Published in final edited form as:

Stat Methods Med Res. 2017 April ; 26(2): 1021–1038. doi:10.1177/0962280214567140.

STATISTICAL INTERACTIONS AND BAYES ESTIMATION OF LOG ODDS IN CASE-CONTROL STUDIES

Jaya M. Satagopan¹, Sara H. Olson¹, and Robert C. Elston²

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

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

Abstract

This paper is concerned with the estimation of the logarithm of disease odds (log odds) when evaluating two risk factors, whether or not interactions are present. Statisticians define interaction as a departure from an additive model on a certain scale of measurement of the outcome. Certain interactions, known as removable interactions, may be eliminated by fitting an additive model under an invertible transformation of the outcome. This can potentially provide more precise estimates of log odds than fitting a model with interaction terms. In practice, we may also encounter non-removable interactions. The model must then include interaction terms, regardless of the choice of the scale of the outcome. However, in practical settings, we do not know at the outset whether an interaction exists, and if so whether it is removable or non-removable. Rather than trying to decide on significance levels to test for the existence of removable and non-removable interactions, we develop a Bayes estimator based on a squared error loss function. We demonstrate the favorable bias-variance trade-offs of our approach using simulations, and provide empirical illustrations using data from three published endometrial cancer case-control studies. The methods are implemented in an R program, available freely at <http://www.mskcc.org/biostatistics/~satagopj>.

Keywords

Bayes estimator; compositional epistasis; logistic link; mean squared error; minimax estimator; non-removable interaction; removable interaction; transformation

INTRODUCTION

One of the main objectives of case-control studies is to estimate the natural logarithm of disease odds (log odds) corresponding to the categorical levels of the risk factors of interest. This is an important epidemiologic parameter, which can facilitate the estimation of odds

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

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ratios and absolute risk of disease. It can also provide insights into the potential benefits of screening high-risk individuals [1]. This is generally done by estimating the effects of the individual risk factors using a logistic regression model. When multiple risk factors are examined, it becomes important to decide whether or not interactions between the risk factors must be included in the model to obtain accurate estimates of the log odds. The purpose of this paper is to develop an optimal estimation procedure for case-control studies when we wish to estimate the log odds corresponding to two risk factors, whether or not interactions are present.

The increasing ability to study multiple environmental and genetic factors for disease is contributing to a proliferation of research works in the evaluation of gene-gene and gene-environment interactions [2–4]. The rapidly growing interests in studies of novel treatments are contributing to investigations of gene-drug interactions [5–6]. A wide range of definitions is used to refer to the term interaction [7]. An additive model is defined as one in which the additive effects on the outcome of the various levels of one risk factor do not depend upon the levels of another risk factor. Whereas epidemiologists describe such a model as one having “additive interaction” [8], statisticians define interaction as a *departure* from an additive model [9,10]. This is commonly referred to as a statistical interaction. Throughout this paper we shall only use the word interaction in this statistical sense.

Certain interactions, referred to as removable interactions, may be eliminated via an invertible transformation of the outcome so that the resulting model is additive on the transformed scale [11]. The results can be back-transformed for clinical interpretation, and the interactions will reappear in the model upon back-transformation [7, 12]. When the disease trait is binary, a transformation corresponds to a link function [13]. When an interaction is removable, accurate and precise estimates of the log odds parameters can be obtained by fitting a parsimonious additive model under a suitable link function [13]. We define an accurate estimate as one having negligible bias, and a precise estimate as one having small standard error. In this paper, we first show that the Guerrero and Johnson [14] (abbreviated, GJ) link function is an appropriate transformation to additivity when an interaction under the logistic link is removable.

Not all interactions are removable. Non-removable interactions are also referred to as qualitative interactions [15, 16]. When a non-removable interaction exists, an additive model will not usually provide accurate estimates of the log odds, regardless of the choice of transformation, and interaction terms must be included in the model to obtain unbiased estimates of the log odds. However, in practical data analysis settings, we cannot know with certainty at the outset whether a non-removable interaction exists. In principle, we may conduct preliminary hypothesis tests for the existence of removable and non-removable interactions. Rather than trying to decide on significance levels for these tests, in this paper we develop a Bayes estimator for the log odds parameters by assuming a squared error loss function. We minimize the loss function subject to the condition that the resulting class of Bayes estimators includes minimax estimators in the limit.

This paper is organized as follows. In the Materials and Methods section, we first introduce some notations and describe the concept of removable interactions. Next, we describe the GJ

link function and show that it is an appropriate link function to additivity when an interaction under the logistic link is removable. We also show that, when the model is additive under the GJ link, the logistic link function can result in a systematic departure from additivity. Thus, a suitable model under the logistic link function may be used to estimate the log odds when the model is additive under the GJ link. Since some interactions may be non-removable, we develop a Bayes estimation approach for obtaining precise estimates of log odds whether or not all interaction is removable. A main advantage of the proposed Bayes estimator is that it does not require preliminary hypothesis tests to determine whether an interaction exists and/or whether it is removable or non-removable, in order to decide how to estimate the log odds parameters. In the Results section, we first demonstrate the favorable bias-variance trade-offs of the Bayes estimator using simulations, and then illustrate our method using published data from three case-control studies of endometrial cancer [17–19]. These data represent three distinct types of interactions: some removable and some non-removable interaction [17], only removable interaction [18], and only non-removable interaction [19]. They help illustrate that the proposed Bayes estimation method gives similar estimates to what would have been otherwise found by testing separately for the presence of removable and non-removable interactions under some choice of significance levels for these tests. We have developed a computer program to implement the proposed methods, which we note in the section Computational Guidance for Practitioners, and conclude with a Discussion. We describe the main results in the paper and refer the reader to the Online Supplementary Material for technical details.

MATERIALS AND METHODS

Consider a case-control study with two risk factors X and Z having L_1 and L_2 levels, respectively, measured on each individual. Let N_{ij} and M_{ij} denote the number of affected cases and unaffected 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} . We assume that there are no empty sub-classes. We generally fit a logistic regression model to case-control data, written as:

$$g_0(p_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij}, \quad (\text{Equation 1})$$

where $g_0(p_{ij}) = \log \left\{ \frac{p_{ij}}{1-p_{ij}} \right\}$ is the logistic link function, μ is the baseline risk, α_i and β_j are the main effects of the i -th level of X and the j -th level of Z, respectively, 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

$\sum_{i=1}^{L_1} \alpha_i = 0$, $\sum_{j=1}^{L_2} \beta_j = 0$, $\sum_{i=1}^{L_1} \gamma_{ij} = 0$, and $\sum_{j=1}^{L_2} \gamma_{ij} = 0$ for $i = 1, \dots, L_1$, and $j = 1, \dots, L_2$. The maximum likelihood estimates (MLEs) of the main and interaction effects can be obtained using the iteratively reweighted least squares method.

Suppose there exists an alternative, but unknown, link function, denoted $f(p_{ij})$, under which the model is additive i.e., $f(p_{ij}) = \mu + \alpha_i + \beta_j$. If $g_0(\cdot)$ is a linear function of $f(\cdot)$, then the model is also additive under $g_0(\cdot)$ i.e., there will be no interaction term in the model under the logistic link function. However, if $g_0(\cdot)$ is non-linear in $f(\cdot)$, then a quadratic approximation may provide a better fit to the data than an additive model under $g_0(\cdot)$: $g_0(p_{ij}) = \eta_0 + \eta_1\{f(p_{ij})\} + \eta_2\{f(p_{ij})\}^2$, where η_0 , η_1 , and η_2 are unknown parameters. When this quadratic polynomial is monotonic in $f(p_{ij})$ in the range defined by the data, we obtain the approximation [10, 11, 13]:

$$g_0(p_{ij}) \approx \mu + \alpha_i + \beta_j + \theta \times \alpha_i \times \beta_j, \quad (\text{Equation 2})$$

where θ is a scalar quantifying non-additivity of the model under the logistic link. In previous work [13] we have shown how to obtain MLEs of the parameters θ , μ , α_i , and β_j in Equation (2).

