et al[8], a similar condition for covariates is assumed. However, we comment that the robustness as in results 1 and 2 does not hold in general under the assumption of independence of X_1 and X_2 conditional on other covariates. Second, our results show that sample size is an important factor in type I error inflation. For continuous outcomes, although n=500 is usually considered relatively large for a model with four parameters when the model is correctly specified, it may not be large enough for robust inference using the sandwich variance estimate when the model is severely misspecified. Usually, the sandwich variance-based score test has better finite sample performance than the corresponding Wald test and extremely large ($> 10^5$) sample size may be needed for some extreme cases for the sandwich Wald test to work well. So small sample modification, for example, the Bootstrap Inference with Corrected Sandwich (BICS) procedure proposed in Sun and et al.[8] may be necessary in practice. Third, the strategy of using GAM is quite appealing in terms of power when outcome is continuous and is almost as powerful as the ideal case where main effects are correctly specified in a parametric model. However, when the outcome is binary, there is still room for improvement in power, representing an important direction for future research. Forth, our simulation study only considers interaction between two variables. When the number of variables in the model increases to, for example, three, the inference on interaction becomes more challenging. The performance of tests on interactions among multiple variables is unknown. Finally, misspecification of the interaction effect needs to be considered in addition to main effect misspecification.

ACKNOWLEDGEMENTS: The results reported herein were supported by grant DMS 1712933 from the National Science Foundation.

DATA AVAILABILITY: Data cannot be shared publicly due to patient confidentiality. The data underlying the results presented in the study are available from University of Michigan Medical

This article is protected by copyright. All rights reserved.

Author Manuscript

School Central Biorepository at https://research.medicine.umich.edu/our-units/central-biorepository/get-access and from the UK Biobank at http://www.ukbiobank.ac.uk/register-apply/ for researchers who meet the criteria for access to confidential data.

Author Manuscript

REFERENCES

1. Bateson W. *Mendel's Principles of Heredity*. Cambridge, United Kingdom: Cambridge University Press, 1909.
2. Vansteelandt S, Vanderweele TJ, Tchetgen E.J., Robins JM. Multiply robust inference for statistical interactions. *J Am Stat Assoc* 2008;**103**(484):1693-1704.
3. Rosenblum M, van der Laan MJ. Using regression models to analyze randomized trials: asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics* 2009;**65**(3):937-45.
4. Tchetgen Tchetgen EJ. Multiple-Robust Estimation of an Odds Ratio Interaction. Harvard University Biostatistics Working Paper Series 2012, paper 142.
5. Tchetgen Tchetgen EJ, Kraft P. On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified. *Epidemiology* 2011;**22**(2):257-61.
6. Cornelis MC, Tchetgen Tchetgen EJ, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *Am J Epidemiol* 2012;**175**(3):191-202.
7. Lin X, Lee S, Christiani DC, Lin X. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* 2013;**14**(4):667-81.
8. Sun R, Carroll RJ, Christiani DC, Lin X. Testing for gene-environment interaction under exposure misspecification. *Biometrics* 2018;**74**(2):653-662.
9. He Z, Zhang M, Lee S, Smith JA, Kardia SLR, Diez Roux AV, Mukherjee B. Set-Based Tests for the Gene-Environment Interaction in Longitudinal Studies. *J Am Stat Assoc* 2017;**112**(519):966-978.

This article is protected by copyright. All rights reserved.

Author Manuscript

10. Voorman A, Lumley T, McKnight B, Rice K. Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PLoS One* 2011;**6**(5):e19416.

11 Dai JY, LeBlanc M, and Kooperberg C. Semiparametric estimation exploiting covariate independence in two-phase randomized trials. Biometrics. 2009 Mar;65(1):178-87.

12. Dai JY, Kooperberg C, Leblanc M, Prentice RL. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. Biometrika. 2012 Dec;99(4):929-944.

13. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. 1st ed Chapman and Hall/CRC, 1990.

14. Duchon J. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In: Schempp W, Zeller K, eds. *Constructive Theory of Functions of Several Variables*. Lecture Notes in Mathematics. Vol. 571 Springer, Berlin, Heidelberg, 1977;85-100.

15. Friedmand JH and Popescu BE. Predictive Learning Via Rule Ensemble. The Annals of Applied Statistics. 2008; 2(3): 916-954.

16. Marjolein Fokkema and Benjamin Christoffersen (2019). pre: Prediction Rule Ensembles. R package version 0.7.1. https://CRAN.R-project.org/package=pre

17. Fritsche LG, Gruber SB, Wu Z, Schmidt EM, Zawistowski M, Moser SE, Blanc VM, Brummett CM, Kheterpal S, Abecasis GR, Mukherjee B. Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *Am J Hum Genet.* 2018 Jun 7;102(6):1048-1061. doi:10.1016/j.ajhg.2018.04.001.

18. Boos DD, Stefanski LA. *Essential Statistical Inference: Theory and Methods* Springer US, 2013.

19. van der Vaart AW. *Asymptotic Statistics*. Cambridge, United Kingdom: Cambridge University Press, 1998.

This article is protected by copyright. All rights reserved.

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

Appendix

(A) Proof of Result 1

Suppose we are interested in testing the interaction between X_1 and X_2 based on data (Y_i, X_{1i}, X_{2i}), $i = 1, \dots, n$, iid across i, where Y_i is the quantitative outcome for subject i, and X_{1i} and X_{2i} are independent variables. Without loss of generality, we suppose Y_i, X_{1i}, X_{2i} are all centered. Suppose under the null hypothesis, the true model is

$$Y_i = h_1(X_{1i}) + h_2(X_{2i}) + \epsilon_i,$$

where h_1 and h_2 are unknown functions, e_i is an error term with mean 0 and independent of X_{1i} and X_{2i}. Suppose instead we assume the following working model

$$E(Y_i|X_{1i}, X_{2i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i},$$

and we test the null hypothesis of no interaction by testing $\mathrm{H}_0: \beta_3 = 0$.

