Published in final edited form as:

Stat Med. 2013 March 30; 32(7): 1164–1190. doi:10.1002/sim.5628.

Evaluation of removable statistical interaction for binary traits

Jaya M. Satagopan¹ and Robert C. Elston²

¹Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

²Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA

Abstract

This paper is concerned with evaluating whether an interaction between two sets of risk factors for a binary trait is removable and fitting a parsimonious additive model using a suitable link function to estimate the disease odds (on the natural logarithm scale) when an interaction is removable. Statisticians define the term “interaction” as a departure from additivity in a linear model on a specific scale on which the data are measured. Certain interactions may be eliminated via a transformation of the outcome such that the relationship between the risk factors and the outcome is additive on the transformed scale. Such interactions are known as removable interactions. We develop a novel test statistic for detecting the presence a removable interaction in case-control studies. We consider the Guerrero and Johnson family of transformations and show that this family constitutes an appropriate link function for fitting an additive model when an interaction is removable. We use simulation studies to examine the type I error and power of the proposed test and to show that an additive model based on the Guerrero and Johnson link function leads to more precise estimates of the disease odds parameters and a better fit when an interaction is removable. The proposed test and use of the transformation are illustrated using case-control data from three published studies. Finally, we indicate how one can check that, after transformation, no further interaction is significant.

Keywords

Analysis of variance; curvature; independence; interaction effect; link function; main effect; residuals; score statistic; Tukey’s test; transformation; unbalanced data

Introduction

There is a long-standing interest in the evaluation of interactions in genetic and epidemiology studies. The ability to conduct candidate-gene and high-throughput association studies has accelerated the efforts to examine gene-gene and gene-environment interactions in relation to disease risk using a variety of statistical models (see [1 – 6] and several other works reviewed in [7 – 9]). The interest in interactions is also stimulated by the hope that significant interaction effects identified using statistical models can provide insights into biological interactions that underpin the disease [7, 10]. This has also caused some confusion regarding the very meaning of the term “interaction” and substantial

Correspondence to: Jaya M. Satagopan, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 307, East 63rd Street, 3rd Floor, New York, NY 10065, USA, Phone: (+1) 646-735-8122, FAX: (+1) 646-735-0010, satagopj@mskcc.org.

confusion as to what is actually being estimated or tested from a biological viewpoint [11, 12].

The term “interaction” is defined in statistical analyses as a departure from additivity in a linear model on a selected scale of measurement [13, 14]. This departure, also referred to as “statistical interaction”, measures the change in the effect on the outcome of one risk factor when the values(s) of another risk factor(s) is (are) altered. Public health burden and policy implications are generally assessed on the scale of risk. A departure from additivity on this scale is an interaction that is often denoted as “synergy” in the epidemiology literature [15]. The term “biological interaction” refers to the joint action of two or more risk factors, which can occur even when there is no statistical interaction [12]. Consider, for example, the etiology of cancer. A normal cell undergoes a malignant transformation through a process termed tumor progression, which is driven by a succession of cellular abnormalities (for example, somatic mutations) that accumulate over multiple stages [16, 17]. Under this paradigm, the risk factors associated with disease must be involved at the biological level. This can occur even in the absence of a statistical interaction. Thus, a simple additive statistical model can still provide useful insights about the joint role of risk factors that are putative contributors to underlying biological interactions. The specific biological interactions, however, may only be understood by conducting additional molecular experiments using human biospecimens or animal models [18, 19].

The term “epistasis”, defined as a masking effect that occurs when one genetic factor prevents another from manifesting its effect [19], is also used to refer to different types of interactions [20]: physical interaction between genes (functional epistasis), masking effect (compositional epistasis), and departure from additivity i.e., statistical interaction (statistical epistasis). Functional epistasis can occur even in the absence of a statistical interaction and may be best understood using laboratory experiments. It has been argued recently that compositional epistasis is a more biological notion than statistical epistasis, in the sense that, under compositional epistasis, the disease risk of some individuals in the population is associated with the simultaneous presence of all the risk factors under consideration [21]. This notion has been used to derive lower bounds for the magnitudes of the interaction parameters in a statistical model, which are used as a non-standard test for compositional epistasis [22].

Statistical interactions (henceforth, simply referred to as “interactions”) correspond to higher order effects. Including interaction terms in a model is equivalent to including higher degree polynomial terms, and the resulting model is non-additive in the risk factors. Large interactions may induce curvature effects [23]. On a selected scale of observation, interaction terms may be required to obtain the best fitting model that can explain the curvature effects. Certain types of interactions may be eliminated by a transformation of the outcome so that the resulting model is additive in the risk factors [24, 25]. Such interactions are referred to as removable interactions. When the disease trait (or outcome) is binary, the transformation corresponds to specifying a link function that is additive in the risk factors. The transformation must be invertible so that the results can be back-transformed for clinical interpretation on the original scale, but the removed interactions will reappear in the model on the original scale [12]. When a removable interaction exists, it is pragmatic to eliminate it, whenever it is possible do so, and fit an additive model on the transformed scale to attain model parsimony and a more precise fit to the data [12].

The ultimate objective of this paper is efficient estimation of disease odds (say, on the natural logarithm scale) using a parsimonious additive model, once we have decided that two factors (for example, single nucleotide polymorphisms and/or environmental exposures) are involved in disease risk. Our objective here is not in detecting interaction once main

effects of two risk factors are known to exist since, as noted above, the existence of main effects automatically implies that there is a biological interaction (see also [11, 12]). We focus on case-control studies where the outcome is binary. We first develop a one degree-of-freedom score test for the null hypothesis that there is no interaction on the logistic scale or that it is not removable. The alternative hypothesis is that there is a removable interaction. Once the existence of a removable interaction on the logistic scale is established, we use the Guerrero and Johnson family of transformations [26] for fitting an additive model to the data and estimate the log disease odds. Our focus here is thus on removable interaction, because if there is no removable interaction there is no point in seeking a change of scale to obtain additivity. In the Discussion, we indicate a test for non-removable interaction; unless this is non-significant, we cannot expect an additive model to fit the data well.

This paper is organized as follows. In Section 2 we describe a novel test statistic for removable interactions and describe its asymptotic distribution. In Section 3 we describe different types of interactions. In Section 4 we use simulations to evaluate the type I error and power of the proposed test statistic under the various types of interactions. In Section 5 we illustrate the proposed test using published data from case-control studies of advanced colorectal adenoma [27], bladder cancer [28], and endometrial cancer [29]. In Section 6 we describe the Guerrero and Johnson [26] transformation. In Section 7, we conduct simulation studies to examine the estimation and model fitting properties of this transformation under various types of interaction. In Section 8 we apply this transformation to the case-control studies described in Section 5 to fit additive models, and examine the properties of these models. We conclude the paper with a discussion in Section 9, including a test to verify that after transformation no significant interaction remains.

2. Methods

2.1 Preliminaries

Consider a case-control study where the cases are affected with a disease of interest and the controls do not have that disease. Two risk factors X and Z having L_1 and L_2 levels, respectively, are measured on each individual. Let N_{ij} and M_{ij} denote the number of cases and controls, respectively, having the i -th level of X and the j -th level of Z ($i = 1, 2, \dots, L_1$; $j = 1, 2, \dots, L_2$). Given the total number of individuals $N_{ij} + M_{ij}$ in the (i,j) -th risk factor sub-class, N_{ij} is distributed as a binomial random variable with disease probability p_{ij} given by:

$$P(N_{ij}|N_{ij}+M_{ij}) = \binom{N_{ij}+M_{ij}}{N_{ij}} \times p_{ij}^{N_{ij}} \times (1-p_{ij})^{M_{ij}} \quad (\text{Equation 1})$$

It is assumed that the disease probability p_{ij} depends upon X and Z , and is modeled as follows. We assume that there are no empty sub-classes i.e., $N_{ij} > 0$ and $M_{ij} > 0$ for all i, j . Let $g_0\{\cdot\}$ denote the logit link function i.e., given a real number $x \in (0,1)$,

$g_0(x) = \text{logit}(x) = \log\left\{\frac{x}{1-x}\right\}$. Thus $g_0(p_{ij})$ is the log disease odds in the (i,j) -th risk factor sub-class. We model p_{ij} as:

$$g_0(p_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij}, \quad (\text{Equation 2})$$

where μ is the baseline risk, α_i is the main effect of the i -th level of X ($i = 1, \dots, L_1$), β_j is the main effect of the j -th level of Z ($j = 1, \dots, L_2$), and γ_{ij} is the effect of the interaction

between the i -th level of X and the j -th level of Z , subject to the constraints $\sum_{i=1}^{L_1} \alpha_i = 0$,

$\sum_{j=1}^{L_2} \beta_j = 0$, $\sum_{i=1}^{L_1} \gamma_{ij} = 0$ for $j = 1, \dots, L_2$, and $\sum_{j=1}^{L_2} \gamma_{ij} = 0$ for $i = 1, \dots, L_1$. In vector notation we write $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_{L_1})$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_{L_2})$, $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \dots, \gamma_{L_1 L_2})$.

To estimate the logarithm of disease odds, we can estimate the parameters μ , $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ and plug these into the right hand side of Equation (2). To obtain the parameter estimates, we can define the design matrix, denoted $\mathbf{W}_1$, for the above model (Equation 2) by specifying orthogonal contrasts corresponding to the main and interaction effects. The unknown parameters μ , $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ can be estimated using the iteratively reweighted least squares (IRLS) approach for a generalized linear model [30]. An asymptotically equivalent approach is the weighted least squares (WLS) method for logistic regression analysis of contingency tables. The WLS method considers outcomes ξ_{ij} that are estimated as

$\hat{\xi}_{ij} = \log\left(\frac{N_{ij}}{M_{ij}}\right)$ for the (i,j)-th risk factor sub-class, where “log” is the natural logarithm. In vector notation, we write $\hat{\boldsymbol{\xi}} = (\hat{\xi}_{11}, \hat{\xi}_{12}, \dots, \hat{\xi}_{L_1 L_2})$. The WLS estimates are given by

$(\mathbf{W}_1^T \mathbf{A}_1^{-1} \mathbf{W}_1)^{-1} \mathbf{W}_1^T \mathbf{A}_1^{-1} \hat{\boldsymbol{\xi}}$, where $\mathbf{A}_1 = \text{diag}\left\{\frac{1}{N_{ij}} + \frac{1}{M_{ij}}\right\}$ denotes the asymptotic variance of $\hat{\xi}_{ij}$ [31, 32] and the notation $\text{diag}\{s_{ij}\}$ denotes a diagonal matrix with diagonal entries equal to s_{ij} , $i = 1, 2, \dots, L_1$ and $j = 1, 2, \dots, L_2$. Both the IRLS and WLS estimates can be obtained using standard statistical software packages.

The definition of an interaction depends upon the definition of what constitutes a main effect, and the interaction terms must be interpreted along with the corresponding main effect terms [12, 33]. Therefore, when a term γ_{ij} is included in the model, the corresponding α_i and β_j terms should also be included in the model. When there is no interaction i.e., when $\gamma_{ij} = 0$ for all $i = 1, \dots, L_1$ and $j = 1, \dots, L_2$, the right hand side of Equation (2) becomes $\mu + \alpha_i + \beta_j$. This is a parsimonious additive model that requires estimation of fewer parameters to estimate the log disease odds relative to a model that includes interaction terms. We can test the null hypothesis $H_0: \gamma_{ij} = 0$ for all $i = 1, \dots, L_1$ and $j = 1, \dots, L_2$ against the alternative $H_A: \gamma_{ij} \neq 0$ for some (i, j) using a likelihood ratio statistic to determine whether a parsimonious model may be applicable to the data. Here H_A is a general family of alternatives. The likelihood ratio test statistic, denoted Λ_G has an asymptotic central chi-squared distribution with $(L_1 - 1) \times (L_2 - 1)$ degrees of freedom under H_0 when all the required regularity conditions hold [30]. The subscript “G” is used in Λ_G to indicate that that this test statistic corresponds to the general family of alternatives.

2.2 Removable interaction

Equation (2) states that, under the known link function $g_0\{\cdot\}$, disease risk is a non-additive function of the risk factors. Suppose there is an alternative link function $f\{\cdot\}$ under which disease risk is additive in the risk factors:

$$f\{p_{ij}\} = \mu_0 + \alpha_{i0} + \beta_{j0}, \quad (\text{Equation 3})$$

where μ_0 , the α_{i0} , and the β_{j0} are unknown parameters. If $f\{\cdot\}$ is known, then we may fit Equation (3) to attain model parsimony and possibly a more precise fit. However, $f\{\cdot\}$ is generally unknown in practice. Hence, we may fit a model using $g_0\{\cdot\}$. If $g_0\{\cdot\}$ is linear in $f\{\cdot\}$, then $g_0\{p_{ij}\}$ is additive in the risk factors, and we may replace $f\{\cdot\}$ with $g_0\{\cdot\}$ on the left hand side of Equation (3). If $g_0\{\cdot\}$ is a non-linear function of $f\{\cdot\}$ then a quadratic approximation may provide a better fit than an additive model when the data are fitted using the link function $g_0\{\cdot\}$ [24]:

$$g_0\{p_{ij}\} \approx \eta_0 + \eta_1 f\{p_{ij}\} + \eta_2 [f\{p_{ij}\}]^2, \quad (\text{Equation 4})$$

where η_0 , η_1 , and η_2 are unknown parameters. Suppose Equation (4) indeed provides a better fit than an additive model. Further, suppose the right hand side is a strictly monotonic function of $f\{p_{ij}\}$ in the range $(\min[f\{p_{ij}\}], \max[f\{p_{ij}\}])$, where the minimum and maximum are taken over the values of $f\{p_{ij}\}$ estimated for each risk factor sub-class. Then there exists a transformation of the outcome such that the model is additive on the transformed scale [24].