A comparison of Equations (1) and (2) shows that we can approximate the interaction contrasts as $\gamma_{ij} \approx \theta \alpha_i \beta_j$ when monotonicity holds. Thus, under monotonicity, there will be $(L_1-1) \times (L_2-1) - 1$ fewer parameters in the model. In practical settings, we do not know whether this approximation is applicable since we do not know at the outset whether monotonicity holds. Here we propose to take advantage of potential monotonicity and write the interaction terms as:

$$\gamma_{ij} = \theta \alpha_i \beta_j + e_{ij}, \quad (\text{Equation 3})$$

where the terms e_{ij} ($i = 1, \dots, L_1$; $j = 1, \dots, L_2$) can be interpreted as the error in representing γ_{ij} as $\theta \alpha_i \beta_j$ when monotonicity does not hold. When $e_{ij} = 0$ for all i and j , and $\theta = 0$ in Equation (3), this would be an indication of monotonicity, and we say that the interaction is **removable**. When $e_{ij} \neq 0$ for at least one i and j , this would be an indication of lack of monotonicity, and we say that there is **non-removable** interaction.

Our objective is to estimate the log odds. The following four scenarios arise depending upon whether an interaction exists and whether it is removable or non-removable:

Scenario 1 (no interaction). When $\theta = 0$ and $e_{ij} = 0$ for all i and j , there is no removable and no non-removable interaction i.e., there is no interaction at all ($\gamma_{ij} = 0$ for all i and j). Hence, the log odds summaries can be estimated using an additive model under the logistic link function i.e., using Equation (1) by setting $\gamma_{ij} = 0$ for all i and j .

Scenario 2 (removable interaction). When $\theta \neq 0$ but $e_{ij} = 0$ for all i and j , there is removable interaction and no non-removable interaction. Hence, there exists a transformation to additivity. We anticipate the log odds summaries estimated using a transformation to additivity to be more precise (i.e., smaller standard error) than estimates based on Equation (1) since an additive model will have fewer parameters than Equation (1).

Scenario 3 (non-removable interaction). When $\theta = 0$ and $e_{ij} = 0$ for at least one i and j , there is no removable interaction and there is only non-removable interaction. Therefore, a transformation to additivity is not feasible, and the log odds summaries must be estimated using the logistic regression model of Equation (1).

Scenario 4 (both removable and non-removable interactions). When $\theta = 0$ and $e_{ij} = 0$ for at least one i and j , there are both removable and non-removable interactions. In this case, the log odds summaries may be estimated using Equation (1), or by making suitable use of the parametric form of γ_{ij} given by Equation (3). Since this latter approach takes advantage of modeling the removable component of the interaction suitably as $\theta\alpha_i\beta_j$, we anticipate this method would estimate the log odds summaries with better precision than Equation (1).

In practical settings, at the outset we do not know which of these four scenarios is applicable. One approach would be to test the null hypothesis $H_0: \gamma_{ij} = 0$ for all $i = 1, 2, \dots, L_1-1$ and $j = 1, 2, \dots, L_2-1$ against the alternative hypothesis $H_A: \gamma_{ij} = 0$ for at least one (i,j) using a likelihood ratio statistic. We may conduct further hypothesis tests by evaluating the null hypothesis of no removable interaction $H_0: \theta = 0$ against the alternative $H_A: \theta \neq 0$. We may also test the null hypothesis of no non-removable interaction $H_0: e_{ij} = 0$ for all $i = 1, 2, \dots, L_1-1$ and $j = 1, 2, \dots, L_2-1$ against $H_A: e_{ij} = 0$ for at least one (i,j) . [Test statistics for evaluating removable and non-removable interactions are given in the Online Supplementary Material.] The results of these hypothesis tests can be used to determine the specific scenario and estimate the log odds summaries using a suitable model.

While such preliminary testing procedures can be useful for selecting a method to estimate the log odds summaries, they rely on the choice of a significance level, can lead to inflated type I errors, and the parameter estimates may have poor precision [20,21]. Therefore, in this paper we develop a Bayes estimator of log odds by accounting for potential removable interaction, but without the need for conducting preliminary hypothesis tests. Before describing our proposed Bayes estimator, we show that the Guerrero and Johnson link function [14] is a suitable transformation to additivity when the interactions under the logistic link function are removable.

The Guerrero and Johnson link function

The Guerrero and Johnson (GJ) link function, indexed by a single transformation parameter λ_G , and denoted as $g(p_{ij}, \lambda_G)$, is a Box-Cox transformation [22] of the disease odds, given by [14]:

$$g(p_{ij}, \lambda_G) = \begin{cases} \frac{1}{\lambda_G} \left\{ \left(\frac{p_{ij}}{1-p_{ij}} \right)^{\lambda_G} - 1 \right\} & \text{if } \lambda_G \neq 0 \\ \log \left(\frac{p_{ij}}{1-p_{ij}} \right) & \text{if } \lambda_G = 0 \end{cases} \quad (\text{Equation 4})$$

The logistic link is a member of the GJ family when $\lambda_G = 0$. When $\lambda_G \neq 0$, the GJ link is not symmetric (see Figure 1) since the disease risk approaches 1 (or 0) more rapidly than it approaches 0 (or 1). Further, the rate at which the disease risk approaches 1 (or 0) is higher

under the GJ link than the logistic link. The identity link function is not a member of the GJ family since there is no value of λ_G under which $g(p_{ij}, \lambda_G) = p_{ij}$.

An additive model under the GJ link is given by:

$$g(p_{ij}, \lambda_G) = \mu^* + \alpha_i^* + \beta_j^*, \quad (\text{Equation 5})$$

where μ^* is the baseline effect, and α_i^* and β_j^* are the main effects of the two risk factors on the GJ scale. In previous work [13] we have developed methods for obtaining MLEs of λ_G and the parameters of Equation (5). The following result establishes the GJ link function as an appropriate transformation to additivity when interactions under the logistic link are removable (proof given in the Online Supplementary Material).

Result

- i. *When the interaction between risk factors is removable under the logistic link function (i.e., Equation 2 holds with an equality sign instead of an approximation sign), there exists a link function, denoted $g(p_{ij})$, under which the model is additive; and this link function takes the form of the GJ link function given by Equation (4) under the boundary conditions $g(0) = -1/\lambda_G$ and $g(1) = 0$.*
- ii. *Conversely, whenever the model is additive under the GJ link function (i.e., Equations 4 and 5 apply) but $\lambda_G \neq 0$, and a quadratic polynomial in $\mu + \alpha_i + \beta_j$ is strictly monotonic over the domain of values of α_i and β_j that fit the range of the data at hand, the logistic link function yields a systematic departure from the additive model. Further, the logistic regression model is given by Equation (2) with $\theta = -\lambda_G$.*

When the model is additive under the GJ link (i.e., Equations 4 and 5 hold), the log odds can be written as:

$$\log \left(\frac{p_{ij}}{1-p_{ij}} \right) = \frac{1}{\lambda_G} \times \log \{ 1 + \lambda_G \times g(p_{ij}, \lambda_G) \} \quad (\text{Equation 6})$$

Taken together, these observations suggest that when the interactions are removable under the logistic link function, we can fit an additive model under the GJ link, and plug the MLEs of the parameters into the right hand side of Equation (6) to estimate the log odds. This is equivalent to fitting the logistic regression model given by Equation (2), obtaining the MLEs of this model, and plugging these into the right hand side of Equation (2) to estimate the log odds.

Bayes' estimator of log odds

Under scenarios 1 to 4 described above, the log odds can be modeled using Equation (1) or using an additive model under the GJ link (Equations 4, 5, and 6; equivalently, the logistic regression model of Equation 2). Let $\mathbf{Y}$, a vector of length $L_1 \times L_2$, denote the log odds.

Under the standard logistic regression model of Equation (1), a general form for the (i,j) -th element of $\mathbf{Y}$ is: $Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$. The MLE of $\mathbf{Y}$, denoted $\hat{\mathbf{Y}}_{MLE}$, can be obtained by using the MLEs of the main and interaction contrasts from a standard logistic regression model. When the design matrix of Equation (1) is of full rank, we have $E(\hat{\mathbf{Y}}_{MLE}) = \mathbf{Y}$ i.e., $\hat{\mathbf{Y}}_{MLE}$ is an unbiased estimate of $\mathbf{Y}$.