Consistency: The ordinary least square estimator $\hat{\beta} = \left[\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right]^T$ satisfies the estimating equation:

$$\frac{1}{n}\sum_i \left\{[1, X_{1i}, X_{2i}, X_{1i}X_{2i}]^T \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{1i} X_{2i}\right)\right\} = 0. \qquad \text{(A1)}$$

Under standard regularity conditions and by a standard M-estimation (also referred to as Z-estimation) theory (Boos and Stefanski, 2013[18]; van der Vaart, 2012[19]), $\hat{\beta}$ converges in probability to $\beta^* = [\beta_0^*, \beta_1^*, \beta_2^*, \beta_3^*]^T$, which satisfies the "population" version of this last estimating equation, i.e.,

$$E\{[1, X_1, X_2, X_1X_2]^T (Y - \beta_0^* - \beta_1^* X_1 - \beta_2^* X_2 - \beta_3^* X_1 X_2)\} = 0. \qquad \text{(A2)}$$

We can derive, by solving the above equation, that $\beta_0^* = 0$, $\beta_1^* = \frac{E(X_1 Y)}{E(X_1^2)}, \beta_2^* = \frac{E(X_2 Y)}{E(X_2^2)}$, and $\beta_3^* = \frac{E(X_1 X_2 Y)}{E(X_1^2 X_2^2)}$. Regarding the numerator of β_3^*, note that

$$E(X_1 X_2 Y) = E[X_1 X_2 \{h_1(X_1) + h_2(X_2) + \epsilon\}]$$
$$= E\{X_1 h_1(X_1) X_2\} + E\{X_1 X_2 h_2(X_2)\} + E(X_1 X_2 \epsilon)$$

This article is protected by copyright. All rights reserved.

$$= E\{X_1 h_1(X_1)\}EX_2 + EX_1 E\{X_2 h_2(X_2)\} + E(X_1X_2)E(\epsilon)$$
$$= 0$$

where the second equality is due to independence of X_1 and X_2, and the last equality is due to $EX_1 = EX_2 = 0$ because of centering. Therefore, $\hat{\beta}_3$ converges in probability to $\beta_3^* = 0$.

Asymptotical normality: Asymptotical normality follows as a standard result from M-estimation theory. Let $X_i = [1, X_{1i}, X_{2i}, X_{1i}X_{2i}]^T$ be the covariate vector for the i-th subject, $i = 1, \dots, n$. Equation (A1) can be written as

$$\frac{1}{n}\sum_i X_i\left(Y_i - X_i^T\hat{\beta}\right) = 0.$$

By a Taylor expansion of the left hand side of the above equation around β^*, we have

$$\frac{1}{n}\sum_i X_i(Y_i - X_i^T\beta^*) - \frac{1}{n}\sum_i X_i X_i^T\left(\hat{\beta} - \beta^*\right) + o_p(1) = 0.$$

Rearranging terms leads to

$$\sqrt{n}\left(\hat{\beta} - \beta^*\right) = \left(\frac{1}{n}\sum_i X_i X_i^T\right)^{-1}\frac{1}{\sqrt{n}}\sum_i X_i(Y_i - X_i^T\beta^*) + o_p(1)\,.$$

By Central Limit Theorem, $\frac{1}{\sqrt{n}}\sum_i\{X_i(Y_i - X_i^T\beta^*)\}$ converges in distribution to a normal distribution with mean $E\{X_i\left(Y_i - \beta^{*T}X_i\right)\} = 0$ and variance

$$E\{X_i X_i^T(Y_i - X_i^T\beta^*)^2\}.$$

By Slutsky Theorem, $\sqrt{n}\left(\hat{\beta} - \beta^*\right)$ converges in distribution to Normal $(0, \Sigma)$, where

$$\Sigma = \{E(X_i X_i^T)\}^{-1}E\{X_i X_i^T(Y_i - X_i^T\beta^*)^2\}\{E(X_i X_i^T)\}^{-1}\,,$$

and Σ can be consistently estimated by

$$\hat{\Sigma} = \left(\frac{1}{n}\sum_i X_i X_i^T\right)^{-1}\left\{\frac{1}{n-p}\sum_i X_i X_i^T\left(Y_i - X_i^T\hat{\beta}\right)^2\right\}\left(\frac{1}{n}\sum_i X_i X_i^T\right)^{-1},$$

Author Manuscript

This article is protected by copyright. All rights reserved.

where p is the dimension of X_i. Therefore, the asymptotic variance of $\hat{\beta}$ can be consistently estimated by $\frac{\Sigma}{\mathrm{n}}$, which equals $\hat{V}_{sandwich}(\hat{\beta}) = \frac{n}{(n-p)}(\sum_i X_i X_i^T)^{-1}(\sum_i X_i X_i^T \hat{\epsilon}_i^2)(\sum_i X_i X_i^T)^{-1}$ defined in the Methods Section. Regardless of whether the model is correctly specified or not, under the null hypothesis, the Wald test statistic with the empirical sandwich variance estimate $\frac{\hat{\beta}_3^2}{\hat{V}_{sandwich}(\hat{\beta}_3)} \sim \chi_1^2$, where $\hat{V}_{sandwich}(\hat{\beta}_3)$ is the diagonal element of $\hat{V}_{sandwich}(\hat{\beta})$ corresponding to β_3.

