Published in final edited form as:

Stat Med. 2017 November 10; 36(25): 4050–4060. doi:10.1002/sim.7405.

A comparison of statistical methods for the study of etiologic heterogeneity

Emily C. Zabor^{a,b,*} and Colin B. Begg^a

^aDepartment of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 485 Lexington Ave, 2nd floor, New York, NY 10017, USA

^bDepartment of Biostatistics, Mailman School of Public Health, Columbia University, 722 W 168th St, New York, NY 10032, USA

Abstract

Cancer epidemiologic research has traditionally been guided by the premise that certain diseases share an underlying etiology, or cause. However, with the rise of molecular and genomic profiling, attention has increasingly focused on identifying subtypes of disease. As subtypes are identified, it is natural to ask the question of whether they share a common etiology or in fact arise from distinct sets of risk factors. In this context, epidemiologic questions of interest include 1) whether a risk factor of interest has the same effect across all subtypes of disease and 2) whether risk factor effects differ across levels of each individual tumor marker of which the subtypes are comprised. A number of statistical models have been proposed to address these questions. In an effort to determine the similarities and differences among the proposed methods, and to identify any advantages or disadvantages, we employ a simplified data example to elucidate the interpretation of model parameters and available hypothesis tests, and we perform a simulation study to assess bias in effect size, type I error, and power. The results show that when the number of tumor markers is small enough that the cross-classification of markers can be evaluated in the traditional polytomous logistic regression framework, then the statistical properties are at least as good as the more complex modeling approaches that have been proposed. The potential advantage of more complex methods is in the ability to accommodate multiple tumor markers in a model of reduced parametric dimension.

Keywords

cancer epidemiology; etiologic heterogeneity; disease subtypes

1. Introduction

The basic goal of most epidemiologic research is to investigate the prevalence and causes of disease. Traditionally, epidemiologists have organized this line of research under the premise that patients with a certain disease share an underlying etiology, or cause. In this framework, the disease is treated as a single entity, and investigators have sought to identify risk factors

*Correspondence to: Emily C. Zabor, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 485 Lexington Ave, 2nd floor, New York, NY 10017, USA. zabore@mskcc.org.

that are associated with the disease. However, in recent years attention is increasingly focused on identifying subtypes of disease. This has been especially true in cancer research because of the growing use of molecular and genomic profiling, which give researchers access to many more ways in which to classify a tumor. It is now widely accepted that many cancers, including but not limited to breast [1, 2, 3], lung [4, 5], colorectal [6], and endometrial [7, 8] cancers, are comprised of specific molecular subtypes. As these subtypes are identified, it is natural to ask the question of whether they share a common etiology or in fact arise from distinct sets of risk factors. The concept of differing risk factors across subtypes of disease is known as etiologic heterogeneity.

The study of etiologic heterogeneity is fraught with many statistical challenges at every level and statistical methods are needed not only to detect the presence of etiologic heterogeneity, but also to quantify the extent of that heterogeneity. One challenge is the possibly high dimension of the data, which may include information from multiple molecular profiling platforms such as expression, copy number, mutation, and methylation data. Further, as more evidence accumulates for subtypes of cancers with distinct risk profiles according to known risk factors, it is natural to ask whether undiscovered risk factors will also exhibit differential associations across subtypes. An epidemiologic investigation of etiologic heterogeneity such as a case-control study would naturally be subject to the constraints of smaller subtype sample sizes as compared to the aggregate case group, as well as the prospect of false discovery due to the increasing number of statistical comparisons being made. However, such investigation also serves to benefit from a potentially larger effect size in at least one subtype and improved risk prediction accuracy for all patients. In previous work we investigated the statistical implications of the trade-offs in the choice between a traditional case-control design versus one that further classifies cases into subtypes through a simulation study to examine statistical power under various study design scenarios [9]. We found that over a range of risk factor prevalences and overall case-control odds ratios, only modest heterogeneity was needed before a study design that accounts for subtypes achieved equivalent power to a traditional case-control approach that considers all cases in aggregate. This result provides practical motivation to pursue development of statistical methods for the study of etiologic heterogeneity.