Assuming that the monotonicity condition holds, we may further approximate Equation (4) as [24]:

$$g_0\{p_{ij}\} \approx \mu + \alpha_i + \beta_j + \theta \times \alpha_i \times \beta_j, \quad (\text{Equation 5})$$

The last term on the right hand side of Equation (5) is a quadratic term taking the form $\gamma_{ij} = \theta \times \alpha_i \times \beta_j$ that characterizes the interaction effects. The parameter θ is a scalar that quantifies non-additivity. When there are main effects (i.e., $\alpha_i \neq 0$ and $\beta_j \neq 0$ for at least one i and j), there is no interaction (i.e., all estimable $\gamma_{ij} = 0$) if and only if $\theta = 0$, and Equation (5) becomes an additive model. When $\theta \neq 0$, the model is non-additive. In this case, since monotonicity holds, the model is intrinsically linear, and we may empirically look for a transformation of the outcome to identify a link function $f\{\cdot\}$ and then fit an additive model on the transformed scale [33].

Monotonicity induces a useful class of alternatives for evaluating interactions – namely, removable interactions. Under the assumption of monotonicity, we can test the null hypothesis of no interaction against the alternative that an interaction exists and is removable, instead of considering the general family of alternatives; i.e., we can test $\gamma_{ij} = 0$ for all i and j against the alternative $\gamma_{ij} \neq 0$ for some (i, j) and that it takes the form $\gamma_{ij} = \theta \times \alpha_i \times \beta_j$. When we assume that the main effects are not all 0, this is equivalent to testing $H_0: \theta = 0$ against the alternative $H_A: \theta \neq 0$. When H_0 is rejected, this would be an indication that there is removable interaction, suggesting that it may be sufficient to fit an additive model to the data under an alternative link function in order to estimate the log disease odds in the various risk factor sub-classes. However, when H_0 is not rejected, this would mean that there is no removable interaction: either there is no interaction at all or there is only non-removable interaction (possibly because the monotonicity assumption fails). In this paper, we focus on the evaluation of removable interaction, but this may need to be complemented by a test for significant non-removable interaction (see Discussion). When there is evidence for a removable interaction, it is pragmatic to eliminate it via a transformation of the outcome to attain model parsimony and a precise fit, especially if all the interactions are removable [12]. In the following subsections, we develop a score test for removable interaction.

2.2.1 Score statistic for evaluating removable interaction—The score function for θ , evaluated at H_0 , is denoted U . It can be obtained using Equation (5) (see Appendix A). The estimate of U , denoted $\hat{U}$, is given by:

$$\hat{U} = \mathbf{1}^T \widehat{\mathbf{D}} \mathbf{A} \hat{\boldsymbol{\gamma}} = \sum_{i=1}^{L_1} \sum_{j=1}^{L_2} \hat{\alpha}_i \times \hat{\beta}_j \times \left(\frac{1}{N_{ij}} + \frac{1}{M_{ij}} \right)^{-1} \times \hat{\gamma}_{ij}, \quad (\text{Equation 6})$$

where

- $\mathbf{1}$ is a column vector of unities of length $L_1 \times L_2$;
- $\hat{\mathbf{D}} = \text{diag}\{\hat{\alpha}_i \times \hat{\beta}_j\}$ is a diagonal matrix of dimension $L_1 \times L_2$, where $\hat{\alpha}_i$ ($i = 1, 2, \dots, L_1$) and $\hat{\beta}_j$ ($j = 1, 2, \dots, L_2$) are the estimated main effects of X and Z obtained from an additive model (by setting $\theta = 0$ in Equation 5);
- $\hat{\boldsymbol{\gamma}}$ are the estimated residuals of the additive model, and provide an estimate of the interaction effects $\boldsymbol{\gamma}$;
- $\mathbf{A}$ is a diagonal matrix of dimension $L_1 \times L_2$, where the diagonal elements represent the inverse of the variance of $g_0\{p_{ij}\}$. When the sample size is large, $\mathbf{A}$ is equivalent to $\mathbf{A}_1^{-1}$, where $\mathbf{A}_1$ is given in Section 2.1.

2.2.2 Estimating the main effects and residuals from an additive model—Data from a case-control study are unbalanced since all the $N_{ij} + M_{ij}$ are not the same. Therefore, the main effects must be defined appropriately to obtain orthogonal estimates of the residuals, as described below.

The elements of $\hat{\boldsymbol{\gamma}}$ are of the form $\hat{\gamma}_{ij} = \hat{\xi}_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$, where $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$ are estimates of the baseline risk and the main effects of the i -th level of X and j -th level of Z, respectively, and $\hat{\xi}_{ij}$ are defined in Section 2.1. Denoting v_{ij} as weights for the (i,j) -th subclass, these estimates are, in general, given by Scheffé [14] as:

$$\hat{\mu} = \frac{\sum_{i=1}^{L_1} \sum_{j=1}^{L_2} v_{ij} \hat{\xi}_{ij}}{\sum_{i=1}^{L_1} \sum_{j=1}^{L_2} v_{ij}}, \hat{\alpha}_i = \frac{\sum_{j=1}^{L_2} v_{ij} \hat{\xi}_{ij}}{\sum_{j=1}^{L_2} v_{ij}} - \hat{\mu}, \hat{\beta}_j = \frac{\sum_{i=1}^{L_1} v_{ij} \hat{\xi}_{ij}}{\sum_{i=1}^{L_1} v_{ij}} - \hat{\mu}. \quad (\text{Equation 7})$$

The standard approach for estimating the parameters of the additive model weights all the study subjects equally. The resulting estimates are the WLS estimates given by $(\mathbf{W}^T \mathbf{A} \mathbf{W})^{-1} \mathbf{W}^T \mathbf{A} \hat{\boldsymbol{\xi}}$, where $\mathbf{W}$ is the design matrix of the additive model with $L_1 \times L_2$ rows and $L_1 + L_2 - 1$ columns. The first column of $\mathbf{W}$ is a vector of unities corresponding to the baseline risk, and the next $L_1 - 1$ and $L_2 - 1$ columns are orthogonal contrasts corresponding to the main effects of X and Z, respectively. When the parameter estimates of Equation (7) are obtained by weighting all the study subjects equally, we have $E(\hat{\gamma}_{ij}) = 0$ even when there is no interaction, i.e., even when $\gamma_{ij} = 0$ for all i and j [11]. [To see this, note that $E(\hat{\xi}_{ij}) = \mu + \alpha_i + \beta_j$ when there is no interaction.] This suggests that $\hat{\gamma}_{ij}$ does not provide an appropriate estimate of the error term under the null hypothesis of no interaction. This issue can be resolved when the weights v_{ij} are chosen to be proportional (or when they occur in proportional frequencies) i.e., when $v_{ij} = v_i \times v_j$, where v_i does not depend upon j and v_j does not depend upon i . A special case of proportional frequencies is the case of equal weights i.e., $v_{ij} = v$ for all i and j , where v is an arbitrary non-zero real number. When the

weights are proportional, it follows from Equation (7) that $\sum_i v_i \hat{\alpha}_i = 0$ and $\sum_j v_j \hat{\beta}_j = 0$. The following results apply for proportional (and, hence, for equal) weights (see Appendix B):

Result 1: When the weights are proportional, we have $E(\hat{\gamma}_{ij}) = 0$ in the absence of an interaction, regardless of the choice of v_i and v_j .

Result 2: The expected value of any contrast in $\hat{\alpha}_i$ (or $\hat{\beta}_j$) is independent of the weights when there is no interaction.

It then follows that the spaces of dimension $L_1 - 1$, $L_2 - 1$, and 1, spanned by $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\mu}_j$, respectively, are mutually orthogonal. These results suggest that, without loss of generality, we can obtain the parameter estimates of the additive model using equal weights by considering $v_i = 1$ for all $i = 1, \dots, L_1$ and $v_j = 1$ for all $j = 1, \dots, L_2$. The parameter estimates can be obtained as $(\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T \hat{\xi}$. The estimated residuals are, therefore, given by $\hat{\gamma} = (\mathbf{I} - \mathbf{H}) \hat{\xi}$, where $\mathbf{H} = \mathbf{W}(\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T$ is a symmetric and idempotent matrix.

2.2.3 Test statistic—We can calculate $\hat{U}$ (Equation 6) using the parameter estimates obtained as described above. We propose the following statistic, denoted Λ_R , for testing $H_0: \theta = 0$ against $H_A: \theta \neq 0$:

$$\Lambda_R = \frac{\hat{U}^2}{\text{Var}(\hat{U}|\hat{\alpha}, \hat{\beta})}, \quad (\text{Equation 8})$$

where $\text{Var}(\hat{U}|\hat{\alpha}, \hat{\beta}) = \mathbf{1}^T \hat{\mathbf{D}} \mathbf{A} (\mathbf{I} - \mathbf{H}) \mathbf{A}^{-1} (\mathbf{I} - \mathbf{H}) \mathbf{A} \hat{\mathbf{D}} \mathbf{1}$ is the conditional variance of $\hat{U}$. The subscript “R” is used in Λ_R to emphasize that the alternative hypothesis represents a family of removable interactions. This statistic has parallels to Tukey’s one-degree of freedom test for interaction in a balanced two-way analysis of variance with one observation per cell [34]. Here Λ_R is a test statistic for evaluating removable interactions in unbalanced data.

2.2.4 Asymptotic distribution of Λ_R

Distribution under the alternative hypothesis: Since the main effects and interactions are defined using orthogonal contrasts, $\hat{\gamma}$ is independent of $\hat{\alpha}$ and $\hat{\beta}$. Suppose the null hypothesis of no interaction does not hold. Under the alternative family of removable interactions;

given $\hat{\alpha}$ and $\hat{\beta}$, $\hat{U}$ has conditional mean $E(\hat{U}|\hat{\alpha}, \hat{\beta}) = \theta \times \sum_{i=1}^{L_1} \sum_{j=1}^{L_2} \hat{\alpha}_i \times \hat{\beta}_j \times \left(\frac{1}{N_{ij}} + \frac{1}{M_{ij}} \right)^{-1} \times \alpha_i \times \beta_j$ (see Appendix C). The conditional variance of $\hat{U}$ is $\text{Var}(\hat{U}|\hat{\alpha}, \hat{\beta})$ given above. Given $\hat{\alpha}$ and $\hat{\beta}$, each component of $\hat{\gamma}$ has an asymptotic normal distribution with mean $\gamma_{ij} = \theta \times \alpha_i \times \beta_j$ under the alternative hypothesis of a removable interaction. Thus, $\hat{U}$ (Equation 6) is a weighted sum of asymptotically normally distributed random variables, where the weights $\hat{\alpha}_i$ and $\hat{\beta}_j$ (and N_{ij} and M_{ij}) are known. Hence, $\hat{U}$ has an asymptotic normal distribution with mean and variance given above. Therefore, the conditional distribution of Λ_R , given $\hat{\alpha}$ and $\hat{\beta}$, is chi-squared with 1 degree of freedom and non-centrality parameter

$\Delta(\hat{\alpha}, \hat{\beta}) = \frac{\{E(\hat{U}|\hat{\alpha}, \hat{\beta})\}^2}{\text{Var}(\hat{U}|\hat{\alpha}, \hat{\beta})}$, which depends on the estimated main effects. Obtaining the unconditional distribution (and, hence, the power of Λ_R) involves obtaining expectations of the form $E\{\Delta^t(\hat{\alpha}, \hat{\beta})\}$ for various values of t . Since the numerator and denominator of Δ

involve terms of the form $\left(\frac{1}{N_{ij}} + \frac{1}{M_{ij}} \right)^{-1}$, evaluating these expectations, and hence obtaining a closed form for Δ , is not straightforward. We use simulations to evaluate the power of Λ_R under different parametric configurations in Section 4.

Null distribution: To obtain the asymptotic null distribution of Λ_R , we first assume that main effects exist i.e., $\alpha_i \neq 0$ and $\beta_j \neq 0$ for at least one (i, j) . Under $H_0: \theta = 0$, the non-centrality parameter is 0. Therefore, the conditional null distribution of Λ_R , given $\hat{\alpha}$ and $\hat{\beta}$, is a central chi-squared distribution with 1 degree of freedom, which is also its unconditional null distribution. Suppose at least one of the risk factors under consideration does not have an effect on the outcome i.e., the true model is $g_0(p_{ij}) = \mu$ (when neither X nor Z has an effect), or $\mu + \alpha_i$ (when Z has no effect), or $\mu + \beta_j$ (when X has no effect). Then the

interaction effect is not estimable. Suppose, however, we fit the additive model $g_0(p_{ij}) = \mu + \alpha_i + \beta_j$, and test for the presence of a removable interaction. In Appendix C we show that, in this case, the asymptotic null distribution of Λ_R has a non-standard form. This non-standard distribution has a slightly heavier tail than a central chi-squared distribution with 1 degree of freedom, resulting in a small inflation in the type I error – particularly when neither risk factor has a main effect and the two risk factors are independent. We observe this phenomenon in our simulations. We also observe that, although the asymptotic null distribution of Λ_R departs from a central chi-squared distribution with 1 degree of freedom, the departure is not severe in our case, suggesting that we may use this distribution as a benchmark for assessing the statistical significance of Λ_R in practical settings.