(B) Proof of Result 2

Unbiasedness of score: The score corresponding to β_3 is $S = \frac{1}{n}\sum_i S_i(\tilde{\beta}) = \frac{1}{n}\sum_i\{X_{1i}X_{2i}(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_{1i} - \tilde{\beta}_2 X_{2i})\}$, where $\tilde{\beta} = [\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2]^T$ is the ordinary least squares estimator under the null working model:

$$E(Y_i|X_{1i}, X_{2i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}.$$

Specifically, $\tilde{\beta} = [\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2]^T$ satisfies the estimating equation

$$\frac{1}{n}\sum_i\{[1, X_{1i}, X_{2i}]^T(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_{1i} - \tilde{\beta}_2 X_{2i})\} = 0,$$

and under standard regularity conditions, by M-estimation theory, it converges in probability to $\beta^{\#} = [\beta_0^{\#}, \beta_1^{\#}, \beta_2^{\#}]^T$, which satisfies the "population" version of the last equation,

$$E\{[1, X_{1i}, X_{2i}]^T(Y_i - \beta_0^{\#} - \beta_1^{\#} X_{1i} - \beta_2^{\#} X_{2i})\} = 0.$$

Solving the equation, we have $\beta_0^{\#} = 0$, $\beta_1^{\#} = \frac{E(X_1 Y)}{E(X_1^2)}$, $\beta_2^{\#} = \frac{E(X_2 Y)}{E(X_2^2)}$. It follows that, by law of large numbers and under regularity conditions, the score S converges in probability to

$$\begin{aligned} & E\{X_1 X_2(Y - \beta_0^{\#} - \beta_1^{\#} X_1 - \beta_2^{\#} X_2)\} \\ & = E[X_1 X_2\{h_1(X_1) - \beta_1^{\#} X_1\}] + E[X_1 X_2\{h_2(X_2) - \beta_2^{\#} X_2\}] + E(X_1 X_2 \epsilon) \qquad \text{(A3)} \\ & = E[X_1\{h_1(X_1) - \beta_1^{\#} X_1\}]EX_2 + EX_1 E[X_2\{h_2(X_1) - \beta_1^{\#} X_1\}] + E(X_1 X_2)E(\epsilon) \\ & = 0. \end{aligned}$$

This article is protected by copyright. All rights reserved.

Author Manuscript

Therefore, the score is unbiased for zero.

Asymptotic normality: By a Taylor expansion around $\beta^{\#}$, we have

$$\frac{1}{\sqrt{n}}\sum_i S_i(\tilde{\beta}) = \frac{1}{\sqrt{n}}\sum_i S_i(\beta^{\#}) - \frac{1}{\sqrt{n}}\left(\sum_i X_{1i}X_{2i}\, X_{o,i}\right)\left(\tilde{\beta} - \beta^{\#}\right) + o_p(1) \qquad \text{(A4)}$$

where $X_{o,i} = [1, X_{1i}, X_{2i}]^T$. By an argument similar to that in the proof for result 1, we have

$$\sqrt{n}\left(\tilde{\beta} - \beta^{\#}\right) = \left(\frac{1}{n}\sum_i X_{0,i}\, X_{0,i}^T\right)^{-1} \frac{1}{\sqrt{n}}\sum_i X_{o,i}\left(Y_i - X_{o,i}^T\beta^{\#}\right) + o_p(1)\,,$$

and substituting this into (A4) we have $\frac{1}{\sqrt{n}}\sum_i S_i(\tilde{\beta})$

$$= \frac{1}{\sqrt{n}}\sum_i S_i(\beta^{\#}) - \left(\frac{1}{n}\sum_i X_{1i}X_{2i}\, X_{o,i}^T\right)\left(\frac{1}{n}\sum_i X_{o,i}\, X_{o,i}^T\right)^{-1} \frac{1}{\sqrt{n}}\sum_i X_{o,i}\left(Y_i - X_{o,i}^T\beta^{\#}\right) + o_p(1)$$

$$= \frac{1}{\sqrt{n}}\sum_i X_{1i}X_{2i}\left(Y_i - X_{o,i}^T\beta^{\#}\right) - E(X_{1i}X_{2i}X_{o,i}^T)\{E\left(X_{o,i}\, X_{o,i}^T\right)\}^{-1} \frac{1}{\sqrt{n}}\sum_i X_{o,i}\left(Y_i - X_{o,i}^T\beta^{\#}\right) + o_p(1)$$

$$= \frac{1}{\sqrt{n}}\sum_i \left[-E(X_{1i}X_{2i}X_{o,i}^T)\{E\left(X_{o,i}\, X_{o,i}^T\right)\}^{-1}, 1\right] X_i\left(Y_i - X_{o,i}^T\beta^{\#}\right) + o_p(1),$$

where $X_i = [1, X_{1i}, X_{2i}, X_{1i}X_{2i}]^T$ as defined before. By Central Limit Theorem, $\frac{1}{\sqrt{n}}\sum_i S_i(\tilde{\beta})$ converges to a normal distribution with mean 0 because $E\{X_{1i}X_{2i}\left(Y_i - X_{o,i}^T\beta^{\#}\right)\} = 0$ as shown above and $E\left\{X_{o,i}\left(Y_i - X_{o,i}^T\beta^{\#}\right)\right\} = 0$ by definition of $\beta^{\#}$, and with variance

$A\, E\left\{X_i X_i^T\left(Y_i - X_{o,i}^T\beta^{\#}\right\}^2\right\} A^T$,

where $A = \left[-E(X_{1i}X_{2i}X_{o,i}^T)\{E\left(X_{0,i}X_{0,i}^T\right)\}^{-1}, 1\right]$. The variance can be consistently estimated by the empirical variance estimator $\tilde{A}\{\frac{1}{n-p}\sum_i X_i X_i^T\left(Y_i - X_{o,i}^T\tilde{\beta}\right)^2\}\tilde{A}^T$, where

Author Manuscript

This article is protected by copyright. All rights reserved.

$$\tilde{A} = \left[-\left(\sum_{i=1}^{n} X_{1i} X_{2i} X_{o,i}^T \right) \left(\sum_{i=1}^{n} X_{o,i}\, X_{o,i}^T \right)^{-1}, 1 \right].$$

Therefore, regardless of whether the model for the main effect of X_1 and X_2 is correctly specified or not, the score test statistic $S^2/\hat{V}_{sandwich}(S)$ follows a χ_1^2 distribution asymptotically, when conditions stated in result 2 are satisfied.