Epidemiologic questions of interest related to the study of etiologic heterogeneity include 1) whether a risk factor of interest has the same effect across all subtypes of disease and 2) whether risk factor effects differ across levels of each individual tumor marker by which the subtypes are defined. Early investigations of etiologic heterogeneity typically divided cases into a small number of pre-determined subtypes, based on a single molecular marker or pathologic feature, or on combinations thereof. Associations of specific subtypes with risk factors could be examined using polytomous logistic regression [10]. In recent years, however, a number of new statistical methods have been proposed for the study of etiologic heterogeneity. Chatterjee [11] proposed a two-stage regression model for use with case-control data. The method can handle a potentially large number of subtypes and allows testing for differences in the effects of risk factors both across subtypes and with respect to individual tumor markers. This method was subsequently extended to the setting of cohort studies [12]. Rosner et al. [13] constructed a strategy employing a model for use with data from a cohort study that permits simultaneous investigation of the individual risk factors and

tumor markers in a single-stage model. Finally, Wang et al. [14] proposed a two-stage model that uses the parameter estimates from a first-stage regression model as inputs in a second-stage model that captures the relationships between individual tumor markers and risk factors. This two-stage approach can be applied to cohort studies, nested case-control studies, and unmatched case-control studies.

These methods have very distinctive parametric structures and it is not immediately straightforward how results using the different methods align with each other. In this article we seek to reconcile the similarities among the methods, and to evaluate their statistical properties. To accomplish this, we employ a simplified data example to elucidate the interpretation of model parameters and available hypothesis tests, and we perform a simulation study to assess bias in effect size, type I error, and power.

2. Analytic framework

We compare four distinct available methods: polytomous logistic regression; the two-stage meta-regression method proposed by Wang et al. [14]; the two-stage regression with simultaneous estimation approach proposed by Chatterjee [11]; and the stratified logistic regression approach of Rosner et al. [13]. We focus solely on methods for the analysis of case-control data, though many of the approaches discussed can be applied in the context of other study designs. Let i index study subjects, $i = 1, \dots, N$, let k index tumor markers, $k = 1, \dots, K$, let m index disease subtypes, $m = 0, \dots, M$, where $m = 0$ denotes control subjects, and let p index risk factors, $p = 1, \dots, P$. Initially, for simplicity, we focus on a setting where we have two binary tumor markers, each of which can be either positive (+) or negative (−). These two tumor markers are cross-classified to form four disease subtypes (−/−, +/−, −/+, and +/+). Additionally, for conceptual simplicity in our primary exposition and simulations, we limit this investigation to the case of a single binary risk factor of interest. The setting explored here has tumor markers $k = 1, 2$, disease subtypes $m = 1, 2, 3, 4$, and risk factor $p = 1$.

The first epidemiologic question of interest to be addressed is whether the risk factor of interest has the same effect across all subtypes of disease. This is typically the primary question of interest in an investigation of etiologic heterogeneity and allows one to determine whether the risk factor of interest is only associated with specific subtypes of disease. From each of the available methods, we can obtain parameters β_{pm} , which represent the log odds ratio for a one-unit change in risk factor p for subtype m disease versus controls. In the case of four subtypes and one binary risk factor, there are four such log odds ratios β_{11} , β_{12} , β_{13} , and β_{14} (Table 1). We are thus interested in a test of the hypothesis $H_{0\beta}$: $\beta_{11} = \beta_{12} = \beta_{13} = \beta_{14}$. A second epidemiologic question of specific interest is whether the risk factor effects differ across levels of each individual tumor marker. This question allows one to evaluate whether a specific tumor marker is in part responsible for observed differences in log odds ratios of the risk factor across the subtypes. To answer this question, we can obtain estimates of parameters γ_{pk} , each of which represents the ratios of the log odds ratios for the risk factor defined by different levels of the k th tumor marker when each level of the other tumor markers is held constant. In the case of two binary tumor markers

and a single binary risk factor, we obtain γ_{11} and γ_{12} (Table 1). Then we can address this question with tests of the hypotheses $H_{0\gamma_{11}}: \gamma_{11} = 0$ and $H_{0\gamma_{12}}: \gamma_{12} = 0$.