Let $\mathbf{Y}_{GJ}$ denote the log odds obtained via an additive model under the GJ link. From the results of the previous section, the (i,j) -th element of $\mathbf{Y}_{GJ}$ can be written as $\mu + \alpha_i + \beta_j + \theta\alpha_i\beta_j$. Let $\hat{\mathbf{Y}}_{GJ}$ denote the MLE of $\mathbf{Y}_{GJ}$. Note that the estimation of $\hat{\mathbf{Y}}_{MLE}$ involves estimating L_1L_2 parameters (baseline risk, L_1-1 and L_2-1 main effects contrasts for risk factors X and Z , respectively, and $(L_1-1)\times(L_2-1)$ interaction contrasts). However, the estimation of $\mathbf{Y}_{GJ}$ involves estimating only $L_1 + L_2$ parameters (baseline risk, L_1-1 and L_2-1 main effects contrasts for the risk factors X and Z , respectively, and the scalar parameter $\theta = -\lambda_G$ i.e., $(L_1-1)\times(L_2-1) - 1$ fewer parameters. As a result of this parsimony, $\hat{\mathbf{Y}}_{GJ}$ is likely to have a smaller standard error than $\hat{\mathbf{Y}}_{MLE}$.

Suppose there is no interaction at all so that the model is additive under the logistic link. This is equivalent to an additive model under the GJ link with $\lambda_G = 0$. Therefore, in principle, we may estimate the log odds using $\hat{\mathbf{Y}}_{GJ}$ either when the interaction between the risk factors is removable or when there is no interaction at all. Denote $E(\hat{\mathbf{Y}}_{GJ}) = \mathbf{Y}_{GJ}$. When all the interactions are removable, we have $E(\hat{\mathbf{Y}}_{GJ}) = \mathbf{Y}_{GJ} = \mathbf{Y}$. Otherwise, $\hat{\mathbf{Y}}_{GJ}$ will be biased, thereby offsetting any precision gains that may be attained by fitting a parsimonious model (i.e., a model with fewer parameters). In this case, $\mathbf{Y}$ should be estimated using $\hat{\mathbf{Y}}_{MLE}$.

In any practical setting, we do not know at the outset whether an interaction exists, and if so whether it is removable or non-removable. Here we propose a Bayes estimator for $\mathbf{Y}$ that does not require us to test for specific types of interactions, yet allows us to take advantage of model parsimony either when the interaction is removable or when there is no interaction at all. Using Equation (3) and the result from the previous section, we can write $\mathbf{Y}$ as:

$$\mathbf{Y} = \mathbf{Y}_{GJ} + \mathbf{e}. \quad (\text{Equation 7})$$

We know how to estimate $\mathbf{Y}_{GJ}$ via Equation (6) using an additive model under the GJ link function (equivalently, Equation 2). If we can estimate $\mathbf{e}$, we can plug this into Equation (7) to estimate $\mathbf{Y}$. We shall obtain an estimator, $\mathbf{e}_B$, so that a desirable criterion is optimized. We choose the squared error loss as our criterion, which is given by the quadratic form $(\mathbf{e} - \mathbf{e}_B)^T (\mathbf{e} - \mathbf{e}_B)$. Below we derive an admissible estimator of $\mathbf{e}$ by minimizing the risk function, which is the expected value of this loss function.

A crude estimate of $\mathbf{e}$ is given by $\hat{\mathbf{e}} = \hat{\mathbf{Y}}_{MLE} - \hat{\mathbf{Y}}_{GJ}$. Relying on the asymptotic normality of the MLEs, given $\mathbf{e}$, $\hat{\mathbf{e}}$ has an asymptotic $N(\mathbf{e}, \Sigma)$ distribution, where Σ can be estimated as $\hat{\Sigma} = \text{Var}(\hat{\mathbf{Y}}_{MLE} - \hat{\mathbf{Y}}_{GJ})$. Denote this normal distribution as $\pi(\hat{\mathbf{e}} | \mathbf{e})$ (we do not show the variance Σ in this notation). We shall identify admissible estimators $\mathbf{e}_B$ as functions of $\hat{\mathbf{e}}$.

The risk function is defined as the expected value of the loss function with respect to $\pi(\hat{\mathbf{e}} | \mathbf{e})$, and is given by $\int (\mathbf{e} - \mathbf{e}_B)^T (\mathbf{e} - \mathbf{e}_B) \pi(\hat{\mathbf{e}} | \mathbf{e}) d\hat{\mathbf{e}}$. This risk function is also known as the mean squared error (MSE). There are several classes of estimators that focus on minimizing this risk function. One such class consists of minimax estimators [23,24]. However, obtaining a minimax estimator is not straightforward in practical scenarios. Therefore, we propose to obtain a class of Bayes estimators such that certain minimax estimators are members of this class under limiting conditions.

Denote $\psi(\mathbf{e})$ as a prior probability density of $\mathbf{e}$. At this time we do not make any assumption about the specific form of $\psi(\cdot)$. Note that the likelihood function, which is closely related to $\pi(\hat{\mathbf{e}} | \mathbf{e})$, does not contain any information about $\psi(\mathbf{e})$. The posterior probability density of $\mathbf{e}$ given $\hat{\mathbf{e}}$, denoted $\pi(\mathbf{e} | \hat{\mathbf{e}})$, can be written as:

$$\pi(\mathbf{e} | \hat{\mathbf{e}}) = \frac{\pi(\hat{\mathbf{e}} | \mathbf{e}) \psi(\mathbf{e})}{\int \pi(\hat{\mathbf{e}} | \mathbf{e}) \psi(\mathbf{e}) d\mathbf{e}}.$$

Thus, $\int \pi(\mathbf{e} | \hat{\mathbf{e}}) d\mathbf{e} = 1$, regardless of the specific form of $\psi(\mathbf{e})$. We can now define the posterior risk as the expectation of the loss function with respect to the posterior density of $\mathbf{e}$, written as: $\int (\mathbf{e} - \mathbf{e}_B)^T (\mathbf{e} - \mathbf{e}_B) \pi(\mathbf{e} | \hat{\mathbf{e}}) d\mathbf{e}$. We shall identify Bayes estimators $\mathbf{e}_B$ that minimize the posterior risk. Clearly, these estimators depend upon both $\hat{\mathbf{e}}$ and the prior density $\psi(\cdot)$.

Taking the derivative of the posterior risk with respect to $\mathbf{e}_B$ and setting it equal to 0 shows that the $\mathbf{e}_B(\hat{\mathbf{e}}; \psi) = E_{\psi}(\mathbf{e} | \hat{\mathbf{e}})$, i.e., the posterior risk is minimized by the posterior mean of $\mathbf{e}$. We have used the notation $E_{\psi}(\cdot)$ to denote that the specific form of the posterior mean depends upon the specific form of the prior density $\psi(\cdot)$. Thus, $\mathbf{e}_B(\hat{\mathbf{e}}; \psi)$ is a class of Bayes estimators, and different members of this class can be obtained by specifying different prior densities $\psi(\cdot)$.

The risk function corresponding to this class of Bayes estimators is given by: $\int \{ \mathbf{e} - \hat{\mathbf{e}}_B(\hat{\mathbf{e}}; \psi) \}^T \{ \mathbf{e} - \hat{\mathbf{e}}_B(\hat{\mathbf{e}}; \psi) \} \pi(\hat{\mathbf{e}} | \mathbf{e}) d\hat{\mathbf{e}}$. If this risk is a constant for a particular $\psi(\cdot)$, then the corresponding Bayes estimator is also a minimax estimator [23]. Therefore, to identify an admissible estimator of $\mathbf{e}$, all we need to do is to identify a prior density $\psi(\cdot)$ under which the risk is a constant. Although it may not be always possible to identify such a prior density, we may be able to identify a $\psi(\cdot)$ that provides a Bayes estimator with constant risk under some limiting conditions.