3. Types of interactions

To evaluate the type I error and power of Λ_R , we simulated risk factors X and Z for cases and controls. We assumed thresholds C_1 and C_2 for the levels of X and Z , respectively, for disease risk through the following model:

$$\text{Logit} \{P(\text{disease} | X, Z)\} = \mu + \alpha_1 \times I(X \geq C_1) + \beta_1 \times I(Z \geq C_2) + \gamma_{11} \times I(X \geq C_1) \times I(Z \geq C_2), \quad (\text{Equation } 9)$$

where $I(\cdot)$ denotes the indicator function. Before proceeding further, it will be useful to understand the different types of interactions that can result from this model. In this model, X and Z contribute to disease risk only when $X \geq C_1$ and $Z \geq C_2$. Therefore, the disease risk factors are intrinsically binary. Denote $I_X = I(X \geq C_1)$ and $I_Z = I(Z \geq C_2)$. The outcome summaries are $\xi_{ij} = \text{logit} \{P(\text{disease} | X, Z)\} = \text{logit} \{P(\text{disease} | I_X=i, I_Z=j)\}$, where $(i,j) = (0,0), (0,1), (1,0), \text{ or } (1,1)$. The ordering of the ξ_{ij} induces different types of interactions [34]. Without loss of generality, we assume that the sub-class $(I_X, I_Z) = (0,0)$ has the smallest disease risk. There are 6 possible orderings of the ξ_{ij} , providing 4 major types of interactions, denoted Type A, Type B, Type C, and Type D ([35]; see Table 1 of Haldane's paper or Figure 1 of this paper); Types A and D each contain two subtypes. These interactions can be interpreted as given below. Although we consider these four types of interactions here, additional types may be obtained by allowing some of the inequalities to be equalities (for example, [36]).

Type A interaction

The outcome summary under condition $I_X=1$ is as large or larger than that under $I_X=0$ for both levels of I_Z , and similarly for the summary under the two conditions for I_Z i.e., $\xi_{00} \leq \xi_{10}, \xi_{01} \leq \xi_{11}, \xi_{00} \leq \xi_{01}, \xi_{10} \leq \xi_{11}$. This is equivalent to: (i) $\mu + \alpha_1 \geq \mu$; (ii) $\mu + \beta_1 \geq \mu$; (iii) $\mu + \alpha_1 + \beta_1 + \gamma_{11} \geq \mu + \beta_1$; (iv) $\mu + \beta_1 + \gamma_{11} \geq \mu + \alpha_1 + \beta_1 + \gamma_{11}$. Therefore, $\alpha_1 \geq 0, \beta_1 \geq 0$, and $\gamma_{11} \geq \max\{-\alpha_1, -\beta_1\}$.

Type B interaction

The condition $I_X=1$ has outcome summary as large as or larger than that under $I_X=0$ when $I_Z=0$. The ordering is reversed when $I_Z=1$. In contrast, the condition $I_Z=1$ has summary as large as or larger than that under $I_Z=0$ for both levels of I_X , i.e., $\xi_{00} \leq \xi_{10}, \xi_{01} \leq \xi_{11}, \xi_{00} \leq \xi_{01}, \xi_{10} \leq \xi_{11}$. Therefore, $\alpha_1 \geq 0, \beta_1 \geq 0$, and $-\beta_1 \geq \gamma_{11} \geq -\alpha_1$ when $\beta_1 \geq \alpha_1$.

Type C interaction

The condition $I_X=1$ has outcome summary as large as or larger than that under $I_X=0$ for both levels of I_Z . However, the condition $I_Z=1$ has summary as large as or larger than that under $I_Z=0$ when $I_X=0$, and the ordering is reversed when $I_X=1$, i.e., $\xi_{00} \leq \xi_{10}, \xi_{01} \leq \xi_{11}, \xi_{00} \leq \xi_{01}, \xi_{10} \leq \xi_{11}$. Therefore, $\alpha_1 \geq 0, \beta_1 \geq 0$, and $-\alpha_1 \geq \gamma_{11} \geq -\beta_1$ when $\alpha_1 \geq \beta_1$. Type C interaction is similar to Type B interaction, with the roles of I_X and I_Z reversed.

Type D interaction

This interaction does not preserve any specific ordering of the outcome summaries, but rather defines a cross-over effect i.e., $\xi_{00} \leq \xi_{10}, \xi_{01} \leq \xi_{11}, \xi_{00} \leq \xi_{01}, \xi_{10} \leq \xi_{11}$. This leads to the following conditions: $\alpha_1 \leq 0, \beta_1 \leq 0$, and $\gamma_{11} \leq \min\{-\alpha_1, -\beta_1\}$.

Chatterjee et al [37] developed a one-degree of freedom test for detecting a causal allele when there is an interaction between two alleles, and demonstrated the properties of their test statistic for a binary trait and two binary risk factors, using the odds ratio as outcome. Of the interaction models considered by these authors, the multiplicative model is equivalent to a Type A interaction, and all their other interaction models can be written as Type A, B, or C, depending upon the odds ratios assumed for the four cells. Their “cross-over” effect gives the smallest odds ratio (<1) to the cell $(I_X=0, I_Z=0)$, the cells $(I_X=0, I_Z=1)$ and $(I_X=1, I_Z=0)$ have odds ratios equal to 1, and the cell $(I_X=1, I_Z=1)$ has an odds ratio greater than 1. Thus, the “cross-over” effect of Chatterjee et al [37] is not a cross-over effect in the sense of a Type D interaction, but can be written as an interaction of Type A, B, or C, depending upon the magnitudes of the cell values.

Cordell [19] and VanderWeele [21, 22] considered compositional epistasis that is quantified as follows: the presence of at most one risk factor does not result in any change in disease risk, while the simultaneous presence of both the risk factors changes (for example, increases) risk. In our notation, assuming $\gamma_{11} > 0$, the compositional epistasis considered by these authors can be quantified as: $\xi_{00} = \xi_{10} = \xi_{01} = \mu$, and $\xi_{11} = \mu + \gamma_{11} > \mu$. This comes under the rubric of Haldane’s Type A interaction since $\xi_{00} \leq \xi_{01}, \xi_{10} \leq \xi_{11}, \xi_{00} \leq \xi_{10}$ and $\xi_{01} \leq \xi_{11}$.

In Appendix D we use the example of two independent binary risk factors to show that Type A, but not Type D interaction, generally satisfies the monotonicity condition. Indeed, Type A would satisfy strict monotonicity when $\alpha_1 > 0, \beta_1 > 0$, and $\gamma_{11} > \max\{-\alpha_1, -\beta_1\}$. We also show that certain kinds of Type B and Type C interactions satisfy the monotonicity condition. Therefore, the test under removable alternatives, Λ_R , is likely to be a powerful approach for identifying interactions of Types A, and certain kinds of interactions of Types B and C, but not interactions of Type D. The test under the general alternatives, Λ_G , is likely to be beneficial mainly for identifying interactions of Type D.

For simplicity and for ease of explanation, we have described the four types of interactions using 2×2 contingency tables. For this setting, since Type A interaction generally obeys the monotonicity condition and is, hence, removable, compositional epistasis may also be largely removable. However, for larger contingency tables, an interaction term may be needed to obtain a better fit to the data even after a transformation in the presence of compositional epistasis.

4. Simulation

We first conducted a simulation study to establish that Λ_R has adequate power to detect a removable interaction.

4.1 Set up

We simulated data for 1000 cases and 1000 controls by considering two risk factors – X having $L_1 = 3$ levels with marginal frequencies $P(X = 1) = 0.70, P(X = 2) = 0.20$, and $P(X = 3) = 0.10$, and Z having $L_2 = 2$ levels with marginal frequencies $P(Z = 1) = 0.9 = 1 - P(Z = 0)$. For each person, we generated X and Z using these marginal probabilities and setting the correlation between I_X and I_Z to be $\rho \in \{0, 0.50, 0.80\}$, where $\rho = 0$ implies that the risk factors are independent. The baseline parameter μ was chosen by setting the marginal

disease probability to be 10%. Given X and Z , the binary disease status for each individual was generated using the model given by Equation (9) with $C_1 = 3$ and $C_2 = 2$. To calculate type I error, we set $\gamma_{11} = 0$ and $\{\alpha_1, \beta_1\} \in \{0, \log(2)=0.6931, \log(3)=1.0986\}$. There are no main effects when $\alpha_1 = \beta_1 = 0$. In this case we anticipate Λ_R to have a non-standard null distribution, which would impact the type I error. To calculate power, we set $\alpha_1 = \beta_1 = \log(2) = 0.6931$, and set values of γ_{11} in the relevant ranges to represent Type A and D interactions (see Section 3). Note that when $\alpha_1 = \beta_1$, interactions of Types B and C are essentially subsumed under Type A interaction. Therefore, we present results for power under Type A and D interactions. We also generated data under compositional epistasis by setting $\alpha_1 = \beta_1 = 0$ and $\gamma_{11} \neq 0$.

A total of 1000 data sets were generated under each parametric configuration. Each data set was analyzed using logistic regression for contingency tables by defining orthogonal contrasts for the main and interaction effects. We tested for removable interaction using Λ_R , and evaluated its significance at the 5% level by using the central chi-squared distribution with one degree of freedom as the null distribution. As a benchmark, we also tested the null hypothesis of no interaction against the family of general alternatives using the test statistic Λ_G , and evaluated its significance using a central chi-squared distribution with $(L_1-1) \times (L_2-1)$ degrees of freedom as the null distribution. Under each parametric configuration, the empirical type I error and power of each test statistic are defined as the fraction of data sets out of 1000 where that test statistic is significant.

4.2 Results

Table 1 shows the empirical type I error rates at the nominal significance level of 0.05. As expected, Λ_G had a nominal type I error rate. When $\alpha_1 = 0$ and $\beta_1 = 0$, Λ_R had type I error close to the nominal rate under all the parametric configurations considered. Any departure of these type I errors from the nominal rate of 5% may be attributed to random variation. This can be seen in Figure 2, which shows the quantile-quantile plots of the empirical distribution of Λ_R versus the quantiles of a central chi-squared distribution with 1 degree of freedom. Figure 2 shows that when $\alpha_1 = 0$, $\beta_1 = 0$, and $\rho = 0$, the values of Λ_R are slightly larger than expected, resulting in an empirical type I error of 0.10. Otherwise, in general, the empirical null distribution of Λ_R is remarkably close to a central chi-squared distribution with 1 degree of freedom.

Table 1 also shows the intriguing result that when the risk factors are correlated ($\rho = 0.50$ and 0.80) and there are no main effects ($\alpha_1 = 0 = \beta_1$), the type I error of Λ_R is close to the nominal level (i.e., similar to a setting where the main effects are non-zero). On reflection, this phenomenon may be explained by considering parameter estimation under a generalized linear regression model with the logit link function. For this explanation, let $\mathbf{X}$ and $\mathbf{Z}$ denote matrices of dimension $N \times (L_1-1)$ and $N \times (L_2-1)$, respectively, corresponding to the two risk factors, whose columns indicate the levels of the risk factors in each individual. Here N denotes the total number of cases and controls. Further suppose that $\mathbf{X}$ and $\mathbf{Z}$ are of full rank, where $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{Z})$. Let $\mathbf{X} \otimes \mathbf{Z}$ denote a matrix of dimension $N \times (L_1-1)(L_2-1)$ whose columns are the product of every column of $\mathbf{X}$ with every column of $\mathbf{Z}$. Then, the main effects of the risk factors may be estimated under the Fisher scoring approach, using the working likelihood for the model $\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \mathbf{X} \otimes \mathbf{Z} \boldsymbol{\gamma} + \text{error}$. In this notation, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are parameter vectors of lengths L_1-1 , L_2-1 , and $(L_1-1)(L_2-1)$, respectively, quantifying the main effects and the interaction effects, and $\mathbf{y}$ is the vector of working responses of length N . Suppose the true value of $\boldsymbol{\beta}$ is 0. Denote $\mathbf{H}_1 = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. Note that $\mathbf{H}_1$ is a symmetric and idempotent matrix and $(\mathbf{I} - \mathbf{H}_1)\mathbf{X} = 0$, where $\mathbf{I}$ is the identity matrix of dimension equal to the total number of cases and controls. Then $\boldsymbol{\beta}$ may be estimated from the model $(\mathbf{I} - \mathbf{H}_1)\mathbf{y} = (\mathbf{I} - \mathbf{H}_1)\mathbf{Z}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H}_1)\mathbf{X} \otimes \mathbf{Z} \boldsymbol{\gamma} + (\mathbf{I} - \mathbf{H}_1)\text{error}$. When $\rho = 0$, the columns

of $\mathbf{X}$ and $\mathbf{Z}$ are orthogonal. Hence, we have $(\mathbf{I} - \mathbf{H}_1)\mathbf{Z} = \mathbf{Z}$. However, when $\rho \neq 0$, $\mathbf{X}$ and $\mathbf{Z}$ span the same vector space. Hence, we have $(\mathbf{I} - \mathbf{H}_1)\mathbf{Z} \approx 0$, regardless of the value of β . Consequently, we may consider some arbitrary non-zero value for β even when its true value is 0, so that even when $\mathbf{X}$ and $\mathbf{Z}$ are correlated, the type I error of L_R is similar to when the main effects are non-zero.