Our work addresses how each of the four methods can be constructed to address these two epidemiologic questions, and compares the statistical properties of the methods. Throughout, it is important to keep in mind the original purpose of each of the four methods. Polytomous logistic regression is constructed in such a way as to naturally address the question of whether risk factor effects differ across subtypes of disease. The β_{pm} parameters are estimated directly in polytomous logistic regression. In section 3.1, we show that the γ_{pk} parameters can then be obtained indirectly as a linear combination of the estimated β_{pm} parameters. Conversely, the two-stage regression with simultaneous estimation approach of Chatterjee [11] and the stratified logistic regression approach of Rosner et al. [13] were originally proposed to address the question of whether risk factor effects differ across levels of each individual tumor marker. As such, the γ_{pk} parameters are estimated directly. Both methods also allow for inclusion of interaction effects between individual tumor markers. In sections 3.3 and 3.4 we show that when all first-order interaction terms are included in the model, the β_{pm} parameters can be obtained indirectly as a linear combination of the estimated γ_{pk} parameters. The two-stage meta-regression approach of Wang et al. [14] was specifically proposed to address both the question of whether risk factor effects differ across subtypes of disease and the question of whether risk factor effects differ across levels of each individual tumor marker. In this approach the β_{pm} parameters are directly estimated in the first-stage model and then the γ_{pk} parameters are directly estimated in the second-stage model. Details of model specification and estimation for each of the four methods follow in section 3.

3. Methods

In this section, we present details of the estimation of model parameters and the hypothesis testing procedure for each approach.

3.1. Polytomous logistic regression

Polytomous logistic regression allows for the simultaneous estimation of subtype-specific regression parameters. Let Y_i denote the disease status for subject i such that $Y_i = 0$ for a non-diseased control subject and $Y_i = m$ for a subject with disease subtype m . X_{1i} denotes the value of risk factor $p = 1$ for subject i . Then a polytomous logistic regression model is specified as

$$\Pr(Y_i = m | X_{1i}) = \frac{\exp(\beta_{0m} + \beta_{1m} X_{1i})}{1 + \sum_{m=1}^4 \exp(\beta_{0m} + \beta_{1m} X_{1i})}, m = 1, 2, 3, 4 \quad (1)$$

where β_{0m} is the intercept parameter for the m th disease subtype. To evaluate whether the risk factor has the same effect across all subtypes of disease we can perform a Wald test of the hypothesis $H_{0\beta}: \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14}$.

Defining w_{km} as the level of the k th tumor marker corresponding to the m th disease subtype we can create a linearly transformed set of parameters using

$$\beta_{jm} = \gamma_{j0} + \gamma_{j1}w_{1m} + \gamma_{j2}w_{2m} + \gamma_{j12}w_{1m}w_{2m} \quad (2)$$

for $j = 0, 1$. It follows that we can obtain estimates of the γ_{pk} parameters associated with the individual tumor marker effects in the case of $m = 4$ disease subtypes as

$$\gamma_{11} = \frac{(\beta_{12} - \beta_{11}) + (\beta_{14} - \beta_{13})}{2} \text{ and } \gamma_{12} = \frac{(\beta_{13} - \beta_{11}) + (\beta_{14} - \beta_{12})}{2}.$$

Note that while here we limit ourselves to the case of $m = 4$ disease subtypes formed by $k = 2$ tumor markers, this transformation is generalizable. Thus tests addressing the second set of questions, whether risk factor effects differ across levels of each individual tumor marker, can be accomplished using Wald tests of $H_{0\gamma_{11}}: \beta_{12} - \beta_{11} + \beta_{14} - \beta_{13} = 0$ and $H_{0\gamma_{12}}: \beta_{13} - \beta_{11} + \beta_{14} - \beta_{12} = 0$.