Consider independent and identical $N(0, \sigma^2)$ prior distributions for the components of $\mathbf{e}$. Denote $\mathbf{I}$ as an identity matrix. Then it is easy to see that the posterior density $\pi(\mathbf{e} | \hat{\mathbf{e}})$ is normal with mean $E_{\psi}(\mathbf{e} | \hat{\mathbf{e}}) = \sigma^2 (\Sigma + \sigma^2 \mathbf{I})^{-1} \hat{\mathbf{e}}$ and variance $\sigma^2 \Sigma (\Sigma + \sigma^2 \mathbf{I})^{-1}$. Different choices of σ^2 will give different Bayes estimators under this prior. For a given σ^2 , the

posterior mean is unique. Denoting $\mathbf{K} = \sigma^2 (\Sigma + \sigma^2 \mathbf{I})^{-1} = \left(\frac{\Sigma}{\sigma^2} + \mathbf{I} \right)^{-1}$ and $\text{Trace}\{\cdot\}$ as the trace of a matrix, the risk function can be written as:

$$\begin{aligned}
& \int \{ \mathbf{e} - \hat{\mathbf{e}}_B(\hat{\mathbf{e}}; \psi) \}^\top \{ \mathbf{e} - \hat{\mathbf{e}}_B(\hat{\mathbf{e}}; \psi) \} \pi(\hat{\mathbf{e}} | \mathbf{e}) d\hat{\mathbf{e}} \\
&= \int \{ \mathbf{e} - \mathbf{K}\hat{\mathbf{e}} \}^\top \{ \mathbf{e} - \mathbf{K}\hat{\mathbf{e}} \} \pi(\hat{\mathbf{e}} | \mathbf{e}) d\hat{\mathbf{e}} \\
&= \mathbf{e}^\top \{ (\mathbf{I} - \mathbf{K})^\top (\mathbf{I} - \mathbf{K}) \} \mathbf{e} + \text{Trace} \{ \mathbf{K}^\top \mathbf{K} \Sigma \} \quad (\text{Equation 8})
\end{aligned}$$

The last step follows from linear algebra results for quadratic forms:

$$\begin{aligned}
E(\hat{\mathbf{e}}^\top \mathbf{K}^\top \mathbf{K} \hat{\mathbf{e}} | \mathbf{e}) &= E(\hat{\mathbf{e}}^\top | \mathbf{e}) \mathbf{K}^\top \mathbf{K} E(\hat{\mathbf{e}} | \mathbf{e}) + \text{Trace} \{ \mathbf{K}^\top \mathbf{K} \text{Var}(\hat{\mathbf{e}} | \mathbf{e}) \} \\
&= \mathbf{e}^\top \mathbf{K}^\top \mathbf{K} \mathbf{e} + \text{Trace} \{ \mathbf{K}^\top \mathbf{K} \Sigma \}.
\end{aligned}$$

When $\sigma^2 \rightarrow \infty$, we have $\mathbf{K} \rightarrow \mathbf{I}$ and the limit of the Bayes estimator $E_\psi(\mathbf{e} | \hat{\mathbf{e}})$ is $\hat{\mathbf{e}}$. It follows from Equation (8) that the risk function then approaches $\text{Trace}\{\Sigma\}$, which is a constant with respect to $\mathbf{e}$. Hence, $\hat{\mathbf{e}}$ is a minimax estimator of $\mathbf{e}$ when $\sigma^2 \rightarrow \infty$. When $\sigma^2 \rightarrow 0$, we have $\mathbf{K} \rightarrow \mathbf{0}$ and the limit of the Bayes estimator $E_\psi(\mathbf{e} | \hat{\mathbf{e}})$ is $\mathbf{0}$ and the risk function is $\mathbf{e}^\top \mathbf{e}$. Note that $\mathbf{e} \rightarrow \mathbf{0}$ since it has an *a priori* normal distribution with mean $\mathbf{0}$ and variance σ^2 that goes to 0. Thus, the risk function approaches $\mathbf{0}$, which is a constant. Therefore, $\mathbf{0}$ is a minimax estimator of $\mathbf{e}$ when $\sigma^2 \rightarrow 0$.

These observations show that an *a priori* normal distribution for $\mathbf{e}$ provides a class of estimators, given by $E_\psi(\mathbf{e} | \hat{\mathbf{e}}) = \sigma^2 (\Sigma + \sigma^2 \mathbf{I})^{-1} \hat{\mathbf{e}}$, that are minimax under the limiting conditions when $\sigma^2 \rightarrow \infty$ and $\sigma^2 \rightarrow 0$, respectively. Plugging this estimator into Equation (7), a Bayes estimator of $\mathbf{Y}$ from this class, denoted $\hat{\mathbf{Y}}_B$, is given by:

$$\begin{aligned}
\hat{\mathbf{Y}}_B &= \hat{\mathbf{Y}}_{GJ} + \hat{\mathbf{E}}(\mathbf{e} | \hat{\mathbf{e}}) \\
&= \hat{\mathbf{Y}}_{GJ} + \hat{\sigma}^2 (\hat{\Sigma} + \hat{\sigma}^2 \mathbf{I})^{-1} \hat{\mathbf{e}} \\
&= \hat{\mathbf{Y}}_{MLE} - \hat{\Sigma} (\hat{\Sigma} + \hat{\sigma}^2 \mathbf{I})^{-1} \hat{\mathbf{e}} \quad (\text{Equation 9})
\end{aligned}$$

The last step of Equation (9) follows from the identities $\hat{\mathbf{e}} = \hat{\mathbf{Y}}_{MLE} - \hat{\mathbf{Y}}_{GJ}$ and $\hat{\Sigma}(\hat{\Sigma} + \hat{\sigma}^2 \mathbf{I})^{-1} + \hat{\sigma}^2 (\hat{\Sigma} + \hat{\sigma}^2 \mathbf{I})^{-1} = \mathbf{I}$. Here we have used the notations $\hat{\sigma}^2$ and $\hat{\Sigma}$ to denote estimated values of

these variances. An empirical estimate of σ^2 can be obtained as $\hat{\sigma}^2 = \frac{\hat{\mathbf{e}}^\top \hat{\mathbf{e}}}{L_1 \times L_2}$. The calculation of $\hat{\Sigma}$ and an approximate formula for the variance of $\hat{\mathbf{Y}}_B$ are given in the Online Supplementary Material.

Properties of the proposed Bayes estimator—In order to obtain $\hat{\mathbf{Y}}_B$, it is not necessary to know whether an interaction exists or whether it is removable or non-removable. However, if there is an underlying interaction, Equation (9) suggests that $\hat{\mathbf{Y}}_B$ will have the following properties depending upon whether the interaction is removable or non-removable. When the interaction is removable, $\mathbf{e}$ will be negligible i.e., σ^2 will be near 0. Hence, $\hat{\mathbf{Y}}_B \approx \hat{\mathbf{Y}}_{GJ}$. Since $\mathbf{0}$ is a minimax estimator of $\mathbf{e}$ in this limiting case, $\hat{\mathbf{Y}}_{GJ}$ is a minimax estimator of $\mathbf{Y}$ when the interaction is removable. When the interaction is non-removable, there is the implication that the true values of the components of $\mathbf{e}$ are not equal to 0.

Equivalently, the prior variance σ^2 is not negligible. In the limiting case, as $\sigma^2 \rightarrow \infty$, we have $\hat{\mathbf{Y}}_B \approx \hat{\mathbf{Y}}_{MLE}$. Since $\hat{\mathbf{e}}$ is minimax in this limiting case, $\hat{\mathbf{Y}}_{MLE}$ is a minimax estimator of $\mathbf{Y}$ when the interaction is non-removable.

When $0 < \sigma^2 < \infty$, $\pi(\mathbf{e})$ is a proper prior i.e., $\int_{-\infty}^{\infty} \pi(\mathbf{e}) d\mathbf{e} = 1$. For a given σ^2 , the posterior mean, i.e., the Bayes estimator, is unique. Since a unique Bayes estimator under a proper prior is also admissible [24, Chapter 4.3], $\hat{\mathbf{Y}}_B$ is an admissible estimator when $0 < \sigma^2 < \infty$. However, since σ^2 is unknown, estimation of this parameter may impact the admissibility property of the Bayes estimator. In the limit when $\sigma^2 \rightarrow 0$ or $\sigma^2 \rightarrow \infty$, we have an improper prior. A Bayes estimator under an improper prior can be inadmissible. In fact, in the limiting cases we obtain minimax estimators that are admissible only when $(L_1-1) \times (L_2-1) \geq 2$ [24, Chapter 4.5]. In the following section, we examine the properties of the proposed Bayes estimator using simulations.