Figure 3 shows the empirical power of Λ_R and Λ_G under Type A and Type D interactions and under compositional epistasis. As expected, Λ_R has remarkably good power to detect both Type A interaction and compositional epistasis as being removable. Further, as expected, Λ_R is much less powerful than Λ_G when the interaction is of Type D.

5. Three illustrative examples

Here we illustrate the proposed test using published data from three case-control studies.

5.1 Advanced colorectal adenoma, smoking, and NAT2 acetylation

Table 2 gives the number of cases with advanced colorectal adenoma and number of disease-free controls, their smoking status and NAT2 acetylation from a published case-control study reported by Moslehi et al [27]. Smoking has $L_1 = 3$ levels (never, past, current smoker) and NAT2 acetylation is considered to have $L_2 = 2$ levels (rapid/intermediate and slow acetylation). The observed log odds values of the 6 risk factor sub-classes are also shown in Table 2 and graphically represented in Figure 4 using bold and dashed black lines. For both never and current smokers, the log odds of NAT2 rapid/intermediate acetylators is smaller than that of the slow acetylators. Although there is a reversal of this trend among past smokers, the difference between the log odds of NAT2 slow versus rapid/intermediate acetylators among past smokers appears to be small. Ignoring this reversal of trend, this figure suggests that an interaction, if it exists, may be of Type A, which is removable.

Moslehi et al [27] showed a significant evidence for an interaction between current smoking and NAT2 acetylation in this study. We tested the null hypothesis of no interaction against the general family of alternatives as follows. A generalized linear model with a logit link function (Equation 2) was considered by defining the main and interaction effects contrasts using an orthogonal design matrix, and the parameter estimates were obtained using the IRLS approach. The test statistic for interaction was $\Lambda_G = 8.68$ (d.f = 2, p-value = 0.01), suggesting significant evidence for an interaction under the general family of alternatives.

In order to determine whether this interaction is removable, we calculated the proposed test statistic Λ_R by estimating the main effects using the equally weighted approach described in Section 2.2.2. The value of the test statistic was $\Lambda_R = 6.77$ (d.f = 1, p-value = 0.01), suggesting significant evidence for a removable interaction. Therefore, we may be able transform the outcome to fit an additive model on the transformed scale, which will be pursued in Section 8.

5.2 Bladder cancer, smoking, and NAT2 acetylation

In a published genome-wide association case-control study of bladder cancer, Rothman et al [28] found significant evidence for an interaction between smoking (consisting of three levels as before – never, past and current smokers) and the NAT2 gene (single nucleotide polymorphism rs1495741 consisting of two levels – AA/AG genotypes representing rapid/intermediate acetylators, and GG genotype representing slow acetylators). The case-control data for smoking and NAT2 gene are shown in Table 3. Figure 5 shows the observed log odds in the 6 risk factor sub-classes using bold and black dashed lines. This figure suggests that an interaction, if it exists, may be of Type A, which is removable. We tested for the presence of a removable interaction. The value of the test statistic was $\Lambda_R = 16.82$ (d.f = 1,

p-value < 0.001), which is significant and suggests the possibility of transforming the outcome to fit an additive model.

5.3 Endometrial cancer, tea intake and CYP19A1 genotype

In an analysis of case-control data from the Shanghai Endometrial Cancer Study, Xu et al [29] reported a significant interaction between tea intake (two levels: low and high intake) and *CYP19A1* genotype based on the single nucleotide polymorphism rs1065779 (three levels: GG, GT, and TT genotypes). These data are shown in Table 4. Column 5 of the table suggests that the log odds decrease monotonically as the number of copies of the T allele increases among individuals having high tea intake. In contrast, there is a lack of monotonic trend among individuals having low tea intake, suggesting that an interaction (if any) may be of Type D, which is not removable. Thus, as expected, the test for removable interaction was not significant at the 5% level ($\Lambda_R = 1.02$, d.f = 1, p-value = 0.31). However, the test under the general alternative suggested a significant interaction ($\Lambda_G = 9.10$, d.f = 2, p-value = 0.01). These results suggest that a transformation to additivity is not appropriate for these data.

6. Transformation of the outcome

Several methods are available for transforming binary outcomes [26, 37–39]. In this paper, we consider the Guerrero and Johnson [26] family of power transformations, henceforth abbreviated GJ. The GJ transformation of the outcome for the (i,j)-th risk factor subclass is given in terms of the disease risk, p_{ij} . Denoting λ as the transformation parameter, the transformation, denoted $g(p_{ij}, \lambda)$, is given by:

$$g(p_{ij}, \lambda) = \begin{cases} \frac{1}{\lambda} \left(\left[\frac{p_{ij}}{1-p_{ij}} \right]^\lambda - 1 \right) & \text{when } \lambda \neq 0 \\ \log \left(\frac{p_{ij}}{1-p_{ij}} \right) & \text{when } \lambda = 0 \end{cases} \quad (\text{Equation 10})$$

6.1 Properties of the GJ transformation

The GJ transformation has similarities to the Box-Cox family of power transformations for continuous outcomes [25]. The logit link function is a member of the GJ family since $g(p_{ij}, \lambda)$ approaches the logit link function as λ approaches 0. This transformation is

invertible using the identity $\frac{p_{ij}}{1-p_{ij}} = \{1 + \lambda \times g(p_{ij}, \lambda)\}^{1/\lambda}$ when $\lambda \neq 0$ and

$\frac{p_{ij}}{1-p_{ij}} = \exp\{g(p_{ij}, 0)\}$ when $\lambda = 0$. However, there is no λ for which $g(p_{ij}, \lambda) = p_{ij}$.

Under the logit transformation, the disease probability is given by $\frac{\exp(x)}{1+\exp(x)}$, where x is a real number. As a function of x , the disease probability is an “S”-shaped curve, and is symmetric around $1/2$ (see Figure 6). For a given λ , the disease probability under the GJ

$$\frac{(1+\lambda x)^{1/\lambda}}{1+(1+\lambda x)^{1/\lambda}}$$

transformation is given by $\frac{(1+\lambda x)^{1/\lambda}}{1+(1+\lambda x)^{1/\lambda}}$. This probability is not symmetric, and its shape is governed by the magnitude of λ (see Figure 6). For some values of λ , the probability is finite only when x falls in a certain range. For example, when $\lambda = 3$, the disease probability is defined when $x > -1/3$. When $\lambda \rightarrow 0$, the probability is attenuated (or inflated) relative to the logit transformation for large positive (or negative) values of x (if the probability is

defined for these values of x). This tail behavior may help characterize certain curvature effects exhibited by disease probabilities.

6.2 Parameter estimation under the GJ transformation

The additive model on the transformed scale is given by:

$$g(p_{ij}, \lambda) = \mu + \alpha_i + \beta_j, \quad (\text{Equation 11})$$

where $g(p_{ij}, \lambda)$ is given by Equation (10). The model parameters μ , α_i ($i = 1, \dots, L_1$), and β_j ($j = 1, \dots, L_2$) and the transformation parameter λ are unknown. For a given λ , the maximum likelihood estimates (MLE) of the model parameters can be obtained using an iterative procedure, as described in Appendix E. We can use these MLEs in Equation (11) to obtain the estimated values of $g(p_{ij}, \lambda)$, then plug these into Equation (10) to obtain the estimated value of p_{ij} , and then, using Equation (1), obtain the maximum log-likelihood of the binomial distribution.

The MLE of the transformation parameter λ can be obtained using a profile log-likelihood approach as follows [25]. Consider various values for λ . For each value, obtain the MLEs of the model parameters and the maximum log-likelihood values, and identify the value of λ that maximizes these maximum log-likelihoods. Denote $L_{\max}(\lambda)$ as the maximum log-likelihood of the binomial distribution (Equation 1 with probabilities given by Equations 10 and 11) for a given λ and $\hat{\lambda}$ as the MLE of λ . Following Box and Cox [25], we can obtain a 95% confidence interval of the form:

$$2 \times \{L_{\max}(\hat{\lambda}) - L_{\max}(\lambda)\} < 3.84 \quad (\text{Inequality 12})$$

From Equation (10), the relationship between the logit and GJ transformations is given by:

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \frac{1}{\lambda} \times \log\{1 + \lambda \times g(p_{ij}, \lambda)\}. \quad (\text{Equation 13})$$

Consider the MLEs of the model parameters of Equation (11), calculated at the MLE of λ . The MLEs of the parameters can be plugged into Equation (13) to obtain the MLE of the log odds, and their standard errors can be obtained using the delta method by accounting for the fact that λ is unknown and is estimated from the observed data (see Appendix E)

6.3 Relationship between the GJ link and removable interaction

Using a Taylor's series expansion of the natural logarithm up to the quadratic term, the right hand side of Equation (13) can be written as:

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) \approx g(p_{ij}, \lambda) - \frac{\lambda}{2} \times \{g(p_{ij}, \lambda)\}^2. \quad (\text{Equation 14})$$

When the model is additive under the GJ link and when disease risk is a strictly monotonic function of the additive effects of the risk factors, the right hand side of Equation (14) may be further approximated as follows (see also [24]):

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) \approx \mu + \alpha_i + \beta_j - \frac{\lambda}{2} \times \{\mu + \alpha_i + \beta_j\}^2 \approx \mu + \alpha_i + \beta_j - \frac{\lambda}{2} \times 2 \times \alpha_i \times \beta_j = \mu + \alpha_i + \beta_j - \lambda \times \alpha_i \times \beta_j \quad (\text{Equation 15})$$

Comparing Equations (5) and (15) suggests that the GJ transformation is appropriate for fitting an additive model when the interaction is removable (i.e., of the form given by

Equation 5) and that the transformation parameter is $\lambda \approx -\theta$. In general, when writing the approximation in the second step of Equation (15), the squared terms such as α_i^2 and β_j^2 may be absorbed into the main effects [24, 34]. However, when the interaction is entirely removable, the term $-\lambda/2 \times (\mu + \alpha_i + \beta_j)^2$ is the interaction term γ_{ij} , and the approximations in Equation (15) will be equalities. In particular, using the identifiability conditions given below Equation (2), we can write $\{\mu + \alpha_i + \beta_j\}^2 = 2 \times \alpha_i \times \beta_j$ to obtain the second step (see, for example, Chapter 4 of [14]).

7. Simulation

We conducted simulation studies to investigate the properties of the following three models: (i) the full logistic regression model (i.e., a logistic regression model that includes main effects and interaction terms); (ii) the additive logistic regression model (i.e., a logistic regression model that does not include any interaction term); and the optimal additive GJ model (i.e., an additive model based on the GJ link using the MLE of λ).

We simulated case-control data with 1000 cases and 1000 controls. Two risk factors, X having L_1 levels and Z having L_2 levels, were simulated for each person. We considered simulations with $(L_1, L_2) = (3, 2)$ and $(5, 5)$, which correspond to 2×3 and 5×5 contingency tables. The binary disease trait was simulated using Equation (9) with $(C_1, C_2) = (3, 2)$ and $(4, 4)$ for 2×3 tables and 5×5 tables, respectively. We simulated data under interactions of Types A and D and compositional epistasis, and the simulation parameters were chosen as described in Section 4. When simulating data with 5×5 tables, we considered the prevalence of each level of the risk factors to be 20%. We simulated 1000 data sets under each parametric configuration.

The full logistic model contained $(L_1-1) \times (L_2-1)$ interaction terms. Thus, the total number of interaction terms was 2 and 16 for the 2×3 and the 5×5 contingency tables, respectively. We obtained the MLEs and standard errors of the log odds of the three models, and evaluated the estimation properties of the models using their mean squared error (MSE), Akaike's AIC, and the BIC (Schwarz criterion). The MSE was calculated as the square of the difference between the observed and estimated log odds plus the square of the standard error, averaged over all the risk factor sub-classes. The AIC and BIC were calculated as $-2 \times \text{maximum log-likelihood} + 2P$, and $-2 \times \text{maximum log-likelihood} + P \times \log(N)$, respectively, where P denotes the number of unknown parameters in the model and N is the total sample size. Thus, $P = L_1 \times L_2$ for the full logistic model, $P = L_1 + L_2 - 1$ for the additive logistic model, and $P = L_1 + L_2$ for the optimal additive GJ model (since λ was estimated under this model).

Figures 7 and 8 shows the results, averaged over the 1000 simulated data sets, with $p = 0.50$, $\gamma_{11} = 2$ for Type A interaction and compositional epistasis and $\gamma_{11} = -2$ for Type D interaction. The values of MSE, AIC, and BIC are shown on the natural logarithm scale for ease of visual inspection. The optimal additive GJ model performs similar to the full logistic model and performs better than the additive logistic model under Type A interaction and compositional epistasis. As expected, the optimal additive GJ model performs worse than the full logistic model under Type D interaction. Further, under Type D interaction, the optimal additive GJ model performs similar to the additive logistic model since the MLE of λ is generally close to 0 in this setting. The results for the 5×5 contingency tables under compositional epistasis suggest that the optimal additive GJ model also performs better than the logistic models in terms of AIC and BIC, but its MSE is larger than that of the full logistic model (average MSEs: 0.114 and 0.073 for the optimal additive GJ and the full logistic models, respectively). Note that, while compositional epistasis satisfies the monotonicity condition, the underlying curvature is not strictly monotonic. This lack of

strict monotonicity impacts the bias-variance trade-off (quantified by the MSE) when the number of risk factor sub-classes becomes large, but it does not impact the overall quality of the fit (quantified by AIC and BIC). Similar results were obtained under all the other parametric configurations.