It is of interest to note that when data are not available on control subjects, the test for etiologic heterogeneity can be obtained using a case-only polytomous logistic regression model [15]. In the polytomous logistic regression model we must select one of the four subtypes to serve as the reference group. Because we do not have data on controls, the main effects of the individual tumor markers cannot be determined, so case-only polytomous logistic regression cannot test whether the effect of a risk factor differs across levels of each individual tumor marker. As this method produces almost identical results to those from polytomous logistic regression with regard to the question of whether a risk factor of interest has the same effect across all subtypes of disease, it will not be investigated in further detail.

3.2. Two-stage meta-regression

The method of Wang et al. [14] is a two-stage approach. As noted in section 2, this method was specifically proposed to address both the question of whether risk factor effects differ across disease subtypes and the question of whether risk factor effects differ across levels of each individual tumor marker. The first stage of the analysis uses the previously introduced polytomous logistic regression model (1). Thus the test of whether the risk factor has the same effect across all four subtypes of disease is identical to the one used in section 3.1 above. A second-stage analysis is then employed to directly estimate the parameters γ_{10} , γ_{11} , and γ_{12} for risk factor $p = 1$ using a weighted linear regression model,

$$\hat{\beta}_{1m} = \gamma_{10} + \gamma_{11}w_{1m} + \gamma_{12}w_{2m} + e_{1m}, \quad (3)$$

where $\hat{\beta}_{1m}$ is the estimated log odds ratio of subtype m versus controls for risk factor $p = 1$ from the polytomous logistic regression model and e_{1m} is within study sampling error such

that $Var(e_{1m}) = \widehat{Var}(\hat{\beta}_{1m})$. We test whether the risk factor effect differs across levels of each individual tumor marker using Wald tests of the hypotheses $H_{0\gamma_{11}}: \gamma_{11} = 0$ and $H_{0\gamma_{12}}: \gamma_{12} = 0$.

Wang et al. [14] also propose that the second stage model in equation 3 can be extended to include a random effect, which could capture variance between subtypes not explained by the included tumor markers. Alternatively, the second stage model in equation 3 can incorporate interaction terms between the individual tumor markers in order to evaluate whether the effect of the risk factor associated with one tumor marker actually depends on the level of another tumor marker. These alternative second-stage model specifications are not examined in depth, but may in fact prove more appropriate in certain study settings.

3.3. Two-stage regression with simultaneous estimation

The method of Chatterjee [11] is also a two-stage approach with a similar model structure. However, unlike the preceding two-stage meta-regression method [14], this approach specifies a joint likelihood and uses a maximum likelihood estimation procedure to simultaneously estimate the first-stage and second-stage regression parameters. When the total number of disease subtypes is moderate, maximum likelihood estimation of the two-stage model is relatively straightforward, though a pseudo-conditional likelihood estimation method is also proposed for the case when the number of disease subtypes is large [11]. This method was proposed in order to address the question of whether risk factor effects differ across levels of each individual tumor marker. The first-stage model is the polytomous logistic regression model defined in equation 1. The second-stage model to address the question of whether the effect of risk factor $p = 1$ differs across levels of each individual tumor marker can be constructed as,

$$\beta_{1m} = \gamma_{10} + \gamma_{11}w_{1m} + \gamma_{12}w_{2m}. \quad (4)$$

In this framework we are interested in testing the independent effect of each tumor marker when all other tumor markers are held constant. We can test whether the risk factor effect differs across levels of each individual tumor marker using score tests of the hypotheses $H_{0\gamma_{11}}: \gamma_{11} = 0$ and $H_{0\gamma_{12}}: \gamma_{12} = 0$.