Comment: For a logistic regression for binary outcomes, the score converges to $E\{X_1 X_2 \text{expit}(Y - \beta_o - \beta_1 X_1 - \beta_2 X_2)\}$ for some $\beta = [\beta_0, \beta_1, \beta_2]^T$, where expit $(\mu) = \exp(\mu) / \{1 + \exp(\mu)\}$, and without making further assumptions we cannot separate expit$(Y - \beta_1 X_1 - \beta_2 X_2)$ into terms that involve only X_1 or X_2 as in (A3) above. As a result, in general $E\{X_1 X_2 \text{expit}(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2)\}$ is not equal to zero when the main effect model is misspecified even under the assumption of independence. Therefore, for a logistic regression model, the score test lacks the robustness against main effect misspecification. Although not as obvious, the reason for non-robustness of the Wald test is similar. As a result, under the null hypothesis when main effects are misspecified, the estimator of β_3 does not converge to 0 without making further assumptions on main effects. Therefore, for logistic regression the robustness of testing for interaction against main effect misspecification does not hold.

(C) Effect of Overfitting the Main Effects

We provide some intuition and explanation for why the use of flexible GAM to model main effects of X_1 and/or X_2 does not reduce power under the independence assumption of X_1 and X_2 for continuous outcomes. The result is not specific to the use of GAM and methods other than GAM can be used to model main effect flexibly. This phenomenon is due to a general result that (informally) overfitting the main effect does not reduce power asymptotically under the independence assumption. Taking a simple setting as an example, we show this explicitly. Specifically, suppose the true model for a continuous outcome is $Y_i = \beta_0 + \beta_{11} X_{1i} +$

Author Manuscript

This article is protected by copyright. All rights reserved.

$\beta_{12}X_{1i}^2 + \cdots + \beta_{1p}X_{1i}^p + \beta_2 X_{2i} + \beta_3 X_{1i}X_{2i} + \epsilon_i$, where variance of ϵ_i is σ^2. Instead one tests interaction using a Wald test based on an overfitted main effect model, specified as $Y_i = \beta_0 + \beta_{11}X_{1i} + \beta_{12}X_{1i}^2 + \cdots + \beta_{1q}X_{1i}^q + \beta_2 X_{2i} + \beta_3 X_{1i}X_{2i} + \epsilon_i$, where $q > p$ such that the main effect of X_{1i} includes unnecessary higher order polynomial terms. Directly applying results in Appendix A, it is easy to check that the estimator for β, denoted by $\hat{\beta}$, based on the overfitted model solves the estimating equation

$$\frac{1}{n}\sum_i \left\{\left[1, X_{1i}, \ldots, X_{1i}^q, X_{2i}, X_{1i}X_{2i}\right]^T \left(Y_i - \hat{\beta}_0 - \hat{\beta}_{11}X_{1i} - \cdots - \hat{\beta}_{1q}X_{1i}^q - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{1i}X_{2i}\right)\right\} = 0.$$

We denote the limit of $\hat{\beta}$ by β^* and it satisfies the population version of the above equation. As in Appendix A, it is easy to check that $\beta_3^* = \frac{E(X_1X_2Y)}{E(X_1^2X_2^2)} = \beta_3$, which is nonzero if the alternative hypothesis is true. In addition, $\sqrt{n}\left(\hat{\beta} - \beta^*\right)$ converges to a normal distribution with variance equal to

$$\Sigma = \{E(X_iX_i^T)\}^{-1}E\{X_iX_i^T(Y_i - X_i^T\beta^*)^2\}\{E(X_iX_i^T)\}^{-1},$$

where $X_i = \left[1, X_{1i}, \ldots, X_{1i}^q, X_{2i}, X_{1i}X_{2i}\right]^T$. By the independence of X_1 and X_2 and assuming Y, X_1, X_2 are centered, we can show that $E(X_iX_i^T) = diag\,(A, E(X_1^2X_2^2))$ for some matrix A because it is easy to check that $E(X_1X_2), E(X_1^2X_2), \ldots, E(X_1^{q+1}X_2)$ all equal to zero. Therefore, $\{E(X_iX_i^T)\}^{-1} = diag\,(A^{-1}, \frac{1}{E(X_1^2X_2^2)})$. The middle term of Σ, $E\{X_iX_i^T(Y_i - X_i^T\beta^*)^2\}$=$\sigma^2E(X_iX_i^T)$. Therefore, $\Sigma = \sigma^2 diag\,(A^{-1}, \frac{1}{E(X_1^2X_2^2)})$. It follows that $\sqrt{n}\left(\hat{\beta}_3 - \beta_3\right)$ converges to a normal distribution with mean zero and variance $\sigma^2/E(X_1^2X_2^2)$. The asymptotically distribution is exactly the same as the one based on a correctly specified model without overfitting and the same as the one had the true main effect been known without having to estimate it. Therefore, the Wald tests based on the overfitted model and the true model have the same asymptotic distribution and therefore lead to the same power. When one uses GAM to flexibly model the main effect of X_1 (and/or X_2), the basis functions used to approximate the main effect are not polynomial functions but linear spline terms. However, regardless it still holds that $E(l(X_1)X_2) = 0$ and $E(l(X_2)X_1) =$

Author Manuscript

This article is protected by copyright. All rights reserved.