Simulation plan

We conducted simulation studies to examine the bias-variance trade-offs of $\hat{\mathbf{Y}}_{MLE}$, $\hat{\mathbf{Y}}_{GJ}$, and $\hat{\mathbf{Y}}_B$. We independently sampled case-control data with two ordinal risk factors, X with L_1 levels and Z with L_2 levels. We assumed a threshold model for disease risk in the following sense: there are thresholds C_1 and C_2 for the levels of X and Z , respectively, such that X and Z confer disease risk only if their levels exceed these values. Thus, the disease probability of each individual was assumed to follow a logistic regression model given by:

$$\text{logit} \{P(\text{disease}|X, Z)\} = \delta_0 + \delta_1 \times I(X \geq C_1) + \delta_2 \times I(Z \geq C_2) + \delta_{12} \times I(X \geq C_1, Z \geq C_2),$$

(Equation 10)

where $I(\cdot)$ is the indicator function. The threshold type of model for generating disease risk is motivated by several practical scenarios where public health messages about disease risk are delivered based on thresholds or cutpoints for risk factors – for example, although every unit increase in body mass index (BMI) may contribute to an increase in the risk of endometrial cancer, public health messages about risk are generally quoted for categories such as normal, overweight and obese, based on relevant cutpoints for BMI.

We considered the following three settings:

1. 2×3 table: $L_1 = 2$ and $L_2 = 3$, with $C_1 = 2$ and $C_2 = 3$;
2. 2×5 table: $L_1 = 2$ and $L_2 = 5$, with $C_1 = 2$ and $C_2 = 4$; and
3. 5×5 table: $L_1 = 5$ and $L_2 = 5$, with $C_1 = 4$ and $C_2 = 4$.

For the risk factor prevalence, we assumed $P(X \geq C_1) = 0.10 = P(Z \geq C_2)$. Given C_1 , the specific levels of X were simulated with probability $P(X=C_1) = P(X=C_1+1) = \dots = P(X=L_1) = 0.10/(L_1-C_1+1)$, and $P(X=1) = P(X=2) = \dots = P(X=C_1-1) = 0.90/(C_1-1)$. The levels of Z

were obtained in a similar manner. We considered the magnitude of δ_0 to be such that the disease prevalence was 0.10.

In previous work we showed that, given δ_1 and δ_2 , an interaction is removable or non-removable when δ_{12} falls in a certain range [13]. In particular, the data contain:

- i. only removable interaction when $\delta_{12} = \max\{-\delta_1, -\delta_2\}$;
- ii. only non-removable interaction when $\delta_{12} = \min\{-\delta_1, -\delta_2\}$; and
- iii. some removable and some non-removable interactions when $\min\{-\delta_1, -\delta_2\} < \delta_{12} < \max\{-\delta_1, -\delta_2\}$.

Another type of interaction, known as compositional epistasis, occurs when the effect of a genetic marker at one locus is masked by a variant at another locus [25]. Thus, under compositional epistasis, we have $\delta_1 = 0 = \delta_2$ and $\delta_{12} \neq 0$. We simulated data under the following parametric configurations:

1. Removable interaction: $\delta_1 = \delta_2 = \log(2) = 0.6931$, $\delta_{12} = \{\log(1.1), \log(1.3), \log(1.5), \log(1.7), \log(1.9), \log(2), \log(2.5), \log(3)\}$ i.e., $\delta_{12} = \max\{-\delta_1, -\delta_2\}$;
2. Non-removable interaction: $\delta_1 = \delta_2 = \log(2) = 0.6931$, $\delta_{12} = \{-\log(2), -\log(2.1), -\log(2.3), -\log(2.5), -\log(2.7), -\log(2.9), -\log(3), -\log(5), -\log(7)\}$ i.e., $\delta_{12} = \min\{-\delta_1, -\delta_2\}$;
3. Some removable and some non-removable interactions: $\delta_1 = \log(2)$, $\delta_2 = 0$, and $\min\{-\delta_1, -\delta_2\} < \delta_{12} < \max\{-\delta_1, -\delta_2\}$ i.e., $\delta_{12} = \{-\log(1.25), -\log(1.5), -\log(1.75), -\log(2), -\log(2.25), -\log(2.5), -\log(2.75)\}$;
4. Compositional epistasis: $\delta_1 = \delta_2 = 0$, $\delta_{12} = \{\log(1.5), \log(2), \log(2.5), \log(3), \log(5), \log(7)\}$.

We generated 1000 case-control data sets, each consisting of 1000 cases and 1000 controls, under each parametric configuration using the true model given by Equation (10). When analyzing the data sets, we assumed that we did not know the true model. To estimate $\hat{\mathbf{Y}}_B$, we first estimated $\hat{\mathbf{Y}}_{MLE}$ using Equation (1) with L_1-1 and L_2-1 main effects contrasts for X and Z, respectively, and $(L_1-1) \times (L_2-1)$ interaction contrasts. Next, we estimated $\hat{\mathbf{Y}}_{GJ}$ using Equation (6) by fitting an additive model under the GJ link with L_1-1 and L_2-1 main effects contrasts for X and Z, respectively, and a scalar $\theta = -\lambda_G$ to represent the transformation parameter. Finally, we estimated $\hat{\mathbf{Y}}_B$ using Equation (9). We calculated the variances of these estimates and their MSEs as the sum of the variances of the log odds and the squared difference between the true (i.e., observed) and the estimated log odds, calculated for each risk factor sub-class and averaged over all the sub-classes. We calculated the root mean squared errors (RMSEs) as the square root of the MSEs. We summarized the interquartile range (IQR) and average of the RMSEs over the 1000 simulated data sets under each parametric configuration.

RESULTS

Simulation

The simulation results are illustrated in Figures 2 and 3. When the interaction was removable (i.e., there was no non-removable interaction), as expected $\hat{Y}_{GJ}$ had smaller RMSE than $\hat{Y}_{MLE}$ (Column 1 of Figure 2). When the interaction was non-removable (i.e., there was no removable interaction), as expected $\hat{Y}_{MLE}$ had smaller RMSE than $\hat{Y}_{GJ}$ (Column 2 of Figure 2). However, for large contingency tables (for example, a 5×5 table), the RMSE of $\hat{Y}_{MLE}$ was larger than that of $\hat{Y}_{GJ}$ for interaction effects of small magnitude. This is because obtaining $\hat{Y}_{MLE}$ required estimation of 16 parameters even when the magnitude of the interaction effect was negligible, resulting in loss of efficiency. However, the RMSE of $\hat{Y}_{GJ}$ increased rapidly and was close to that of $\hat{Y}_{MLE}$ as the magnitude of the interaction effect increased. When the data contained some removable and some non-removable interactions, $\hat{Y}_{GJ}$ had smaller RMSE than $\hat{Y}_{STD}$ (Column 3 of Figure 2). Under compositional epistasis, $\hat{Y}_{GJ}$ had higher RMSE than $\hat{Y}_{MLE}$, particularly when the magnitude of the interaction effect (δ_{12}) was large. In contrast, $\hat{Y}_B$ had the smallest RMSE, for the most part, under all the scenarios, demonstrating the remarkable bias-variance trade off attained by the Bayes estimation method. The trade off was best realized for larger contingency tables, i.e., for larger values of $(L_1-1) \times (L_2-1)$.

Figure 2 also illustrates the admissibility properties of the various estimators. Consider the first row of Figure 2, which corresponds to 2×3 contingency tables i.e., $(L_1-1) \times (L_2-1) = 2$. We noted earlier that, in this case, $\hat{Y}_{GJ}$ is minimax and admissible when the interaction is removable, and $\hat{Y}_{MLE}$ is minimax and admissible when the interaction is non-removable. Thus, as expected, $\hat{Y}_{GJ}$ and $\hat{Y}_{MLE}$ had smaller MSE than $\hat{Y}_B$ when the interactions were removable and non-removable, respectively (Columns 1 and 2 in Row 1 of Figure 2). There was only a modest increase in the RMSE of $\hat{Y}_B$, with the difference between the RMSEs of $\hat{Y}_B$ and the minimax estimators falling between 0.001 and 0.028. For large contingency tables (for example, 5×5 tables), $\hat{Y}_B$ had the smallest MSE under a wide range of parametric configurations considered. For example, in a 5×5 contingency table with removable interactions, the RMSEs of $\hat{Y}_{MLE}$, $\hat{Y}_{GJ}$, and $\hat{Y}_B$ were in the range 0.60–0.62, 0.52–0.54, and 0.50–0.51, respectively, under various values of δ_{12} considered in our simulations.