However, this model also allows for inclusion of interaction effects between individual tumor markers. If we incorporate all interaction effects we would utilize

$$\beta_{1m} = \gamma_{10} + \gamma_{11}w_{1m} + \gamma_{12}w_{2m} + \gamma_{112}w_{1m}w_{2m}, \quad (5)$$

where $\gamma_{pk_1k_2}$ is a measure of the interaction effect between the k_1 th and k_2 th tumor markers with respect to the p th risk factor. Note that this is equivalent to equation 2 for the case of $j = 1$. Analogous to section 3.1, in this setting the β_{pm} parameter estimates can be obtained based on linear combinations of the γ_{pk} parameter estimates using the fact that $\beta_{11} = \gamma_{10}$,

$\beta_{12} = \gamma_{10} + \gamma_{11}$, $\beta_{13} = \gamma_{10} + \gamma_{12}$, and $\beta_{14} = \gamma_{10} + \gamma_{11} + \gamma_{12} + \gamma_{112}$. Thus we can test whether the risk factor has the same effect across all four disease subtypes by performing a Wald test of the hypothesis $H_{0\beta}: \gamma_{11} = \gamma_{12} = \gamma_{112} = 0$. One could also test the hypothesis $H_{0\gamma_{pk_1k_2}}: \gamma_{112} = 0$ in order to determine whether the effect of risk factor $p = 1$ associated with tumor marker $k = 1$ actually depends on the level tumor marker $k = 2$.

3.4. Stratified logistic regression

As an alternative to a two-stage approach, Rosner et al. [13] proposed a single-stage regression method. This method was originally designed to address the question of whether risk factor effects differ across levels of each individual tumor marker using a computational structure for which software is readily available. Let Z_{mi} indicate the disease status for subject i specific to subtype m disease such that

$$Z_{mi} = \begin{cases} 1 & \text{if } Y_i = m \\ 0 & \text{if } Y_i \neq m, \end{cases}$$

for $m = 1, \dots, M$. In control subjects $Z_{mi} = 0$ for all m . In contrast to all previously discussed methods, here a data augmentation approach is used, such that each case contributes m correlated outcomes, one for each combination of tumor markers, i.e. each disease subtype m [13]. This approach was originally designed for use in the setting of cohort studies [13] and was implemented using a Cox regression model stratified by the disease subtype. However, using the fact that a stratified Cox regression model is equivalent to a stratified logistic regression model [16], also known as a conditional logistic regression model, we can easily apply the method in the setting of a case-control study when time is constant for all included subjects and data are structured as described. The same data augmentation approach is used, and the logistic regression model is stratified by disease subtype.

To address the question of whether a risk factor of interest, X_{1i} has the same effect across levels of each individual tumor marker, the stratified logistic regression model can be specified,

$$\Pr(Z_{mi}=1|X_{1i}, \mathbf{w}_m) = \frac{\exp(\alpha_m + \gamma_{10}X_{1i} + \gamma_{11}X_{1i}w_{1m} + \gamma_{12}X_{1i}w_{2m})}{1 + \exp(\alpha_m + \gamma_{10}X_{1i} + \gamma_{11}X_{1i}w_{1m} + \gamma_{12}X_{1i}w_{2m})}. \quad (6)$$

where $\mathbf{w}_m = \{w_{1m}, \dots, w_{km}\}$ is the vector of tumor markers for the m th subtype and α_m is the stratum-specific intercept term, which cancels out in the conditional likelihood [17]. We can use this model to test whether the risk factor effect differs across levels of each individual tumor marker using Wald tests of the hypotheses $H_{0\gamma_{11}}: \gamma_{11} = 0$ and $H_{0\gamma_{12}}: \gamma_{12} = 0$.