Taken together, the above results suggest that the optimal additive GJ model is a parsimonious approach for obtaining precise estimates of the disease odds and for obtaining a good fit to the observed data when the interactions are removable.

8. Illustrative examples - revisited

We now illustrate the GJ transformation using the example data sets of Section 5. We provide a detailed illustration for the advanced colorectal adenoma and bladder cancer data, and only a brief summary of the method for the endometrial cancer data.

8.1 Advanced colorectal adenoma, smoking, and NAT2 acetylation

Table 2 shows the estimated log odds and their standard errors under the three models. The additive model provided a poor fit to the data relative to the full logistic model. In particular, the log odds for the current smokers were poorly estimated (see Figure 4).

For the GJ link, the profile MLE of λ was -4.62 (95% CI: $-12.50 - -1.02$, which excludes $\lambda=0$, the logit transformation). The optimal additive GJ model provided a better fit to the data than the additive logistic model (Table 2 and Figure 4). In particular, the log odds of the current smokers was remarkably well estimated relative to the additive logistic model. The log odds of the past smokers was not well estimated, because of a slight lack of monotonic trend in the log odds for this group, as noted in Section 5.1. For the current smokers, the standard errors of the estimated log odds were similar to the within class standard errors, suggesting a remarkably good information recovery for these sub-classes even without considering interactions in the model. For the remaining sub-classes, the standard errors of the estimated log odds were smaller than the within class standard errors. The optimal GJ model had the smallest MSE, AIC, and BIC, suggesting that it provides the best fit to the data among the three models considered (see Table 2), and its 95% confidence intervals for the log odds covered the observed values (Table 5).

Table 6 shows the parameter estimates, standard errors, and p-values under the three models. NAT2 acetylation was significant under the full logistic model. It had marginal significance under the additive logistic model, but was not significant under the optimal additive GJ model. This raises the question as to whether NAT2 has a significant effect on the risk of advanced colorectal adenoma. To investigate this, we considered a GJ model that included smoking, but not NAT2 acetylation, as the risk factor. The profile MLE of λ was 0. Therefore, we considered a logistic regression model with smoking alone as the risk factor in the model and estimated the log odds summaries. The MSE, AIC, and BIC of this model were 0.077, 1770.793, and 1791.455, respectively, which were substantially larger than those of the optimal additive GJ model, suggesting that there may be a small, though non-significant, additive effect of NAT2.

These results suggest that the seemingly significant interaction between smoking and NAT2 observed under the full logistic model reflects non-additivity or curvature effect that cannot be adequately accounted for by an additive logistic model, but one that can be accounted for by modeling the disease risk through an additive model under the GJ link. NAT2 slow acetylators who are current smokers have much larger log odds relative to the other risk factor sub-classes (Table 2). This large log odds is a result of not using a scale that can remove most of the statistical interaction. Although the main effect of NAT2 is not

statistically significant under the optimal additive GJ model (p -value = 0.24; Table 6), it does not preclude such an effect. Cigarette smoking results in exposure to aryl and heterocyclic amines. The NAT2 enzyme plays an important role in the metabolism of aromatic and heterocyclic amines. NAT2 may not be a primary contributor to disease risk at a biological level, but it may increase risk solely based on its ability to detoxify herterocyclic and aromatic amines when a person is exposed to these amines. Ochs-Balcom et al [39], who provide a detailed evaluation of NAT2 acetylation in breast cancer, suggest that such molecular explanation may be an important reason for the lack of significant main effect of the NAT2 phenotype in relation to certain diseases.

8.2 Bladder cancer, smoking, and NAT2 acetylation

The profile MLE of λ for the bladder cancer data was -3.2 (95% CI: $-8.62 - -1.15$). Table 3 shows the estimated log odds for the full logistic model, additive logistic model, and the optimal additive GJ model, the standard errors, MSE, AIC, and BIC. The optimal additive GJ model had the smallest MSE, AIC, and BIC and, thus, provided the best fit among the three models considered.

Under the optimal additive GJ model, the 95% CIs for the log odds included the observed log odds values in all the risk factor sub-classes (Table 5). In contrast, under the additive logistic model, the confidence intervals for never smokers and current smokers who were rapid/intermediate NAT2 acetylators did not include the observed values. Table 6 shows the parameter estimates, standard errors, and p -values under the three models. Both smoking and NAT2 acetylation were significantly associated with the risk of bladder cancer under all the models considered. As in the previous example, we note that the optimal additive GJ model provided log odds estimates that are close to the observed values (Figure 5) and standard errors that were generally close to the within class standard errors.

8.3 Endometrial cancer, tea intake and CYP19A1 genotype

Although the endometrial data do not show significant evidence for a removable interaction (Section 5.3), we briefly summarize the properties of an additive GJ model for these data without providing a detailed analysis. The profile MLE of λ was 0.01 (95% CI: $-3.01 - 4.50$). Since λ was near 0, the MSE, AIC, and BIC of the additive GJ model were similar to those of an additive logistic model (see Table 4). As expected, the full logistic model had an overall better performance than the additive GJ and the additive logistic models. These results are also consistent with the simulation findings of Section 7 for Type D interaction.

9. Discussion

The primary objective of this paper is to demonstrate that fitting an additive model to the data by eliminating a removable interaction via a transformation of the outcome may provide more precise estimates of the disease odds and a better fit to the data. To demonstrate this, we first developed a test statistic for evaluating whether an interaction is removable, and used the Guerrero and Johnson family of transformation to eliminate a removable interaction.

In developing the test for a removable interaction, we have assumed that there are no empty risk factor sub-classes. This is an important assumption since complete testing for interactions is not possible when there are empty sub-classes. To see this, consider X and Z to be just two binary risk factors. The interaction effect is defined as $\xi_{00} + \xi_{11} - \xi_{01} - \xi_{10}$. Suppose we do not have observations (neither cases nor controls) in the $(0,0)$ sub-class in the observed data, but these may occur in the population. Then the summary ξ_{00} is not observed, and we cannot estimate the interaction effect. When X and Z have more than two

levels, there are several γ_{ij} terms representing interactions. We need the relevant risk factor sub-classes to be non-empty in order to be able to estimate these γ_{ij} terms. Although we cannot estimate interaction effects for empty cells, it is perhaps reasonable to assume that they would be non-significant if the majority of (estimable) interactions are non-significant. Here “significance” relates more to finding a parsimonious model that leads to good estimates of main effects than to statistical significance. When there are empty risk factor sub-classes, we cannot estimate the main effects unless we assume that some of the interactions do not exist [42], which may be an unwarranted assumption. Elston and Bush [43] discuss in some detail what hypotheses can be tested in this situation when the outcome is continuous. Alternatively, when there are empty sub-classes it may be pragmatic to remove the estimable interactions through a transformation of the outcome to estimate the main effects. When some cells are small, but not empty, we may add one-half to the cells as an adjustment factor to account for small sample size and test for removable interactions. However, the circumstances under which this approach is adequate for the asymptotic distribution of the test to hold require further investigations.

Figures 7 and 8 suggest that the efficiency gains of the additive GJ model (in terms of minimum MSE, AIC, and BIC) may be better for large contingency tables. Testing the null hypothesis of no interaction under the full logistic model involves $(L_1-1) \times (L_2-1)$ degrees of freedom. The test for removable interaction has one degree of freedom, which leaves $(L_1-1) \times (L_2-1) - 1$ degrees of freedom for non-removable interaction. There is no point in trying to find a transformation in the case of a 2×2 table since the full logistic model has $(L_1-1) \times (L_2-1) = 1$ degrees of freedom, precluding a separation of the interaction into removable and non-removable components.

Once evidence for a removable interaction is obtained, it is generally useful to establish that there is no non-removable interaction i.e., there is no further interaction after transformation. In this paper we have focused solely on removable interaction. In practice, it may be fairly obvious that the interaction is largely removable (see Figure 4) or non-removable (see Table 4). But if there is any doubt, a complementary step would be to test for significant non-removable interaction after finding a parsimonious additive model for estimation. This amounts to testing the null hypothesis that $\gamma_{ij} = \theta\alpha_i\beta_j$ ($i = 1, \dots, L_1; j = 1, \dots, L_2$) against the alternative $\gamma_{ij} \neq \theta\alpha_i\beta_j$ for some (i,j) . Variance components method may be used to develop this test by first viewing $e_{ij} = \gamma_{ij} - \theta\alpha_i\beta_j$ as having *a priori* i.i.d $N(0, \delta^2)$ distribution and then testing whether $\sigma^2 = 0$ (see, for example [44]). Alternatively and as also noted by a reviewer, a likelihood ratio test may be used. A detailed evaluation of these tests is outside the scope of this paper.

There is also an increasing interest in testing for the presence of interactions or in identifying disease risk factors by incorporating interactions into relevant models in genome-wide association studies (GWAS). While our proposed method is not initially suitable for use in GWAS, it may be used to determine whether a statistical interaction, once it is found, is removable.

In this paper, we have considered the Guerrero and Johnson family of transformations. Using a Taylor's series expansion, we have shown in Section 6 that this is an appropriate family of transformation for fitting an additive model when the interaction is removable in the sense of Equation (5). Several alternative transformations such as the Aranda-Ordaz family, Burr's transformation, the identity transformation, and non-parametric link functions are available for fitting binary data [37, 38, 45–47]. A detailed evaluation of alternative transformations to additivity for our illustrative examples is outside the scope of this paper.

The importance of evaluating interactions has been discussed extensively in the literature [7]. One of the stimulating factors for considering interactions is the belief that interactions can provide useful insights about biological mechanisms underlying the disease. However, biological interactions can arguably occur regardless of whether or not there is a statistical interaction [12, 18]. Statistical models, on the other hand, can provide useful ways for fitting data to estimate the roles of risk factors on the outcome. In this work we have shown that model parsimony can be attained via transformation when the risk factors exhibit a curvature effect that obeys certain monotonicity properties, resulting in precise estimation and better fit. It has been argued that transformations must not be pursued because eliminating interactions may limit our ability to obtain biological insights about the risk factors [48]. However, while interaction is eliminated under the transformed scale to attain parsimony, the curvature (i.e., interaction) is indeed retained on the original scale and can be obtained via back-transformation. This is demonstrated in our illustrative examples, where the curvature-type effects are re-found via back transformation after estimating the log odds summaries using the parameter estimates of the additive model under the GJ transformation.

Our illustrative examples show that an additive model on the transformed scale is parsimonious and can provide a precise fit when an interaction is removable, making this an appealing model fitting approach. Our proposed method can be applied to data from a variety of application settings for evaluating removable statistical interactions. We hope that our work will contribute to further research in the area of fitting parsimonious models by removing interactions via a transformation of the outcome.

Acknowledgments

The authors thank Professor Duncan Thomas and an anonymous reviewer for several insightful comments that helped improve this manuscript. Satagopan's work was supported by research grants R01CA137420 from the National Cancer Institute, USA and UL1RR024996 of the Clinical and Translational Science Center at Weill Cornell Medical College, New York, USA. Elston's work was supported by supported by a grant from the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-220-C00004), Cancer Center Support grant P30CAD43703 from the National Cancer Institute, and grant UL1TR000439 from the National Human Genome Research Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*. 1994; 13:153–162. [PubMed: 8122051]
2. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genetic Epidemiology*. 2003; 24:150–157. [PubMed: 12548676]
3. Millstein J, Conti DV, Gilliland FD, Gauderman WJ. A testing framework for identifying susceptibility genes in the presence of epistasis. *American Journal of Human Genetics*. 2005; 78:15–27. [PubMed: 16385446]
4. Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. *Biostatistics*. 2008; 9:30–50. [PubMed: 17429103]
5. Chen GK, Thomas DC. Using biological knowledge to discover higher order interactions in genetic association studies. *Genetic Epidemiology*. 2010; 34:863–878. [PubMed: 21104889]
6. Wakefield J, De Vocht F, Hung RJ. Bayesian mixture modeling of gene-environment and gene-gene interactions. *Genetic Epidemiology*. 2010; 34:16–25. [PubMed: 19492346]
7. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*. 2009; 10:392–404.

8. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *American Journal of Human Genetics*. 2010; 86:6–22. [PubMed: 20074509]
9. Thomas DC. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annual Review of Public Health*. 2010; 21:21–36.
10. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*. 2003; 56:73–82. [PubMed: 14614241]
11. Wang X, Elston RC, Zhu X. The meaning of interaction. *Human Heredity*. 2010a; 70:269–277. [PubMed: 21150212]
12. Wang X, Elston RC, Zhu X. Statistical interaction in human genetics: how should we model it if we are looking for biological interaction? *Nature Reviews Genetics*. 2010b; 12:74.
13. Yates F. The principles of orthogonality and confounding in replicated experiments. *The Journal of Agricultural Science*. 1933; 23:108–145.
14. Scheffé, H. *The Analysis of Variance*. New York: Wiley; 1959.
15. Rothman KJ. Synergy and antagonism in cause-effect relationships. *American Journal of Epidemiology*. 1974; 99:385–388. [PubMed: 4841816]
16. Doll R. The age distribution of cancer: Implications for models of carcinogenesis. *Journal of the Royal Statistical Society, Series A (General)*. 1971; 134:133–166.
17. Whittemore AS, Keller JB. Quantitative theories of carcinogenesis. *SIAM Review*. 1978; 20:1–30.
18. Siemiatycki J, Thomas DC. Biological models and statistical interactions: an example from multi-stage carcinogenesis. *International Journal of Epidemiology*. 1981; 10:383–387. [PubMed: 7327838]
19. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*. 2002; 11:2463–2468. [PubMed: 12351582]
20. Phillips PC. Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*. 2008; 9:855–867.
21. VanderWeele TJ. Sufficient cause interactions and statistical interactions. *Epidemiology*. 2009; 20:6–13. [PubMed: 19234396]
22. VanderWeele TJ. Epistatic interactions. *Statistical Applications in Genetics and Molecular Biology*. 2010; 9 Article 1.
23. Box, GEP. *Improving Almost Anything: Ideas and Essays*. Revised Edition. New York: Wiley; 2006.
24. Elston RC. On additivity in the analysis of variance. *Biometrics*. 1961; 17:209–219.
25. Box GEP, Cox DR. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*. 1964; 26:211–252.
26. Guerrero VM, Johnson RA. Use of the Box-Cox transformation with binary response models. *Biometrika*. 1982; 69:309–314.
27. Moslehi R, Chatterjee N, Church TR, Chen J, Yeager M, Weissfeld J, Hein DW, Hayes RB. Cigarette smoking, N-acetyltransferase genes and the risk of advanced colorectal adenoma. *Pharmacogenomics*. 2006; 7:819–829. [PubMed: 16981843]
28. Rothman N, Garcia-Closas M, Chatterjee N, et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nature Genetics*. 2010; 42:978–984. [PubMed: 20972438]
29. Xu WH, Dai Q, Xiang YB, et al. Interaction of soy food and tea consumption with *CYP19A1* genetic polymorphisms in the development of endometrial cancer. *American journal of Epidemiology*. 2007; 166:1420–1430.
30. Agresti, A. *Categorical Data Analysis*. New York: Wiley; 2002.
31. Woolf B. On estimating the relation between blood group and disease. *Annals of Human Genetics*. 1955; 19:251–253. [PubMed: 14388528]
32. Cornfield J. A statistical problem arising from retrospective studies. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. 1956; 4:135–148.
33. Finney DJ. Main effects and interactions. *Journal of the American Statistical Association*. 1948; 43:566–571. [PubMed: 18107568]

34. Tukey JW. One degree of freedom for non-additivity. *Biometrics*. 1949; 5:232–242.
35. Haldane JBS. The interaction of nature and nurture. *Annals of Eugenics*. 1946; 13:197–205. [PubMed: 20282564]
36. Ottman R. An epidemiologic approach to gene-environment interaction. *Genetic Epidemiology*. 1990; 7:177–185. [PubMed: 2369997]
37. Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *American Journal of Human Genetics*. 2006; 79:1002–1016. [PubMed: 17186459]
38. Aranda-Ordaz FJ. On two families of transformations to additivity for binary regression data. *Biometrika*. 1981; 68:357–363.
39. Whittemore AS. Transformations to linearity in binary regression. *SIAM Journal on Applied Mathematics*. 1983; 43:703–710.
40. Stukel TA. Generalized logistic models. *Journal of the American Statistical Association*. 1988; 83:426–431.
41. Ochs-Balcom HM, Wiesner G, Elston RC. A meta-analysis of the association of N-acetyltransferase-2 gene (NAT2) variants with breast cancer. *American Journal of Epidemiology*. 2007; 166:246–254. [PubMed: 17535831]
42. Searle, SR. *Linear Models for Unbalanced Data*. New York: Wiley; 1987.
43. Elston RC, Bush N. The hypotheses that can be tested when there are interactions in an analysis of variance model. *Biometrics*. 1964; 20:681–698.
44. Stram DO, Lee JW. Variance components testing in the longitudinal mixed effects model. *Biometrics*. 1994; 50:1171–1177. [PubMed: 7786999]
45. Pregibon D. Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*. 1980; 29:15–23.
46. Czado C, Santner TJ. Orthogonalizing parametric link transformation families in binary regression analysis. *Canadian Journal of Statistics*. 1992; 20:51–61.
47. Newton MA, Czado C, Chappell R. Bayesian inference for semiparametric binary regression. *Journal of the American Statistical Association*. 1996; 91:142–153.
48. Rothman KJ. Synergy and antagonism in cause-effect relationships. *American Journal of Epidemiology*. 1974; 99:385–388. [PubMed: 4841816]
49. Sen PK. Asymptotic properties of maximum likelihood estimators based on conditional specification. *Annals of Statistics*. 1979; 7:1019–1033.

APPENDIX A. Derivation of the score statistic U

The log-likelihood of the binomial distribution (Equation 1) is given by :

$$L = \sum_{i=1}^{L_1} \sum_{j=1}^{L_2} N_{ij} \times \log(p_{ij}) + M_{ij} \times \log(1 - p_{ij}),$$

, where p_{ij} is given by Equation (5). Denote the linear predictor as $\eta_{ij} = \mu + \alpha_i + \beta_j + \theta\alpha_i\beta_j$. Let y_{ij} denote the first derivative of L with respect to η_{ij} . In vector notation, write $\boldsymbol{\eta} = (\eta_{11}, \eta_{12}, \dots, \eta_{L_1, L_2})$ and $\mathbf{y} = (y_{11}, y_{12}, \dots, y_{L_1, L_2})$. Let v_{ij} be the negative second derivative of L with respect to η_{ij} . Let $\mathbf{A}^{-1} = \text{diag}\{v_{ij}\}$. The parameters μ , α , β , and θ are estimated via the Fisher scoring approach in an iteratively reweighted least squares manner by considering the working model: $y_{ij} = \mu + \alpha_i + \beta_j + \theta\alpha_i\beta_j + \text{error}$, where $\text{Var}(y_{ij}) = v_{ij}$. The working likelihood function is $-1/2 \times (\mathbf{y} - \boldsymbol{\eta})^T \mathbf{A} (\mathbf{y} - \boldsymbol{\eta})$. The normal equations for estimating the unknown parameters are obtained by taking the derivative of this working likelihood with respect to the parameters. The score function for θ , evaluated at the null hypothesis, is obtained by taking the derivative of this working likelihood with respect to θ and evaluating it at $\theta = 0$. The resulting score function, denote U , is given by $U = \mathbf{1}^T \mathbf{D} \mathbf{A} (\mathbf{y} - \mu \mathbf{1} - \alpha - \beta)$,

where $\mathbf{1}$ is a vector of unities and $\mathbf{D}$ is a diagonal matrix with elements $\alpha_i\beta_j$. The term $\mathbf{y} - \mu\mathbf{1} - \boldsymbol{\alpha} - \boldsymbol{\beta}$ is the vector of residuals. All these quantities can be estimated using the observed data via an additive model (i.e., a model devoid of any interaction terms) as part of the iterative procedure.

APPENDIX B. Proof of Results 1 and 2

In the absence of an interactions, we have $E(\hat{\xi}_{ij}) = \xi_{ij} = \mu + \alpha_i + \beta_j$. When the weights are proportional and when there is no interaction, it follows from Equation 7 that $E(\hat{\mu}) = \mu$, $E(\hat{\alpha}_i) = \alpha_i$, and $E(\hat{\beta}_j) = \beta_j$. Therefore, we have for $E(\hat{\gamma}_{ij}) = 0$ Result 1.

To obtain Result 2, note that a contrast in $\hat{\alpha}_i$ is given by $\sum_{i=1}^{L_1} c_i \hat{\alpha}_i$ where c_i are real numbers

such that not all c_i are zero and $\sum_{i=1}^{L_1} c_i = 0$. When there is no interaction, the expected value of this contrast is obtained from Equation 7 as:

$$E\left(\sum_{i=1}^{L_1} c_i \hat{\alpha}_i\right) = \sum_{i=1}^{L_1} c_i \times \frac{\sum_{j=1}^{L_2} v_{ij} \times E(\hat{\xi}_{ij})}{\sum_{j=1}^{L_2} v_{ij}} = \sum_{i=1}^{L_1} c_i \times \frac{\sum_{j=1}^{L_2} v_{ij} \times (\mu + \alpha_i + \beta_j)}{\sum_{j=1}^{L_2} v_{ij}} = \sum_{i=1}^{L_1} c_i \alpha_i$$

, which does not depend upon the weights. Similarly, the expected value of any contrast in $\hat{\beta}_j$ can be shown to be independent of the choice of the weights. Scheffé [14] has derived similar results for two-way analysis of variance for balanced data. Our focus is on unbalanced data, but assuming that the risk factors arise in proportional frequencies.

APPENDIX C. Conditional expectation of the score under the alternative hypothesis, and the asymptotic null distribution of ΛR when at least one of the risk factors is not associated with the outcome

Under the alternative hypothesis, the conditional expectation of the score is given by:

$$\begin{aligned} E\left(\hat{U} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right) &= \sum_{i=1}^{L_1} \sum_{j=1}^{L_2} E\left\{\hat{\alpha}_i \times \hat{\beta}_j \times \left(\frac{1}{N_{ij}} + \frac{1}{M_{ij}}\right)^{-1} \times \hat{\gamma}_{ij} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right\} = \sum_{i=1}^{L_1} \sum_{j=1}^{L_2} \hat{\alpha}_i \times \hat{\beta}_j \times \left(\frac{1}{N_{ij}} + \frac{1}{M_{ij}}\right)^{-1} \times E\left(\hat{\gamma}_{ij} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right) \\ &= \theta \times \sum_{i=1}^{L_1} \sum_{j=1}^{L_2} \hat{\alpha}_i \\ &\times \hat{\beta}_j \times \left(\frac{1}{N_{ij}} + \frac{1}{M_{ij}}\right)^{-1} \times \alpha_i \times \beta_j. \end{aligned}$$

The first equality follows from Equation 6. The second equality is obtained by taking the expectation within the summation. The third equality follows from the assumption of removable interaction under the alternative hypothesis, under which $\gamma_{ij} = \theta \times \alpha_i \times \beta_j$.

The required asymptotic null distribution is obtained as follows (also see [47]). Suppose both the risk factors exhibit a main effect. Then, under the null hypothesis of no interaction, we have $\boldsymbol{\gamma} = 0$. The MLE $\hat{\boldsymbol{\gamma}}$ then satisfies $\sqrt{N}\hat{\boldsymbol{\gamma}} = \mathbf{P}_0^{-1} \mathbf{S}_0$, where N is the total sample size, $\mathbf{P}_0$ is the matrix of second derivatives of the log-likelihood with respect to the interaction effects γ_{ij} (keeping the main effects fixed at their values estimated under the additive model), and $\mathbf{S}_0$ is the vector of first derivatives of the log-likelihood with respect to the γ_{ij} (keeping the main effects fixed as above), and both $\mathbf{P}_0$ and $\mathbf{S}_0$ are evaluated under the null, i.e., at $\boldsymbol{\gamma}_{ij} = 0$

(for all $i = 1, \dots, L_1$ and $j = 1, \dots, L_2$). $\mathbf{P}_0^{-1}\mathbf{S}_0$ is a normal random variable with mean 0 under the null hypothesis of no interaction.

Suppose the risk factors X and Z are independent and neither is associated with the outcome. Then we have $\alpha_i = 0$ for all $i = 1, \dots, L_1$ and $\beta_j = 0$ for all $j = 1, 2, \dots, L_2$. Let $\mathbf{e}_1$ and $\mathbf{e}_2$ denote vectors of arbitrary real numbers of length L_1 and L_2 respectively. Now suppose we fit an additive model and examine evidence for the presence of an interaction - assuming that, perhaps, an interaction indeed exists. Thus, we assume that $\boldsymbol{\gamma}$ is not actually equal to zero. Now suppose the true value of $\boldsymbol{\gamma}$ is $\boldsymbol{\gamma}_0$, and it is near a boundary of 0 such that the null hypothesis of no interaction does not hold. Following Sen [45], we can conceive of

sequences of the form $\frac{\varepsilon_1}{N^{1/4}}$ corresponding to $\boldsymbol{\alpha}$ and $\frac{\varepsilon_2}{N^{1/4}}$ corresponding to $\boldsymbol{\beta}$ and a real number $\theta_0 \neq 0$ such that $\gamma_0 = \frac{\theta_0 \times \varepsilon_1 \times \varepsilon_2}{\sqrt{N}}$, where $\mathbf{e}_1 \times \mathbf{e}_2$ denotes a vector of length $L_1 \times L_2$ consisting of the product of every element of $\mathbf{e}_1$ with every element of $\mathbf{e}_2$.

Therefore, when the true value of $\boldsymbol{\gamma}$ is on a boundary of 0, we have:

$$\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) = \sqrt{N} \times \hat{\boldsymbol{\gamma}} - \sqrt{N} \times \boldsymbol{\gamma}_0 = \mathbf{P}_0^{-1}\mathbf{S}_0 - \theta_0 \times \mathbf{e}_1 \times \mathbf{e}_2.$$

suggesting that the asymptotic distribution of $\hat{\boldsymbol{\gamma}}$ takes the form of a normal random variable plus a real number. Hence, given the estimated parameters of the additive model, the asymptotic distribution of $\hat{U}$ takes the form of a normal variate plus a real number when the true value of $\boldsymbol{\gamma}$ is $\boldsymbol{\gamma}_0$, which may be on a boundary of 0 when neither risk factor is associated with the outcome. Therefore, the null distribution of Λ_R will take the form of a central (one d.f) chi-squared random variable plus a positive constant, resulting in an inflation in the type I error.