0, where l is an arbitrary function. Therefore, the argument above still applies. Specifically, when X_1 is modeled using $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_{11}(X_{1i} - \tau_1)_+ + \cdots + \beta_{1p}\left(X_{1i} - \tau_p\right)_+ + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i$ using penalized regression, where $(X_{1i} - \tau_k)_+$, $k = 1, \ldots, p$, are linear spline terms, then the estimator for $\beta = \left(\beta_0, \beta_1, \beta_{11}, \ldots, \beta_{1p}, \beta_2, \beta_3\right)^T$ has variance and covariance matrix proportional to $\sigma^2 \{E(X_i X_i^T) + \lambda^2 D\}^{-1} E(X_i X_i^T)\{E(X_i X_i^T) + \lambda^2 D\}^{-1}$, where $X_i = (1, X_{1i}, (X_{1i} - \tau_1)_+, \ldots, \left(X_{1i} - \tau_p\right)_+, X_{2i}, X_3)$, λ is a tuning parameter for roughness, and D is a diagonal matrix where the diagonal terms corresponding to the linear spline terms are one and the other terms are zero. Using results that $E(l(X_1)X_2) = 0$ and $E(l(X_2)X_1) = 0$ and similar arguments as above, it can be checked that the asymptotic variance of $\hat{\beta}_3$ is again$\sigma^2 / E(X_1^2 X_2^2)$. The above derivations and arguments provide an explicit and intuitive explanation for why overfitting the main effect model does not reduce power for continuous outcomes under the independence assumption of X_1 and X_2. However, this result does not hold in general without the independence assumption, although our simulation studies show that the impact on power is small. Finally, we comment that in general overfitting the interaction term usually does significantly affect power.

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

Table 1. Guidelines for choosing method for interaction analysis under misspecification of main effects. We bold the method that is preferred under each scenario.

Method	Outcome: Continuous Factors Independent	Outcome: Continuous Factors Correlated	Outcome: Binary Factors Independent	Outcome: Binary Factors Correlated
Wald Model Based	Type 1 error: inflated Power comparison not valid	Type I error: inflated Power comparison not valid	Type I error: inflated Power comparison not valid	Type I error: inflated Power comparison not valid
Wald Sandwich	Type 1 error: Nominal Power: loss of power depending on the degree of misspecification	Type I error: inflated Power comparison not valid	Type I error: inflated Power comparison not valid	Type I error: inflated Power comparison not valid
Score Sandwich	Type I error: Nominal Power: loss of power depending on the degree of misspecification	Type I error: inflated Power comparison not valid	Type I error: inflated Power comparison not valid	Type I error: inflated Power comparison not valid
GAM1	**Type I error: Nominal if main effect of X_2 is linear** **Power: almost as powerful as the correct model if main effect of X_2 is linear**	**Type I error: Nominal if main effect of X_2 is linear** **Power: almost as powerful as the correct model if main effect of X_2 is linear**	**Type I error: Nominal if main effect of X_2 is linear** **Power: some loss of power relative to the correct parametric model**	**Type I error: Nominal if main effect of X_2 is linear** **Power: some loss of power relative to the correct parametric model**
GAM2	**Type I error: Nominal** **Power: almost as powerful as the correct model**	**Type I error: Nominal** **Power: almost as powerful as the correct model**	**Type I error: Nominal** **Power: more loss of power relative to GAM1 when the extra smooth term is unnecessary**	**Type I error: Nominal** **Power: more loss of power relative to GAM1 when the extra smooth term is unnecessary**

This article is protected by copyright. All rights reserved.

Author Manuscript

	Models			
Variables	0	1	2	3
Age	0.018 (**2.46×10^{-16}**)	0.026 (**5.71×10^{-16}**)	-	-
Sex	0.230 (**1.37×10^{-3}**)	0.233 (**1.21×10^{-3}**)	0.096 (1.76×10^{-1})	0.100 (1.59×10^{-1})
Age-sex interaction	-	-0.015 (**8.53×10^{-4}**) **[5.14×10^{-4}]**	-	-0.020 (**5.52×10^{-6}**)
MSE	48.403	48.390	47.087	47.063

Note: sex variable is coded as an indicator for female sex. P-values less than 0.05 are bolded. P-values in parentheses and brackets are computed using model-based and sandwich variance, respectively.

Model 0**:** BMI = $\boldsymbol{\beta_0}$ + $\boldsymbol{\beta_A}$Age + $\boldsymbol{\beta_S}$Sex + $\boldsymbol{\epsilon}$

Model 1: BMI = $\boldsymbol{\beta_0}$ + $\boldsymbol{\beta_A}$Age + $\boldsymbol{\beta_S}$Sex + $\boldsymbol{\beta_{AS}}$Age*Sex + $\boldsymbol{\epsilon}$

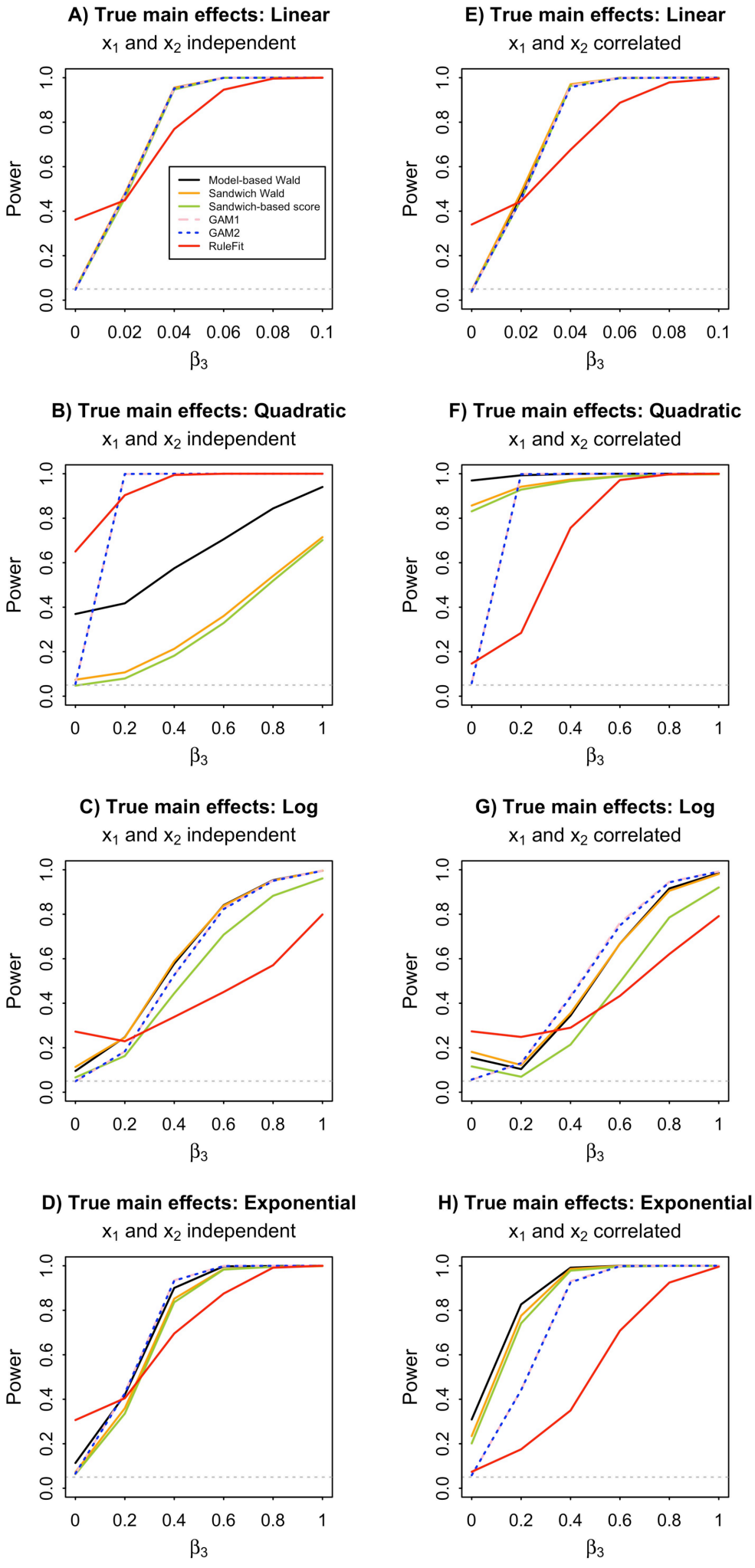
Model 2: BMI = $\boldsymbol{\beta_0}$ + s(Age) + $\boldsymbol{\beta_S}$Sex + $\boldsymbol{\epsilon}$