Rosner's [13] stratified logistic regression approach also allows for inclusion of interaction effects between individual tumor markers. When all interaction effects are included, we obtain the model

$$\Pr(Z_{mi}=1|X_{1i}, \mathbf{w}_m) = \frac{\exp(\alpha_m + \gamma_{10}X_{1i} + \gamma_{11}X_{1i}w_{1m} + \gamma_{12}X_{1i}w_{2m} + \gamma_{112}X_{1i}w_{1m}w_{2m})}{1 + \exp(\alpha_m + \gamma_{10}X_{1i} + \gamma_{11}X_{1i}w_{1m} + \gamma_{12}X_{1i}w_{2m} + \gamma_{112}X_{1i}w_{1m}w_{2m})} \quad (7)$$

and as in section 3.3 the β_{pm} parameters can be obtained indirectly as a linear combination of the γ_{pk} parameters to test whether the risk factor has the same effect across all four disease subtypes as described in section 3.3.

3.5. Software

All statistical analyses were conducted using R software version 3.3.1 (R Core Development Team, Vienna, Austria). For polytomous logistic regression, the `multinom` function from the `nnet` package was used for estimation and the `wald.test` function from the `aod` package was used for significance testing. For the second-stage model in the two-stage meta-regression method of Wang et al. [14], the `rma.mv` function from the `metafor` package was used for estimation and significance testing. Estimation and significance testing for the two-stage regression with simultaneous estimation method of Chatterjee [11] was conducted using an R function provided by the authors, except in the case of the test of $H_{0\beta}$ which was conducted using the `wald.test` function from the `aod` package. Finally, following data augmentation, the `clogit` function from the `survival` package was used for estimation and testing for the stratified logistic regression method of Rosner et al. [13]. For all methods besides that of Chatterjee [11], other standard software packages that support the underlying statistical models could be used. However all require transformation of the results from one parametric configuration to another and use the parameter estimates and variance-covariance matrices for hypothesis testing.

4. Data example

To illustrate the methods we use data from a previous study in which we combined data from two large breast cancer case-control studies, the Cancer and Steroid Hormone (CASH) study and the Womens' Contraceptive and Reproductive Experiences (CARE) study, leading to a total of 984 cases and a corresponding 1592 controls [18]. Our goal in this section is not to conduct a detailed analysis of etiologic heterogeneity. Instead, we focus on a simplified strategy that addresses the etiologic heterogeneity of breast cancer classified into subtypes described by estrogen receptor (ER) and progesterone receptor (PR) status from the perspective of a single risk factor, oral contraceptive (OC) use. The purpose is simply to contrast the various modeling strategies.

The primary results from the four methods are presented in Table 2. The top portion of the table contains results relevant to the question of whether OC use has the same effect across the four disease subtypes. This question is addressed with equation 1 for polytomous logistic regression and the method of Wang et al. [14], equation 5 for the method of Chatterjee [11], and equation 7 for the method of Rosner et al. [13]. All methods lead to rejection of the null

hypothesis (all p-values < 0.05), so regardless of the method we conclude that the effect of OC use differs across the four disease subtypes. We note that parameter estimates for polytomous logistic regression, the method of Wang et al. [14], and the method of Chatterjee [11] are identical. This is expected as the first-stage model for the method of Wang et al. [14] is simply the polytomous logistic regression model, and when all first order interaction effects are included in the method of Chatterjee [11] and maximum likelihood estimation is used, this model should produce results that are nearly identical to those from the polytomous logistic regression model. Finally, we see that there are some small differences between the parameter estimates from the method of Rosner et al. [13] as compared to the other three methods, in that the parameter estimates are all less positive in magnitude.

The lower portion of Table 2 displays results related to the questions of whether the effect of OC use differs across levels of ER status when PR status is held constant, and whether the effect of OC use differs across levels of PR status when ER status is held constant. These questions are addressed with equation 1 for polytomous logistic regression, equation 3 for the method of Wang et al. [14], equation 4 for the method of Chatterjee [11], and equation 6 for the method of Rosner et al. [13]. Again, the parameter estimates and p-values are similar. Regardless of the method, we conclude that the effect of OC use on breast cancer risk differs by ER status, but conclude that the effect of OC use on breast cancer risk does not differ by PR status.