The interquartile ranges (IQRs) are shown in Figure 3. The RMSE of $\hat{Y}_{GJ}$ had the largest IQR and that of $\hat{Y}_{MLE}$ had the smallest IQR under all the scenarios considered, suggesting that the empirical distribution of the RMSE of $\hat{Y}_{GJ}$ has a wider spread relative to the distributions of the RMSEs of $\hat{Y}_{MLE}$ and $\hat{Y}_B$. The RMSE of $\hat{Y}_B$ had IQRs closer to those of the RMSE of $\hat{Y}_{MLE}$. For example, in a 5×5 contingency table with removable interactions, the IQRs of the RMSEs of $\hat{Y}_{MLE}$, $\hat{Y}_{GJ}$, and $\hat{Y}_B$ were in the range 0.04–0.05, 0.10–0.11, and 0.06–0.07, respectively.

In summary, these results suggest that our proposed Bayes estimator is a useful approach for estimating the log odds summaries, regardless of the type of interaction and regardless of the size of the contingency table.

Data applications – three case-control studies of endometrial cancer

Application 1: BMI, CYP19A1, and endometrial cancer—A case-control study within the Epidemiology of Endometrial Cancer Consortium [17] reported a significant interaction between SNP rs727479 in the *CYP19A1* gene (two levels: CC, and AC or AA genotypes) and body mass index (BMI; three levels: normal, overweight and obese) among post-menopausal women (age ≥ 55 years) based on a logistic regression analysis. These data are shown in Table 1.

For illustrative purposes, it will be useful to understand the properties of this interaction, although our proposed Bayes estimation method does not require preliminary tests for removable and non-removable interactions. Using the test statistics outlined in the Online Supplementary Material, there was significant evidence for removable and non-removable interactions between BMI and *CYP19A1* at 5% significance level for each test (test statistic for removable interaction = 7.64, degrees of freedom = 1, p-value = 0.006; test statistic for non-removable interaction = 4.85, degrees of freedom = 1, p-value = 0.03). In the AC/AA genotype group, the observed log odds increased with increasing level of BMI. Although an increasing trend occurred in the CC genotype group, the normal and overweight individuals with CC genotype had fairly similar log odds. Further, the observed log odds were higher for the AC/AA genotypes relative to the CC genotypes among the overweight and obese individuals, but not among those with normal BMI. These observations suggest a lack of strict monotonicity, which is consistent with the presence of a non-removable interaction.

The estimated log odds are shown in Table 1 [See Supplementary Figure 1 for a visual representation of these estimates]. For these data, the model based on Equation (1) is a saturated model since we do not have any additional risk factors for consideration in these analyses. Therefore, $\hat{Y}_{MLE}$ is the same as the observed log odds. For $\hat{Y}_{GJ}$, the estimated log odds for the normal and overweight BMI categories in the CC genotype group were not close to the observed values, possibly due to the non-monotonic interaction. The Bayes estimates $\hat{Y}_B$ were closer to $\hat{Y}_{MLE}$, demonstrating shrinkage towards estimates based on the standard logistic regression. It is of interest that the standard errors of $\hat{Y}_{GJ}$ were smaller than those of $\hat{Y}_{MLE}$, but the RMSE is a better indicator of estimate variability. All three estimators had fairly similar RMSE (0.098 for $\hat{Y}_{MLE}$, 0.106 for $\hat{Y}_{GJ}$ and 0.102 for $\hat{Y}_B$) which is consistent with the results of the simulation study that shows that when monotonicity is not strict, as in compositional epistasis (Column 4 in Row 1 of Figure 2), the RMSEs of $\hat{Y}_{MLE}$ and $\hat{Y}_B$ are similar and are, for the most part, slightly smaller in magnitude than that of $\hat{Y}_{GJ}$.

Application 2: BMI, diabetes, and endometrial cancer—This case-control study [18] reported a significant interaction between BMI (three levels: normal, overweight, and obese) and diabetes (two levels: present and absent). The log odds increased with increasing level of BMI both in the absence and presence of diabetes (Table 2). Further, the log odds were higher in the presence of diabetes, regardless of the level of BMI. These observations illustrate strictly monotonic properties of the data. At 5% significance levels, there was significant evidence for a removable interaction (test statistic = 4.46, degrees of freedom = 1, p-value = 0.035) and no significant evidence for a non-removable interaction (test statistic =

0.280, degrees of freedom = 1, p-value = 0.60). $\hat{\mathbf{Y}}_{\text{GJ}}$ had smaller standard errors and smaller RMSE than $\hat{\mathbf{Y}}_{\text{MLE}}$ (RMSE = 0.153 and 0.199 for $\hat{\mathbf{Y}}_{\text{GJ}}$ and $\hat{\mathbf{Y}}_{\text{MLE}}$, respectively). In the presence of diabetes, the standard errors of $\hat{\mathbf{Y}}_{\text{B}}$ for different levels of BMI were between those of $\hat{\mathbf{Y}}_{\text{MLE}}$ and $\hat{\mathbf{Y}}_{\text{GJ}}$. In the absence of diabetes, the standard errors of all the estimators for different levels of BMI were small. The RMSE of $\hat{\mathbf{Y}}_{\text{B}}$ was 0.183, which lies between the RMSEs of $\hat{\mathbf{Y}}_{\text{GJ}}$ and $\hat{\mathbf{Y}}_{\text{MLE}}$. This is also consistent with the results of the simulation study (Column 1 in Row 1 of Figure 2), which demonstrated that $\hat{\mathbf{Y}}_{\text{GJ}}$, a minimax estimator when the interaction is removable and there is no non-removable interaction, is also admissible when $(L_1-1) \times (L_2-1) = 2$, but $\hat{\mathbf{Y}}_{\text{B}}$ is still a useful estimator in the sense that its RMSE is smaller than that of $\hat{\mathbf{Y}}_{\text{MLE}}$. [See Supplementary Figure 2 for a visual representation of $\hat{\mathbf{Y}}_{\text{MLE}}$, $\hat{\mathbf{Y}}_{\text{GJ}}$, and $\hat{\mathbf{Y}}_{\text{B}}$.]

Application 3: Tea intake, CYP19A1, and endometrial cancer—This case-control study, the Shanghai Endometrial Cancer Study [19], reported a significant interaction between tea intake (two levels: low and high intake) and *CYP19A1* genotype based on the SNP rs1065779 (three levels: GG, GT, and TT genotypes). The observed log odds decreased monotonically with increasing number of the T allele among those with high tea intake (Table 3). However, there was no monotonic trend among individuals with low tea intake, suggesting that the interaction was not removable (test statistic for removable interaction = 1.02, degrees of freedom = 1, p-value = 0.31). Indeed the data contained evidence for significant non-monotonic interaction at 5% significance level (test statistic = 10.65, degrees of freedom = 1, p-value = 0.001). Hence, a transformation to additivity is not appropriate for these data. As expected, $\hat{\mathbf{Y}}_{\text{B}}$ was closer to the observed log odds than $\hat{\mathbf{Y}}_{\text{GJ}}$, which also illustrates shrinkage towards $\hat{\mathbf{Y}}_{\text{MLE}}$ due to the presence of significant non-monotonic interaction. Further, $\hat{\mathbf{Y}}_{\text{B}}$ had RMSE comparable to that of $\hat{\mathbf{Y}}_{\text{MLE}}$ (RMSE = 0.129, 0.178, and 0.136 for $\hat{\mathbf{Y}}_{\text{MLE}}$, $\hat{\mathbf{Y}}_{\text{GJ}}$, and $\hat{\mathbf{Y}}_{\text{B}}$, respectively). This is also consistent with the results of the simulation study (Column 2 in Row 1 of Figure 2), which showed that the minimax estimator $\hat{\mathbf{Y}}_{\text{MLE}}$ is also admissible when $(L_1-1) \times (L_2-1) = 2$. [See Supplementary Figure 3 for a visual representation of $\hat{\mathbf{Y}}_{\text{MLE}}$, $\hat{\mathbf{Y}}_{\text{GJ}}$, and $\hat{\mathbf{Y}}_{\text{B}}$.]

COMPUTATIONAL GUIDANCE FOR PRACTITIONERS

In this paper we have developed a Bayes estimator for log odds, which can be calculated using the following steps:

1. Obtain $\hat{\mathbf{Y}}_{\text{MLE}}$ as the right hand side of Equation (1) by plugging in the MLEs of that model.
2. Obtain $\hat{\mathbf{Y}}_{\text{GJ}}$ as the right hand side of Equation (6) by plugging in the MLEs of an additive model under the GJ link function.
3. Obtain $\hat{\Sigma}$ and $\hat{\sigma}^2$ (formula given in Online Supplementary Material).
4. Obtain $\hat{\mathbf{Y}}_{\text{B}}$ using Equation (9).
5. Obtain their standard errors (formula given in Online Supplementary Material).