Model 3: BMI = $\boldsymbol{\beta_0}$ + s(Age) + $\boldsymbol{\beta_S}$Sex + $\boldsymbol{\beta_{AS}}$Age*Sex + $\boldsymbol{\epsilon}$

Table 2. Example 2: comparing models for BMI as functions of age and sex.

This article is protected by copyright. All rights reserved.

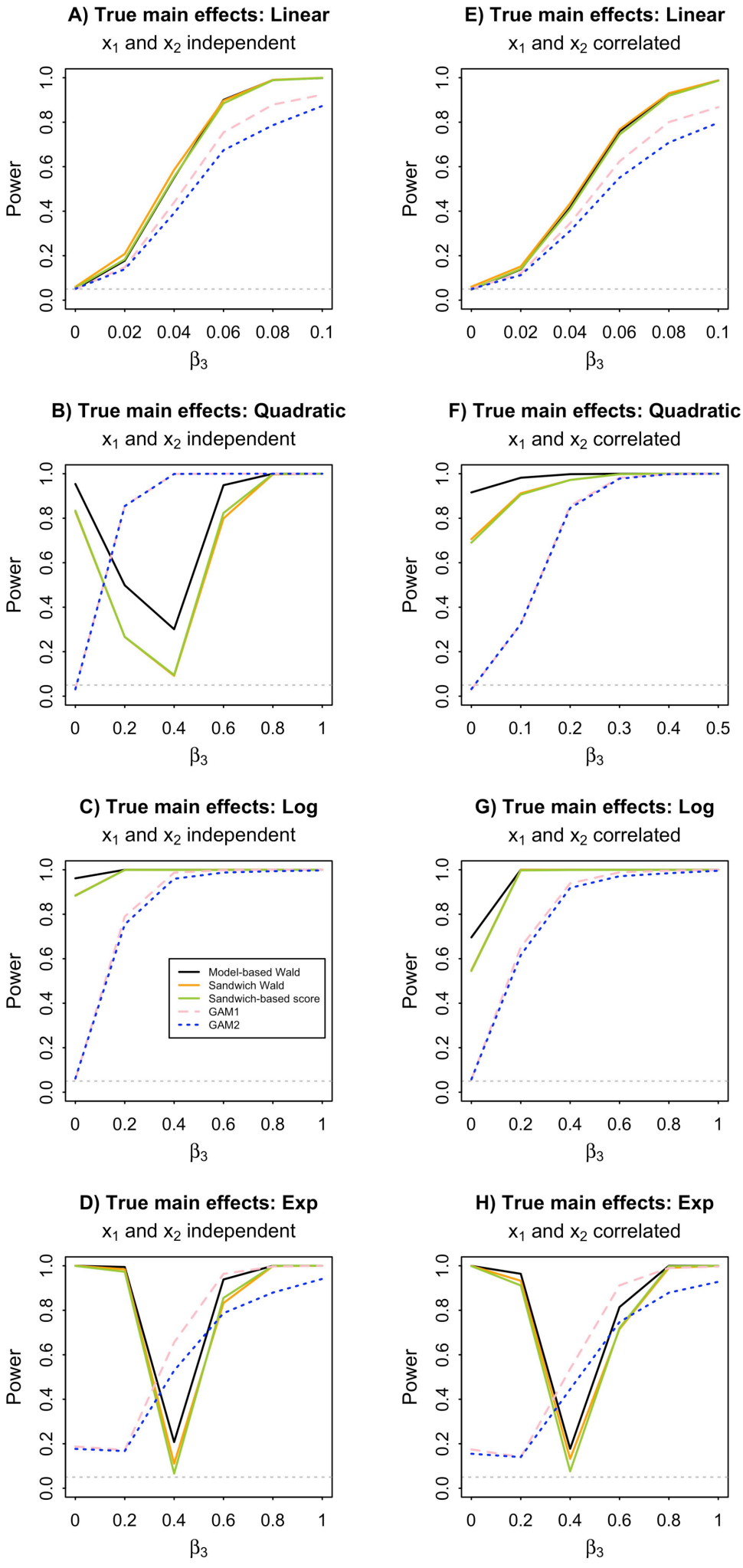
Author Manuscript



SIM_8505_Figure1.tiff

This article is protected by copyright. All rights reserved.