5. Simulation study

The simulation study is conducted using a similar framework to the data example, with four disease subtypes formed by cross-classification of two tumor markers as described in section 2. There is a single binary risk factor of interest, with a prevalence among control subjects of $q = 0.3$. Each simulation uses 1000 controls and 1000 cases, with the cases divided equally among the four disease subtypes. The true regression coefficients are fixed at β_{1m} for subtype m disease, $m = 1, 2, 3, 4$. Risk factor data are randomly generated for each subject with disease subtype m from a binomial distribution with probability $\exp(q\beta_{1m})/[1 + \exp(q\beta_{1m})]$ and for each control subject with probability $\exp(q)/[1 + \exp(q)]$. For each simulation setting, we generate 1000 simulated data sets.

To address the question of whether the risk factor effects differ across the disease subtypes, the simulation study employs equation 1 for polytomous logistic regression and the method of Wang et al. [14], equation 5 for the method of Chatterjee [11], and equation 7 for the method of Rosner et al. [13]. To address the question of whether the risk factor effect differs across levels of each individual tumor marker, the simulation study uses equation 1 for polytomous logistic regression, equation 3 for the method of Wang et al. [14], equation 4 for the method of Chatterjee [11], and equation 6 for the method of Rosner et al. [13].

It is important to note that some of the simulation settings imply an interaction effect between the individual tumor markers whereas some of the simulation settings imply a main effects model with no interaction effect. When there is no interaction between the individual tumor markers, i.e. when $\gamma_{112} = 0$, then $\beta_{14} = \beta_{12} + \beta_{13} - \beta_{11}$ and we can test whether risk factor effects differ across the disease subtypes with a test of $H_{0\beta}: \gamma_{11} = \gamma_{12} = 0$. In settings

where there is truly no interaction effect, we explore the method of Chatterjee [11] using both equation 5 and equation 4 to determine whether there is an efficiency gain from using a model that does not incorporate an interaction effect as compared to a model that does.

5.1. Data simulated under the null hypothesis

The first set of simulations is conducted under the null hypothesis for the question of whether the risk factor effect differs across the four disease subtypes, and under the null hypothesis for the question of whether the risk factor effect differs across levels of each individual tumor marker. We set $\beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0.1$ and therefore $\gamma_{11} = \gamma_{12} = 0$. Equivalent disease subtype effects such as this implies no interaction effect between the individual tumor markers. For the question of whether the risk factor effect differs across the four disease subtypes, we see that the size of the test is 0.051 for polytomous logistic regression, the method of Wang et al. [14], and the method of Chatterjee [11] whereas the method of Rosner et al. [13] has an inflated type I error of 0.089 (Table 3, upper portion). The biases in parameter estimates are small for all methods except that of Rosner et al. [13]. When we apply the main effects model of Chatterjee using equation 4, we find similarly small biases but slightly inflated type I error of 0.068 (data not shown).

For the question of whether the risk factor effect differs across levels of each individual tumor marker, polytomous logistic regression and the method of Chatterjee [11] have very similar type I errors for γ_{11} of 0.063 and for γ_{12} of 0.051 and 0.050, respectively (Table 3, lower portion). The method of Wang et al. [14] has lower type I errors, 0.037 and 0.031 for testing γ_{11} and γ_{12} , respectively; conversely, the method of Rosner et al. [13] has inflated type I errors of 0.077 and 0.073. In this setting all methods produce parameter estimates with comparably small biases.