We have prepared a computer program to implement these methods using the R programming language (<http://cran.r-project.org>). This program, along with instructions for

use, can be downloaded freely from the first author's academic web page: (<https://www.mskcc.org/biostatistics/~satagopj>). This program takes as input the case-control status and the values of the two risk factors of interest. The output contains $\hat{Y}_{MLE}$, $\hat{Y}_{GJ}$, and $\hat{Y}_B$, their standard errors and RMSEs.

DISCUSSION

There is a large body of literature on estimating interaction effects and on conducting a hypothesis test for the presence of a significant interaction for binary traits [2–4]. In contrast, the emphasis of our work is on accurate and precise estimation of the log odds parameters. Our work is based on the thesis that, when there is removable interaction under the logistic link function, the model relating the binary disease trait and the risk factors is additive on some scale of risk. We have shown that the GJ link function is an appropriate scale for additivity. Under the GJ link function, disease risk increases (or decreases) at a higher rate than that under an additive logistic link function (Figure 1). When this happens, it means that interaction terms will be needed in a logistic regression model to obtain a better fit to the data. In contrast, disease risk may be characterized accurately and more parsimoniously using an additive model under the GJ link. This would provide a better fitting model, and would facilitate accurate and more precise estimation of epidemiologic parameters such as log odds, especially in the extremes of the risk distribution, using fewer parameters. To attain this, we have developed a Bayes estimator that exploits model parsimony while simultaneously accounting for potential non-removable interactions. Our simulations show that this method has remarkable bias-variance trade-off under a wide range of parametric configurations and is, hence, a valid method for use in practical settings.

All our empirical examples are case-control data from 2×3 contingency tables. Even in this small contingency table setting, the proposed Bayes estimation approach has good bias-variance trade-off. This can also be observed in the second example [18], where the data exhibit strict monotonicity properties, suggesting that the interaction between BMI and diabetes in this study is removable. When diabetes is present, the sample size in this data set is modest for all three levels of BMI. Even in this setting, the RMSE of the Bayes approach is intermediate between those of the additive GJ model and the standard logistic model, though considerable reduction in RMSE would be attained for larger contingency tables as seen in our simulations.

The topic of interaction continues to garner much attention in epidemiology. There is a long-standing debate as to whether interactions should be examined under a logistic link (referred to in epidemiology as multiplicative interaction) or under an identity link, i.e., on the scale of disease risk (referred to as additive interaction), since the latter is anticipated to be more relevant from a public health perspective [8]. To some extent, this has also led to an anticipation that additive interactions can help understand biological interactions underpinning disease risk. However, biological interactions can occur regardless of the presence of a statistical interaction [12, 26]. Further, the absence of a statistical interaction on one scale of the outcome may imply its presence on another scale. Therefore, finding a parsimonious model, even if it is not on the logistic or the identity scale, should be useful for

obtaining practically relevant insights about the risk factors in relation to disease etiology [27, 28].

We set out to obtain an admissible estimator of log odds by minimizing the expected value of the squared error loss function, i.e., the risk function. This would provide minimax estimators, but minimizing the risk function is not straightforward. Therefore, we minimized the posterior risk function to obtain Bayes estimators. Our efforts to obtain a Bayes estimator that has a constant risk in the limit resulted in a normal prior distribution for the parameters $\mathbf{e}$. This turns out to be a conjugate prior when the MLE $\hat{\mathbf{e}}$ has a normal distribution, as expected asymptotically. This choice of a normal prior is not motivated by its property of being a conjugate prior. Instead, it is a natural choice of prior to obtain a class of Bayes estimators that include minimax estimators as its members in the limit [23, 24]. It is not necessary to know what the parameters of this normal distribution are, and we estimate the prior variance empirically from our observed data. Although our proposed Bayes estimator is based on the squared error loss function, other criteria (for example, the absolute difference between $\mathbf{e}$ and $\mathbf{e}_B$) may also be considered. A comprehensive evaluation of other optimality criteria is outside the scope of this paper.

This methodology can be extended to accommodate models adjusting for additional variables. We briefly outline this here. If we are solely interested in the main effect of an additional variable, but not in its interaction with the other risk factors, then the main effect, denoted δ_k , for this additional variable can be added to the right hand sides of Equations (1) and (2). In other words, if the interaction between the risk factors of interest is removable, then the effects of these risk factors can be represented parsimoniously as the logarithm of their additive effects, as shown in the right hand side of Equation (6) using arguments pertaining to transformation to additivity. The main effect of the new variable can be added to the right hand side of Equation (6). Further approximation would lead to Equation (2). However, suppose we are also interested in the interaction between this new variable and the other risk factors. If all the pair-wise interactions are removable, then the corresponding interaction effects may be written parsimoniously as $\theta\alpha_i\beta_j$, $\theta\alpha_i\delta_k$, $\theta\beta_j\delta_k$ etc. A comprehensive evaluation of the operating characteristics of these extensions, including evaluation of higher order interactions, will be pursued elsewhere.

Whether the precision gains of our proposed Bayes estimator of log odds leads to better risk prediction remains an open question. Further research is needed to measure and evaluate the discriminatory or predictive performance of our proposed approach, and to quantify the statistical significance of the improvements in the performance. Another important use of the log odds parameters (equivalently, odds ratios) is for projecting the benefits of interventions in screening studies. Further research is also needed to evaluate the accuracy of the projected benefits of interventions based on the log odds estimated via our proposed method.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding Acknowledgment:

We are grateful to the reviewers for their insightful comments that helped improve this manuscript. Satagopan and Olson were supported by research grants R01CA137420 and R01CA83918, respectively, from the National Cancer Institute, USA, Cancer Center Support Grant P30CA008748, and grant UL1RR024996 from the Clinical and Translational Science Center at Weill Cornell Medical College, New York, USA. Elston's work was supported by a grant from the National Research Foundation of Korea 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 Center for Advancing Translational Sciences (NCATS).

References

1. Gail MH. Personalized estimates of breast cancer risk in clinical practice and public health. *Stat Med.* 2011; 30(10):1090–1104. [PubMed: 21337591]
2. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet.* 2009; 10(6):392–404. [PubMed: 19434077]
3. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am J Hum Genet.* 2010; 86(1):6–22. [PubMed: 20074509]
4. Thomas DC. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annu Rev Publ Health.* 2010; 21:21–36.
5. Garnett MJ, Edelman EJ, Heidorn SJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature.* 2012; 483(7391):570–575. [PubMed: 22460902]
6. Masica DL, Karchin R. Collections of simultaneously altered genes as biomarkers of cancer cell drug response. *Cancer Res.* 2013; doi: 10.1158/0008-5472.CAN-12-3122
7. Wang X, Elston RC, Zhu X. The meaning of interaction. *Hum Hered.* 2010; 70(14):269–277. [PubMed: 21150212]
8. Rothman, KJ., Greenland, S., Lash, TL. *Modern Epidemiology.* 3. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
9. Finney DJ. Main effects and interactions. *J Am Stat Assoc.* 1948; 43(244):566–571. [PubMed: 18107568]
10. Scheffé, H. *The Analysis of Variance.* New York, NY: Wiley; 1959.
11. Elston RC. On additivity in the analysis of variance. *Biometrics.* 1961; 17(2):209–219.
12. Wang X, Elston RC, Zhu X. Statistical interaction in human genetics: how should we model it if we are looking for biological interaction? *Nat Rev Genet.* 2011; 12(1):74. [PubMed: 21102529]
13. Satagopan JM, Elston RC. Evaluation of removable statistical interaction for binary traits. *Stat Med.* 2013; 32(7):1164–1190. [PubMed: 23018341]
14. Guerrero VM, Johnson RA. Use of the Box-Cox transformation with binary response models. *Biometrika.* 1982; 69(2):309–314.
15. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics.* 1985; 41(2):361–372. [PubMed: 4027319]
16. Piantadosi S, Gail MH. A comparison of the power of two tests for qualitative interactions. *Stat Med.* 12(13):1239–1248. [PubMed: 8210823]
17. Setiawan VW, Doherty JA, Shu XO, et al. Two estrogen-related variants in CYP19A1 and endometrial cancer risk: a pooled analysis in the Epidemiology of Endometrial Cancer Consortium. *Cancer Epidemiol Biomarkers Prev.* 2009; 18(1):242–247. [PubMed: 19124504]
18. Shoff SM, Newcomb PA. Diabetes, body size, and risk of endometrial cancer. *Am J Epidemiol.* 1998; 148(3):234–240. [PubMed: 9690359]
19. 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. *Am J Epidemiol.* 2007; 166(12):1420–1430. [PubMed: 17827443]
20. Sen PK. On the asymptotic distributional risks of shrinkage and preliminary test versions of maximum likelihood estimators. *Sankhya Ser A.* 1986; 48(3):354–371.