APPENDIX D. Monotonicity properties of the different types of interactions – illustrated for two independent binary risk factors

Suppose we have two independent binary risk factors X and Z . Let $\pi_{1.} = P(X=1)$ and $\pi_{.1} = P(Z=1)$. The model relating X and Z is given by $g\{E(Y|X,Z)\} = \mu + \beta X + \delta Z + \theta_{11}XZ$. Suppose we fit an additive model, ignoring the interaction term. The main effects of X and Z , obtained by weighting the study subjects equally, are $\beta + \pi_{.1}\theta_{11}$ and $\delta + \pi_{1.}\theta_{11}$, respectively. Our goal is to determine whether $g\{E(Y|X,Z)\}$ is a monotonic function of $C = X(\beta + \pi_{.1}\theta_{11}) + Z(\delta + \pi_{1.}\theta_{11})$ in the observed range of C . Without loss of generality, we assume that $\beta \geq 0$ and $\delta \geq 0$. The table below gives the expected values of C and $g\{E(Y|X,Z)\}$ for all the (X,Z) combinations.

(X,Z)	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
C	0	$\beta + \pi_{.1}\theta_{11}$	$\delta + \pi_{1.}\theta_{11}$	$\beta + \delta + (\pi_{1.} + \pi_{.1})\theta_{11}$
$g\{E(Y X,Z)\}$	0	β	δ	$\beta + \delta + \theta_{11}$

When $\beta = 0 = \delta$ and $\theta_{11} \neq 0$, we have one form of compositional epistasis. This table shows that, under this setting, $g\{E(Y|X,Z)\}$ is a monotonic function of C .

Type A interaction

Suppose $\beta > 0$ and $\delta = 0$. Then, under Type A interaction, we have $\theta_{11} > 0$. Since $\beta > \delta$ in this scenario, we have $g\{E(Y|X=1, Z=0)\} > g\{E(Y|X=0, Z=1)\}$. Then, monotonicity occurs when $\beta + \pi_{.1}\theta_{11} > \beta + \pi_{1.}\theta_{11}$ i.e., when $\pi_{.1} > \pi_{1.}$.

Suppose $\beta = \delta > 0$. Under Type A interaction, we have $\theta_{11} > -\beta$. In this scenario, we have $g\{E(Y|X=1, Z=0)\} = g\{E(Y|X=0, Z=1)\}$ and $g\{E(Y|X=1, Z=1)\} > g\{E(Y|X=1, Z=0)\}$. Monotonicity clearly holds in this case since $\pi_{1.}$ and $\pi_{.1}$ are between 0 and 1.

Type B (or C) interaction

Consider the scenario where $\beta > 0$ and $\delta = 0$. Under Type B interaction, we have $-\beta < \theta_{11} < 0$. When $\theta_{11} = -\beta$, monotonicity clearly holds. This is also true when $\theta_{11} = 0$. Hence, monotonicity is applicable in this scenario.

Suppose $\beta = \delta > 0$. For Type B interaction, we have $\theta_{11} = -\beta$. Monotonicity is clearly applicable in this case.

Type D interaction

Suppose $\beta > 0$ and $\delta = 0$. Under Type D interaction, we have $\theta_{11} < -\beta$. Suppose $\theta_{11} = -\tau\beta$, where $\tau > 1$. The expected values of C are 0, $\beta(1-\tau\pi_{.1})$, $-\tau\beta\pi_{1.}$, and $\beta(1-\tau\pi_{.1}-\tau\pi_{1.})$ when $(X, Z) = (0, 0), (1, 0), (0, 1),$ and $(1, 1)$, respectively. The corresponding values of $g\{E(Y|X, Z)\}$ are 0, β , 0, and $\beta(1-\tau)$, respectively. Clearly, monotonicity holds when $\tau = 1$. When $\tau > 1$, monotonicity holds only when $\beta(1-\tau\pi_{.1}-\tau\pi_{1.}) < -\tau\beta\pi_{1.}$ i.e., only when $\tau > 1/\pi_{.1}$, and not otherwise.

Suppose $\beta = \delta > 0$. Under Type D interaction, we have $\theta_{11} < -\beta$. As above, let $\theta_{11} = -\tau\beta$, where $\tau > 1$. The expected values of C are 0, $\beta(1-\tau\pi_{.1})$, $\beta(1-\tau\pi_{1.})$, and $\beta(2-\tau\pi_{.1}-\tau\pi_{1.})$ when $(X, Z) = (0, 0), (1, 0), (0, 1),$ and $(1, 1)$, respectively. The corresponding values of $g\{E(Y|X, Z)\}$ are 0, β , β , and $\beta \times (2-\tau)$, respectively. Monotonicity holds when $\tau = 1$. Suppose $1 < \tau < 2$. Then, monotonicity holds only when $\tau > \max\{1/\pi_{.1}, 1/\pi_{1.}\}$, and not otherwise. Suppose $\tau > 2$. In this case, monotonicity holds only when $\tau > 2/(\pi_{.1} + \pi_{1.})$, and not otherwise.

APPENDIX E. Parameter estimation and standard errors under the GJ transformation

For a given λ let Ω_λ denote the parameters of the model given by Equation (11). Thus, Ω_λ is a column vector of length $L_1 + L_2 - 1$, consisting of the baseline effect and the main effect contrasts. Let $\mathbf{W}$ denote the design matrix of the additive model of Equation 11. Let $\mathbf{W}_{ij}$ denote the row of $\mathbf{W}$ corresponding to the (i, j) -th risk factor sub-class. Denote

$$\pi_{ij} = 1 - \left\{ 1 + \left(1 + \lambda \mathbf{W}_{ij} \Omega_\lambda \right)^{1/\lambda} \right\}^{-1} \quad \text{and let } \Sigma \text{ denote a diagonal matrix of dimension } L_1 \times L_2 \text{ with}$$

diagonal elements of the form $\Sigma_{ij} = \left(\frac{M_{ij}}{N_{ij}} \right)^{2\lambda} \times (N_{ij} + M_{ij}) \times \pi_{ij} \times (1/\pi_{ij})$. Finally, let $\mathbf{y}$ be a column vector of length $L_1 \times L_2$ with elements of the form

$$\mathbf{W}_{ij} \Omega_\lambda + \left(\frac{M_{ij}}{N_{ij}} \right)^\lambda \times \{ N_{ij} - (N_{ij} + M_{ij}) \pi_{ij} \} \times \Sigma_{ij}^{-1}.$$

For a given λ , the MLEs of the unknown parameters Ω_λ can be obtained via the following iteratively reweighted least squares procedure. Let $\Omega_\lambda^{(t)}$ denote the estimates in the t -th iteration. In the $(t+1)$ -th iteration, the updated value of the parameters is given by:

$$\Omega_\lambda^{(t+1)} = (\mathbf{W}^T \Sigma^{(t)} \mathbf{W})^{-1} \mathbf{W}^T \Sigma^{(t)} \mathbf{y},$$

where the superscript (t) for Σ and $\mathbf{y}$ on the right hand side indicates that these quantities are calculated by plugging in the value of Let $\Omega_\lambda^{(t)}$ for Ω_λ . These calculations are repeated until convergence of the log-likelihood function. In practice, we have found this approach to be quite rapid, with convergence occurring within 5 iterations.

We obtain the variance of the MLEs by inverting the information matrix, which is calculated as the negative second derivative of the log-likelihood function with respect to the

parameters Ω_λ and λ . The information matrix is given by $\mathfrak{I} = \begin{pmatrix} \mathfrak{I}_{\Omega\Omega} & \mathfrak{I}_{\Omega\lambda} \\ \mathfrak{I}_{\Omega\lambda}^T & \mathfrak{I}_{\lambda\lambda} \end{pmatrix}$, where

$\mathfrak{I}_{\Omega\Omega} = \mathbf{W}^T \Sigma \mathbf{W} + \lambda \times \mathbf{W}^T \text{diag} \left\{ \frac{N_{ij} - (N_{ij} + M_{ij}) \pi_{ij}}{(1 + \lambda \mathbf{W}_{ij} \Omega_\lambda)^2} \right\} \mathbf{W}$ is a square matrix of dimension $L_1 + L_2 - 1$, $\mathfrak{I}_{\Omega\lambda}$ is a column vector of length $L_1 + L_2 - 1$ with elements of the form:

$$\sum_{i=1}^{L_1} \sum_{j=1}^{L_2} \mathbf{W}_{ij}^T \left\{ \frac{\mathbf{W}_{ij} \Omega_\lambda \times \{N_{ij} - (N_{ij} + M_{ij}) \pi_{ij}\}}{(1 + \lambda \mathbf{W}_{ij} \Omega_\lambda)^2} + \frac{(N_{ij} + M_{ij}) \times \pi_{ij} \times (1 - \pi_{ij})}{1 + \lambda \mathbf{W}_{ij} \Omega_\lambda} \times \left(\frac{-1}{\lambda^2} \log(1 + \lambda \mathbf{W}_{ij} \Omega_\lambda) + \frac{1}{\lambda} \times \frac{\mathbf{W}_{ij} \Omega_\lambda}{1 + \lambda \mathbf{W}_{ij} \Omega_\lambda} \right) \right\},$$

and

$$I_{\lambda\lambda} = \sum_{i=1}^{L_1} \sum_{j=1}^{L_2} (N_{ij} + M_{ij}) \times \pi_{ij} \times (1 - \pi_{ij}) \times \left(\frac{-1}{\lambda^2} \log(1 + \lambda \mathbf{W}_{ij} \Omega_\lambda) + \frac{1}{\lambda} \times \frac{\mathbf{W}_{ij} \Omega_\lambda}{1 + \lambda \mathbf{W}_{ij} \Omega_\lambda} \right)^2 - \left\{ N_{ij} - (N_{ij} + M_{ij}) \pi_{ij} \right\} \times \left\{ \frac{2}{\lambda^3} \log(1 + \lambda \mathbf{W}_{ij} \Omega_\lambda) - \frac{2}{\lambda^2} \times \frac{\mathbf{W}_{ij} \Omega_\lambda}{1 + \lambda \mathbf{W}_{ij} \Omega_\lambda} - \frac{1}{\lambda} \times \left(\frac{\mathbf{W}_{ij} \Omega_\lambda}{1 + \lambda \mathbf{W}_{ij} \Omega_\lambda} \right)^2 \right\}$$

is a scalar. Further, the components of the information matrix are calculated by plugging in the MLE of Ω_λ in the above quantities. The covariance matrix of the MLE of Ω_λ is obtained

by inverting the information matrix and is given by $\text{Var}(\widehat{\Omega}_\lambda) = \left(\mathfrak{I}_{\Omega\Omega} - \frac{\mathfrak{I}_{\Omega\lambda} \mathfrak{I}_{\Omega\lambda}^T}{\mathfrak{I}_{\lambda\lambda}} \right)^{-1}$. The variances of the parameter estimates are obtained as the diagonal elements of this matrix.

Let $f(\Omega, \lambda) = \frac{1}{\lambda} \times \log(1 + \lambda \mathbf{W} \Omega_\lambda)$ denote the log odds values, estimated by plugging in the MLE of Ω_λ . Note that $f(\cdot)$ is a column vector of length $L_1 \times L_2$. Using the delta method, the covariance matrix of the estimated log odds is given by

$\text{Var}(f) = \left(\frac{\partial f}{\partial \Omega_\lambda} \frac{\partial f}{\partial \lambda} \right) \text{Var}(\widehat{\Omega}_\lambda) \left(\frac{\partial f}{\partial \Omega_\lambda} \frac{\partial f}{\partial \lambda} \right)^T$, where the derivatives are evaluated at the MLEs of the respective parameters. The variance of the log odds estimate for the (i,j) -th risk factor sub-class is then obtained as the relevant diagonal element of this matrix.

Figure 1.

The different types of interactions as classified by Haldane [34]. I_X and I_Z are two binary risk factors. ξ_{00} , ξ_{01} , ξ_{10} , and ξ_{11} are the outcomes in the four risk factor sub-classes. In each panel, the horizontal axis shows the two values of I_X , and the vertical axis shows the ξ_{ij} values, ordered differently under the different types of Interactions. The black and red lines show the outcomes as a function of I_X when $I_Z = 0$ and $I_Z = 1$, respectively. Type A interaction has two subtypes – one of them is shown in the top left panel, and the second subtype can be obtained by switching the ordering of ξ_{10} and ξ_{01} , which will not alter the shape of the figure. Type D interaction has two subtypes in a similar manner.

Figure 2.

Quantile-quantile plots of the empirical values of Λ_R obtained under the null hypothesis of no interaction and expected values under a central chi-squared distribution with one degree of freedom under various parametric configurations. The 3 rows correspond to data generated with correlation between the risk factors as $\rho = 0$ (uncorrelated), 0.50, and 0.80. The columns correspond to null simulations with different magnitudes of the main effects.

Figure 3. Empirical power of Λ_R and Λ_G obtained with $\alpha_1 = \beta_2 = \log(2)$ under Type A interaction (Column 1) and Type D interaction (Column 2) under different levels of correlations between the risk factors (Rows 1, 2, and 3). Empirical power under compositional epistasis is shown in Column 3. In each panel, the horizontal axis represents the magnitude of the interaction parameter γ_{11} used in the simulations. The black lines denote the empirical power of Λ_R and the red lines are the empirical power of Λ_G .