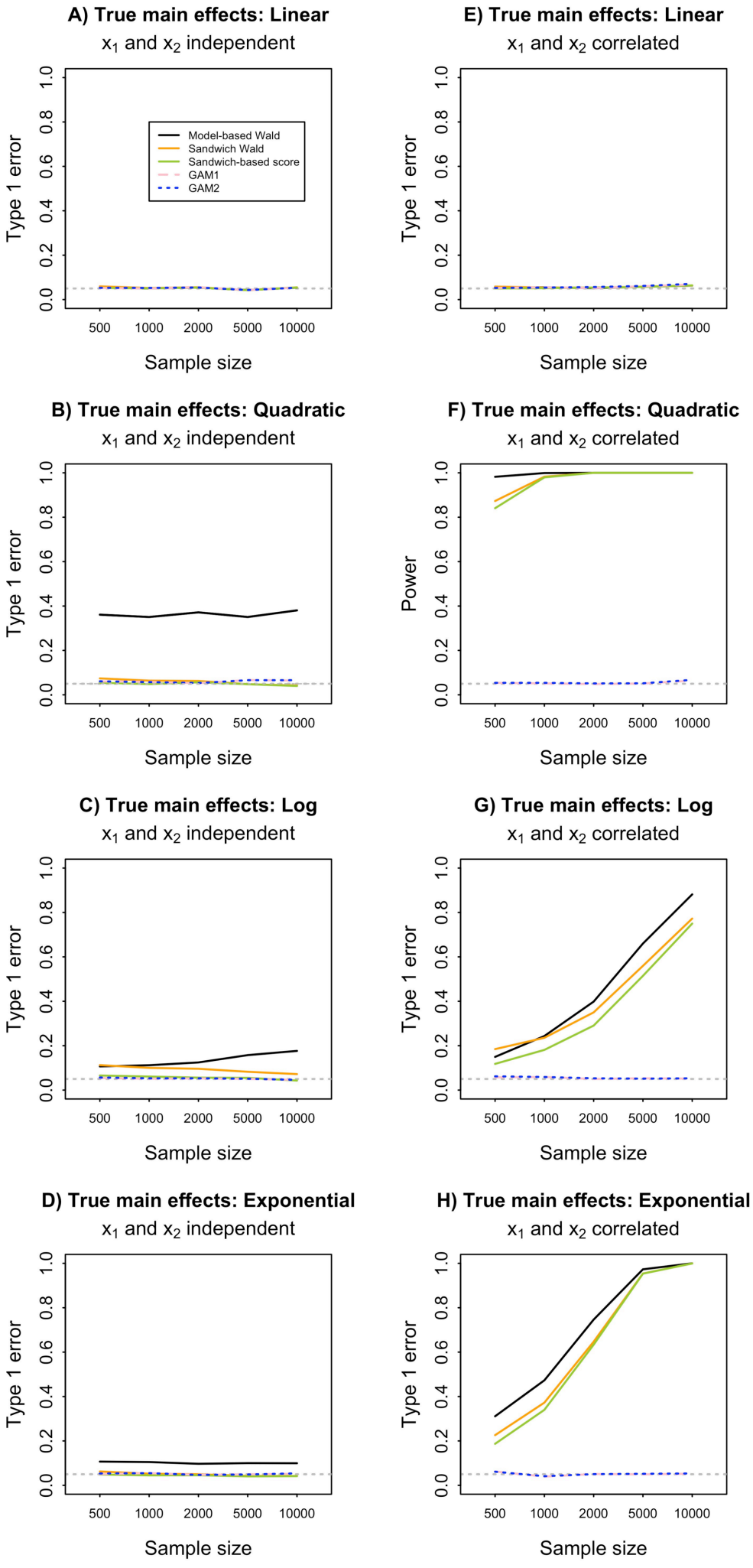
Author Manuscript



SIM_8505_Figure2.tiff

This article is protected by copyright. All rights reserved.