21. Satagopan JM, Zhou Q, Oliveria SA, Duszka SW, Weinstock MA, Berwick M, Halpern AC. Properties of preliminary test estimators and shrinkage estimators for evaluating multiple exposures - Application to questionnaire data from the SONIC study. *J R Stat Soc Ser C Appl Stat.* 2011; 60(4):619–632.
22. Box GEP, Cox DR. An analysis of transformations. *J R Stat Soc Ser B Stat Methodol.* 1964; 26(2): 211–252.
23. Wald A. Contributions to the theory of statistical estimation and testing hypotheses. *Ann Math Stat.* 1939; 10(4):299–326.
24. Lehmann, EL. *Theory of Point Estimation.* New York: Wiley; 1983.
25. Cordell HJ. Epistasis – what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet.* 2002; 11(20):2463–2468. [PubMed: 12351582]
26. Siemiatycki J, Thomas DC. Biological models and statistical interactions: an example from multistage carcinogenesis. *Int J Epidemiol.* 1981; 10(4):383–387. [PubMed: 7327838]
27. Weinberg CR. Interaction and exposure modification: are we asking the right questions? *Am J Epidemiol.* 2012; 175(7):602–605. [PubMed: 22306562]
28. Weinberg CR. Commentary: Thoughts on assessing evidence for gene by environment interaction. *Int J Epidemiol.* 2012; 41(3):705–707. [PubMed: 22596932]

Figure 1.

The shape of the Guerrero and Johnson (GJ) link function and the logistic link function. The link is shown for $\lambda_G = -1.5$ and 1.5 . The logistic link corresponds to $\lambda_G = 0$. The horizontal axis shows the additive effect (say, A). The vertical axis plots the disease risk, calculated as $B/(1+B)$, where $B = \exp(A)$ for the logistic link, and otherwise $B = (1 + \lambda_G A)^{1/\lambda_G}$.

Figure 2.

Simulation results showing the comparison of mean squared errors (MSE) of the three estimators $\hat{Y}_{STD}$ (standard logistic; black line with "S"), $\hat{Y}_{GJ}$ (additive GJ; purple line with "G"), and $\hat{Y}_B$ (Bayes; red line with "B") for data simulated under removable interactions (Column 1), non-removable interactions (Column 2), some removable and some non-removable interactions (Column 3), and compositional epistasis (Column 4), with the risk factors based on 2x3 (Row 1), 2x5 (Row 2), and 5x5 (Row 3) contingency tables.

Figure 3.

Simulation results of the interquartile ranges (IQRs) of the MSEs of $\hat{Y}_{STD}$ (standard logistic; black line with "S"), $\hat{Y}_{GJ}$ (additive GJ; purple line with "G"), and $\hat{Y}_B$ (Bayes; dashed and red line with "B") for data simulated under removable interactions (Column 1), non-removable interactions (Column 2), some removable and some non-removable interactions (Column 3), and compositional epistasis (Column 4), with the risk factors based on 2×3 (Row 1), 2×5 (Row 2), and 5×5 (Row 3) contingency tables.

Results for the Setiawan et al [17] data

Table 1

Columns 1 and 2 indicate the risk factor classes and jointly identify the sub-classes. Columns 3 and 4 provide the number of cases and controls. The estimated log odds and their standard errors (in parentheses) corresponding to $\hat{\Psi}_{STD}$, $\hat{\Psi}_{GJ}$, and $\hat{\Psi}_B$ are given in Columns 5, 6, and 7, respectively. The last row shows the root mean squared errors of the three estimation methods. For the additive model under the GJ link, the transformation parameter is $\hat{\lambda}_G = -2.91$.

BMI	rs727479	Cases	Controls	$\hat{\Psi}_{STD}$ (SE)	$\hat{\Psi}_{GJ}$ (SE)	$\hat{\Psi}_B$ (SE)
Normal	CC	143	328	-0.8302 (0.1002)	-0.9339 (0.0505)	-0.8575 (0.1010)
Overweight	CC	78	175	-0.8081 (0.1361)	-0.6443 (0.0396)	-0.7613 (0.1284)
Obese	CC	72	101	-0.3385 (0.1542)	-0.3362 (0.1411)	-0.3579 (0.1531)
Normal	AC/AA	1004	2475	-0.9022 (0.0374)	-0.8848 (0.0343)	-0.8749 (0.0380)
Overweight	AC/AA	874	1456	-0.5104 (0.0428)	-0.5174 (0.0425)	-0.5572 (0.0426)
Obese	AC/AA	881	758	0.1504 (0.0495)	0.1503 (0.0495)	0.1699 (0.0496)
Root Mean Squared Error				0.098	0.106	0.102

Table 2

Results for the Shoff and Newcomb [18] data

Columns 1 and 2 indicate the risk factor classes. Columns 3 and 4 provide the number of cases and controls. The estimated log odds and their standard errors (in parentheses) corresponding to $\hat{\mathbf{Y}}_{STD}$, $\hat{\mathbf{Y}}_{GJ}$, and $\hat{\mathbf{Y}}_B$ are given in Columns 5, 6, and 7, respectively. The last row shows the root mean squared errors of the three estimation methods. For the additive model under the GJ link, the transformation parameter is $\hat{\lambda}_G = -1.07$.

BMI	Diabetes	Cases	Controls	$\hat{\mathbf{Y}}_{STD}$ (SE)	$\hat{\mathbf{Y}}_{GJ}$ (SE)	$\hat{\mathbf{Y}}_B$ (SE)
Normal	Absent	373	1633	-1.4766 (0.0574)	-1.4827 (0.0558)	-1.4985 (0.0650)
Overweight	Absent	85	262	-1.1257 (0.1248)	-1.0917 (0.1053)	-1.0634 (0.1090)
Obese	Absent	178	253	-0.3516 (0.0978)	-0.3555 (0.0955)	-0.3921 (0.1061)
Normal	Present	20	81	-1.3987 (0.2497)	-1.2957 (0.0722)	-1.3768 (0.2471)
Overweight	Present	16	31	-0.6614 (0.3078)	-0.7912 (0.1826)	-0.7237 (0.2397)
Obese	Present	51	31	0.4978 (0.2277)	0.5063 (0.2232)	0.5383 (0.2264)
Root Mean Squared Error				0.199	0.153	0.187

Results for the Xu et al [19] data

Table 3

Columns 1 and 2 indicate the risk factor classes. Columns 3 and 4 provide the number of cases and controls. The estimated log odds and their standard errors (in parentheses) corresponding to $\hat{\mathbf{Y}}_{STD}$, $\hat{\mathbf{Y}}_{GJ}$, and $\hat{\mathbf{Y}}_B$ are given in Columns 5, 6, and 7, respectively. The last row shows the root mean squared errors of the three estimation methods. For the additive model under the GJ link, the transformation parameter is $\hat{\lambda}_G = 0.01$.

rs1065779	Tea intake	Cases	Controls	$\hat{\mathbf{Y}}_{STD}$ (SE)	$\hat{\mathbf{Y}}_{GJ}$ (SE)	$\hat{\mathbf{Y}}_B$ (SE)
GG	Low	211	226	-0.0687 (0.0957)	0.0636 (0.0847)	-0.0205 (0.0994)
GT	Low	382	322	0.1709 (0.0757)	0.0981 (0.0693)	0.1144 (0.0759)
TT	Low	126	153	-0.1942 (0.1203)	-0.2193 (0.1056)	-0.1858 (0.1194)
GG	High	117	90	0.2624 (0.1402)	-0.0196 (0.1019)	0.2142 (0.1452)
GT	High	148	171	-0.1445 (0.1123)	0.0149 (0.0909)	-0.0880 (0.1117)
TT	High	45	65	-0.3677 (0.1939)	-0.3027 (0.1235)	-0.3760 (0.1925)
Root Mean Squared Error				0.129	0.178	0.136