Figure 4.

Estimated log odds ratios for the **advanced colorectal adenoma study**. The horizontal axis shows the three levels of smoking. The vertical axis shows the log odds. The observed log odds for NAT2 rapid/intermediate acetylators and slow acetylators are shown in black bold and dashed lines, respectively. The estimated log odds under the full logistic model coincide with the observed values. The estimates obtained under the additive logistic model for NAT2 rapid/intermediate acetylators and slow acetylators are shown in green bold and dashed lines, respectively. The estimates based on the optimal additive GJ model for NAT2 rapid/intermediate and slow acetylators are shown in red bold and dashed lines, respectively.

Figure 5.

Estimated log odds ratios for the **bladder cancer study**. The horizontal axis shows the three levels of smoking. The vertical axis shows the log odds. The observed log odds for NAT2 rapid/intermediate acetylators and slow acetylators are shown in black bold and dashed lines, respectively. The estimated log odds under the full logistic model coincide with the observed values. The estimates obtained under the additive logistic model for NAT2 rapid/intermediate acetylators and slow acetylators are shown in green bold and dashed lines, respectively. The estimates based on the optimal additive GJ model for NAT2 rapid/intermediate and slow acetylators are shown in red bold and dashed lines, respectively.

Figure 6.

Shape of the logit and GJ transformations. The horizontal axis shows a real number x . The vertical axis shows the disease probability as a function of x (see Equation 10). Setting $\lambda = 0$ in Equation 10 gives the logit function, shown as a black curve. The red (green) bold and dashed curves show the disease probability as a function of x based on the GJ transformation with $\lambda = -3$ (-0.5) and 3 (0.5), respectively.

Figure 7.

Smoothed kernel density plots of MSE (row 1), AIC (row 2) and BIC (row 3) for data simulated based on a **2×3 contingency table** under interactions of Type A (left column), Type D (middle column) and compositional epistasis (right column). These results correspond to the parametric configuration where $\beta = \log(2) = 0.6931 = \delta$ (Type A and Type D interactions), $\beta = 0 = \delta$ (compositional epistasis), $\theta_{11} = 2$ (Type A interaction and compositional epistasis), $\theta_{11} = -2$ (Type D interaction), and $\rho = 0.50$. The smoothed kernel density estimates were obtained using the `density` function in the R programming language with the default settings for bandwidth. Under Type D interaction, the MSEs of the

optimal additive GJ model and the additive logistic model are similar. Hence, the two curves overlap. In order to distinguish the two colors (green and red), we have shown these two curves using dashed lines for ease of visualization. Similar approach is used to show the AIC curves for the full logistic and the optimal additive GJ models under Type A interaction and compositional epistasis.

Figure 8.

Smoothed kernel density plots of MSE (row 1), AIC (row 2) and BIC (row 3) for data simulated based on a **5×5 contingency table** under interactions of Type A (left column), Type D (middle column) and compositional epistasis (right column). These results correspond to the parametric configuration where $\beta = \log(2) = 0.6931 = \delta$ (Type A and Type D interactions), $\beta = 0 = \delta$ (compositional epistasis), $\theta_{11} = 2$ (Type A interaction and compositional epistasis), $\theta_{11} = -2$ (Type D interaction), and $\rho = 0.50$. The smoothed kernel density estimates were obtained using the `density` function in the R programming language with the default settings for bandwidth. Under Type D interaction, the MSEs of the

optimal additive GJ model and the additive logistic model are similar. Hence, the two curves overlap. In order to distinguish the two colors (green and red), we have shown these two curves using dashed lines for ease of visualization.

Table 1

Empirical type I errors of Λ_R and Λ_G based on 1000 simulations under each parametric configuration. Here α_1 and β_1 are the main effects of the two risk factors, and ρ is the correlation between the risk factors.

ρ	α_1	β_1	Λ_R	Λ_G
0	$\log(1) = 0$	0	0.100	0.047
	$\log(2) = 0.6931$	$\log(2)$	0.031	0.048
	$\log(3) = 1.0986$	$\log(3)$	0.042	0.055
0.50	$\log(1) = 0$	0	0.049	0.045
	$\log(2) = 0.6931$	$\log(2)$	0.056	0.048
	$\log(3) = 1.0986$	$\log(3)$	0.050	0.057
0.80	$\log(1) = 0$	0	0.035	0.046
	$\log(2) = 0.6931$	$\log(2)$	0.071	0.049
	$\log(3) = 1.0986$	$\log(3)$	0.067	0.046

Case-control data from the **advanced colorectal adenoma** study of Moslehi et al [27]. Columns 1 and 2 show the levels of smoking and NAT2 acetylation (based on SNPs rs1495741). Columns 5, 6, and 7 show the estimated log odds and the standard errors in parentheses based on three models. Note that the full logistic model is a saturated model. Therefore, for this model, the estimated log odds is equal to the observed log odds, and the estimated standard error is the square root of the sum of the inverse of the number of cases and the number of controls. We refer to these standard errors as the “within class” standard errors. The value of λ used to fit the additive GJ model is shown in parentheses in the last column. The last three rows of the table show the maximum log-likelihood, mean squared error (MSE), Akaike’s AIC, and the Bayes information criterion for the three models.

Table 2

1	2	3	4	5–7 Estimated log odds (std error) from three models		
Smoking	NAT2 acetylation	Case	Control	Full logistic	Additive logistic	Optimal additive GJ ($\lambda = -4.62$)
Never	Rapid/Intermediate	92	124	−0.298 (0.138)	−0.319 (0.111)	−0.197 (0.078)
Never	Slow	140	158	−0.121 (0.116)	−0.106 (0.101)	−0.153 (0.095)
Past	Rapid/Intermediate	108	116	−0.071 (0.134)	−0.226 (0.109)	−0.132 (0.080)
Past	Slow	134	153	−0.133 (0.118)	−0.013 (0.102)	−0.070 (0.094)
Current	Rapid/Intermediate	60	52	0.143 (0.189)	0.499 (0.143)	0.163 (0.172)
Current	Slow	115	42	1.007 (0.180)	0.712 (0.137)	1.007 (0.180)
Maximum log-likelihood				−875.308	−879.646	−875.868
MSE				0.0221	0.0561	0.0186
AIC				1762.617	1767.293	1761.735
BIC				1793.61	1787.955	1787.563

Table 3

Case-control data from the **bladder cancer study** of Rothman et al [28]. Columns 1 and 2 show the levels of smoking and NAT2 acetylation(based on SNP rs1495741). Columns 5, 6, and 7 show the estimated log odds and the standard errors in parentheses based on three models. Note that the full logistic model is a saturated model. Therefore, for this model, the estimated log odds is equal to the observed log odds, and the estimated standard error is the square root of the sum of the inverse of the number of cases and the number of controls. We refer to these standard errors as the “within class” standard errors. The value of λ used to fit the additive GJ model is shown in parentheses in the last column. The last three rows of the table show the maximum log-likelihood, mean squared error(MSE), Akaike’s AIC, and the Bayes information criterion for the three models.

1	2	3	4	5–7 Estimated log odds (std error) from three models		
Smoking	NAT2 acetylation	Case	Control	Full logistic	Additive logistic	Optimal additive GJ($\lambda = -3.2$)
Never	Rapid/Intermediate	760	1679	-0.793 (0.044)	-0.912 (0.032)	-0.825 (0.027)
Never	Slow	1202	2758	-0.831 (0.035)	-0.756 (0.029)	-0.808 (0.028)
Past	Rapid/Intermediate	1859	2300	-0.213 (0.031)	-0.194 (0.026)	-0.199 (0.023)
Past	Slow	3455	3559	-0.030 (0.024)	-0.041 (0.021)	-0.034 (0.023)
Current	Rapid/Intermediate	1165	1254	-0.074 (0.041)	-0.004 (0.030)	-0.086 (0.033)
Current	Slow	2258	1865	0.191 (0.031)	0.150 (0.027)	0.194 (0.031)
Maximum log-likelihood				-16178.38	-16186.88	-16179.04
MSE				0.0012	0.0052	0.0011
AIC				32368.76	32381.76	32368.08
BIC				32417.30	32414.12	32408.52

Case-control data from the **endometrial cancer study** of Xu et al [29]. Columns 1 and 2 show the *CYP19A1* genotypes based on rs**** and the levels of tea intake. Columns 5, 6, and 7 show the estimated log odds and the standard errors in parentheses based on three models. Note that the full logistic model is a saturated model. Therefore, for this model, the estimated log odds is equal to the observed log odds, and the estimated standard error is the square root of the sum of the inverse of the number of cases and the number of controls. We refer to these standard errors as the “within class” standard errors. The value of λ used to fit the additive GJ model is shown in parentheses in the last column. The last three rows of the table show the maximum log-likelihood, mean squared error(MSE), Akaike’s AIC, and the Bayes information criterion for the three models.

Table 4

1	2	3	4	5–7 Estimated log odds (std error) from three models		
<i>CYP19A1</i>	<i>Tea Intake</i>	Case	Control	Full logistic	Additive logistic	Optimal additive GJ ($\lambda = 0.01$)
GG	Low	211	226	−0.069 (0.096)	0.064 (0.085)	0.064 (0.085)
GG	High	117	90	0.262 (0.140)	−0.019 (0.102)	−0.020 (0.102)
GT	Low	382	322	0.171 (0.076)	0.098 (0.069)	0.098 (0.069)
GT	High	148	171	−0.144 (0.112)	0.015 (0.091)	0.015 (0.091)
TT	Low	126	153	−0.194 (0.120)	−0.220 (0.106)	−0.219 (0.106)
TT	High	45	65	−0.368 (0.194)	−0.302 (0.123)	−0.303 (0.123)
Maximum log-likelihood				−1416.561	−1421.109	−1421.109
MSE				0.016	0.032	0.032
AIC				2845.121	2850.217	2852.217
BIC				2878.892	2872.731	2880.36

Table 5

95% confidence intervals for the log odds estimated under the additive logistic model and the optimal additive GJ model for two example data sets.

Smoking	NAT2	Observed log odds	95% confidence interval	
			Additive logit model	Optimal additive GJ model
ADVANCED COLORECTAL ADENOMA STUDY				
Never	Rapid/Intermediate	−0.298	−0.537, −0.102	−0.352, −0.042
Never	Slow	−0.121	−0.303, 0.091	−0.339, 0.033
Past	Rapid/Intermediate	−0.071	−0.441, −0.011	−0.289, 0.024
Past	Slow	−0.133	−0.212, 0.187	−0.254, 0.115
Current	Rapid/Intermediate	0.143	0.218, 0.780	−0.174, 0.499
Current	Slow	1.007	0.443, 0.981	0.653, 1.361
BLADDER CANCER STUDY				
Never	Rapid/Intermediate	−0.793	−0.975, −0.849	−0.879, −0.772
Never	Slow	−0.831	−0.815, −0.702	−0.862, −0.753
Past	Rapid/Intermediate	−0.213	−0.245, −0.144	−0.245, −0.154
Past	Slow	−0.030	−0.083, 0.002	−0.079, 0.010
Current	Rapid/Intermediate	−0.074	−0.063, −0.055	−0.150, −0.021
Current	Slow	0.191	0.098, 0.202	0.134, 0.254

Table 6

Maximum likelihood estimates (MLE), standard errors (Std. Err), and p-values of the parameters of three models for the two example data sets.

Parameter	Full Logistic Model		Additive Logistic Model		Optimal Additive GJ Model	
	MLE (Std. Err)	p-value	MLE (Std. Err)	p-value	MLE (Std. Err)	p-value
ADVANCED COLORECTAL ADENOMA STUDY						
Intercept	0.088 (0.061)	0.148	0.091 (0.060)	0.131	-0.080 (0.081)	0.332
Main Effects						
Smoking: Past vs Never	0.054 (0.063)	0.396	0.047 (0.063)	0.456	0.070 (0.115)	0.543
Smoking: Current vs Never	0.244 (0.048)	4.9e-07	0.257 (0.048)	6.3e-08	0.122 (0.039)	0.002
NAT2: Slow vs Rapid/Intermediate	0.163 (0.061)	0.007	0.107 (0.057)	0.062	0.050 (0.042)	0.240
Interaction						
Past Smoking × Slow NAT2	-0.059 (0.063)	0.347	-	-	-	-
Current Smoking × Slow NAT2	0.134 (0.048)	0.006	-	-	-	-
BLADDER CANCER STUDY						
Intercept	-0.291 (0.014)	< 0.001	-0.293 (0.014)	< 0.001	-1.362 (0.516)	0.008
Main Effects						
Smoking: Past vs Never	0.345 (0.017)	< 0.001	0.359 (0.017)	< 0.001	1.896 (0.763)	0.013
Smoking: Current vs Never	0.175 (0.010)	< 0.001	0.183 (0.009)	< 0.001	0.693 (0.253)	0.006
NAT2: Slow vs Rapid/Intermediate	0.068 (0.014)	< 0.001	0.077 (0.014)	< 0.001	0.121 (0.021)	< 0.001
Interaction						
Past Smoking × Slow NAT2	0.055 (0.017)	0.001	-	-	-	-
Current Smoking × Slow NAT2	0.032 (0.010)	0.002	-	-	-	-