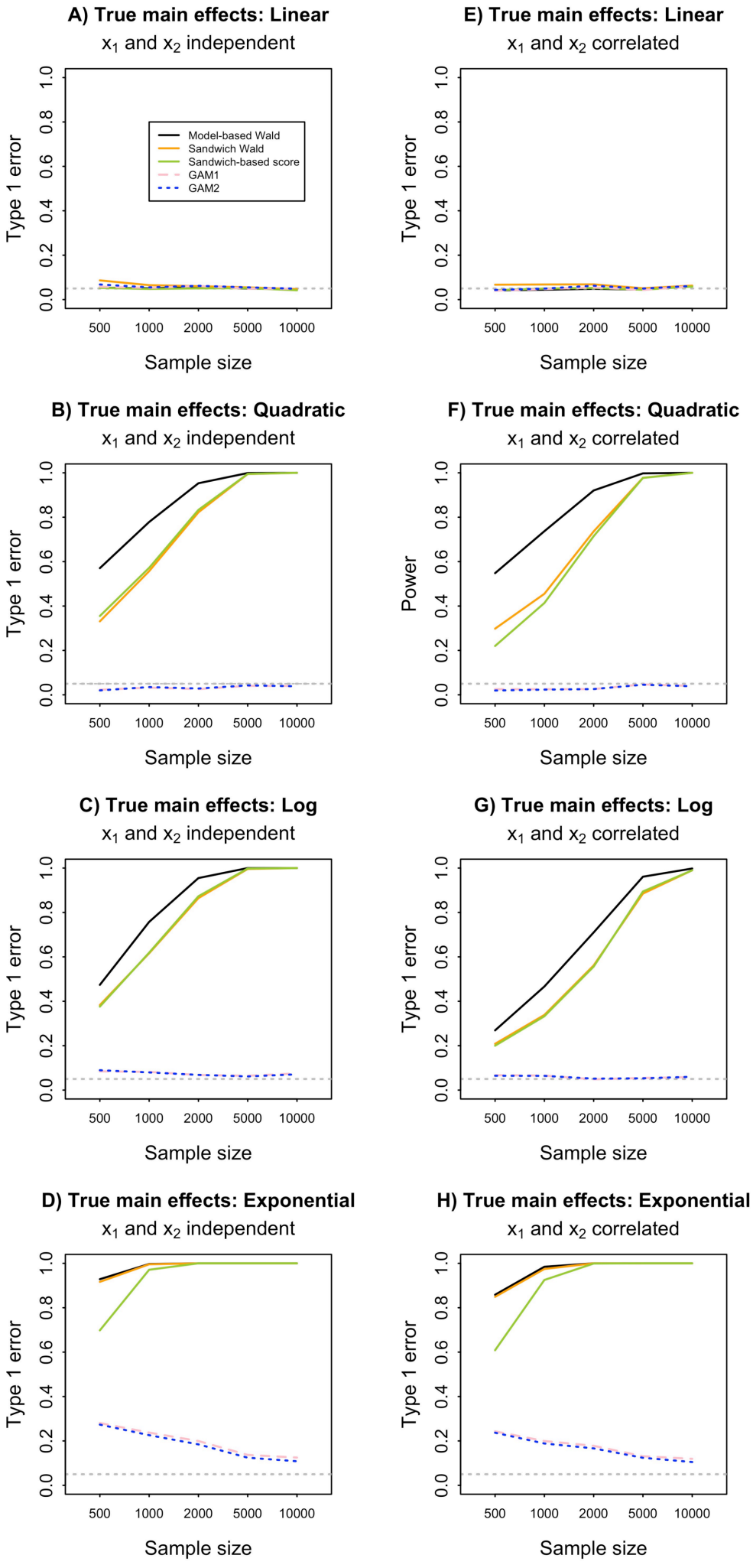
Author Manuscript



SIM_8505_Figure3.tiff

This article is protected by copyright. All rights reserved.

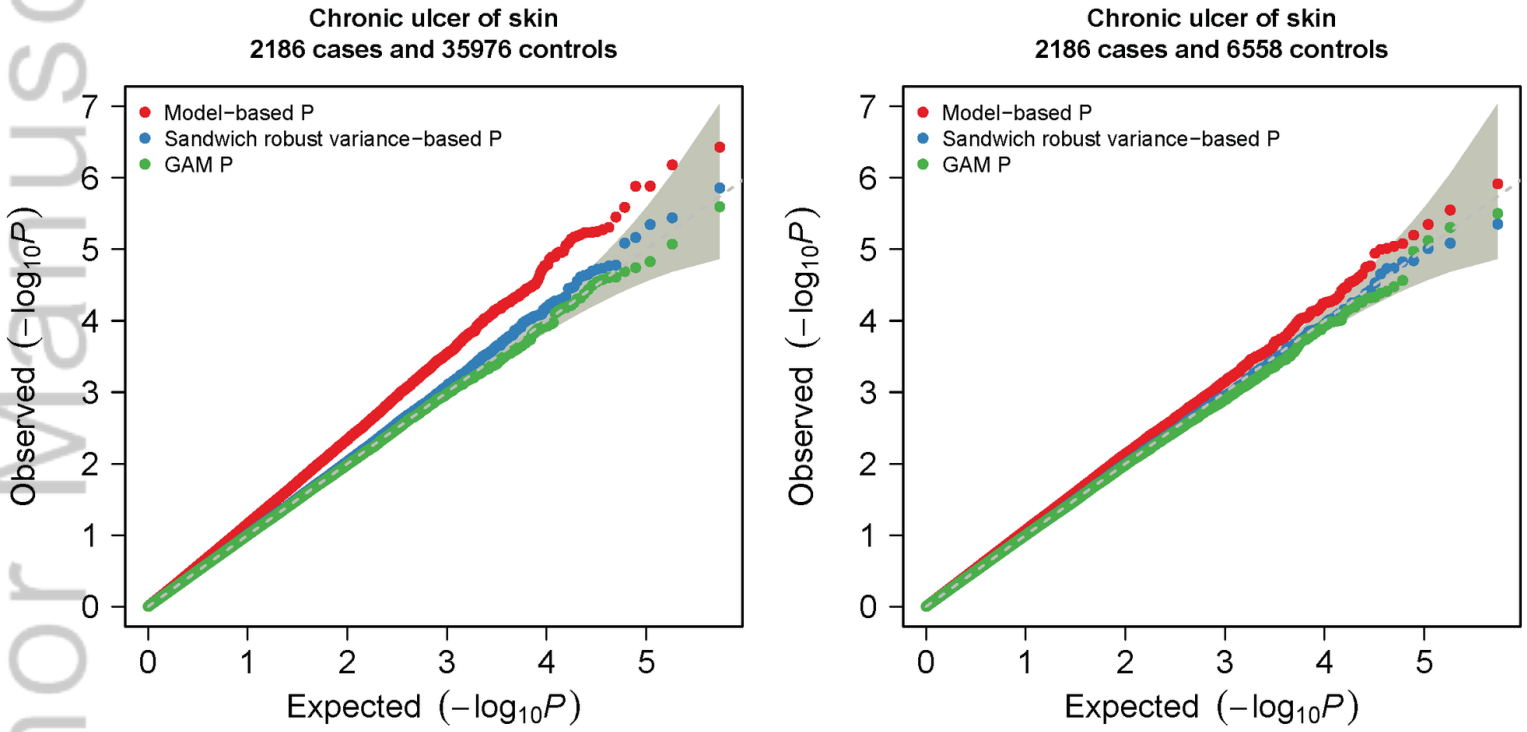
Author Manuscript



SIM_8505_Figure4.tiff

This article is protected by copyright. All rights reserved.

Author Manuscript



SIM_8505_QQ_plots.tif

This article is protected by copyright. All rights reserved.