5.2. Data simulated under the alternative hypothesis

The second set of simulations is conducted under the alternative hypothesis for the question of whether the risk factor effect differs across the four disease subtypes, and under the alternative hypothesis for the question of whether the risk factor effect differs across levels of each individual tumor marker. Here we let $\beta_{11} = 0.2$, $\beta_{12} = \beta_{13} = 0.3$, and $\beta_{14} = 0.8$ so that $\gamma_{11} = \gamma_{12} = 0.3$. For the question of whether the effect of the risk factor differs across the four subtypes, we see that polytomous logistic regression, the method of Wang et al. [14] and the method of Chatterjee [11] all have power of 0.822 whereas the method of Rosner et al. [13] has higher power of 0.862 (Table 3, upper portion). However, recall that the method of Rosner et al. [13] had higher type I error than the other methods, and so calibration is needed to truly compare the signal detection strengths of the methods. We also find that while biases are generally very small for most methods, there is substantial bias in parameter estimates for the method of Rosner et al. [13].

For the question of whether the risk factor effect differs across levels of each individual tumor marker, polytomous logistic regression and the method of Chatterjee [11] again have similar power (Table 3, lower portion). The method of Wang et al. [14] has lower power whereas the method of Rosner et al. [13] has slightly higher power. However, again we must

recall that the method of Rosner et al. [13] had inflated type I error. All methods produce parameter estimates with small biases.

In order to compare the power of the methods in a calibrated manner, we did the following. First we varied the effect size by fixing $\beta_{11} = 0.2$ and $\beta_{12} = \beta_{13} = 0.3$, and incrementally increasing β_{14} from 0.3 to 0.9. This allowed us to determine how large the subtype four effect size, β_{14} , needs to be in order to achieve various levels of power to address whether the risk factor effect differs across the four disease subtypes. We calibrated the comparison by ranking the simulated p-values under the null hypothesis and choosing the critical value that ensured the test size was exactly 0.05, then used this critical value to determine power. Figure 1A shows the resulting power curves for the different methods. We can see that after calibration of type I error, the four methods have indistinguishable power. Note that one of these cases, when $\beta_{14} = 0.4$, implies no interaction effect between the individual tumor markers. Whereas the calibrated power using Chatterjee's equation 5 results in a power of 0.122, equation 4 results in slightly lower calibrated power of 0.119 (data not shown).

The power to address whether the risk factor effect differs across levels of each individual tumor marker is similarly compared. Figure 1B shows the power to detect an effect for γ_{1k} . We find that the results are similar across the four methods.

5.3. Data simulated under different configurations

In the Supplemental Materials, results from additional simulations are presented. These encompass additional examples of the simulation configurations from the preceding two sections using different parametric values β_{1m} and risk factor prevalences q (Supplemental Table S1, Supplemental Figures S1 and S2). In addition, we present a number of simulation results when there are $M = 16$ subtypes defined on the basis of $K = 4$ binary tumor markers (Supplemental Tables S2 and S3). The results generally show the same patterns as those presented in Table 3 and Figure 1. The polytomous logistic regression approach and the method of Chatterjee [11] have very similar properties, while in contrast the method of Wang et al. [14] is conservative for inferences related to the γ_{pk} parameters and the method of Rosner et al. [13] is somewhat anti-conservative and demonstrates biases in parameter estimation.

6. Discussion

We have defined two key questions that epidemiologists seek to answer in studies of etiologic heterogeneity and have shown how to address these questions using each of the methods that have been proposed. We demonstrated the distinctions of the methods by creating a unified notation. Our simulations show that the stratified logistic regression method of Rosner et al. [13] results in substantial biases in parameter estimation for addressing whether risk factor effects differ across levels of the disease subtype, although we acknowledge that this was not a stated goal of the method by the authors. Additionally, the method is anti-conservative. All other methods have type I error close to the nominal level. In the simplified setting examined here, whereas the other methods all estimate eight parameters, the method of Rosner et al. [13] conditions out the constant terms and only involves estimation of four parameters. The conditional nature of this model clearly has

implications for the validity of parameter estimates and hypothesis tests related to the question of heterogeneity across disease subtypes. For addressing whether risk factor effects differ across levels of each individual tumor marker, polytomous logistic regression and the two-stage regression with simultaneous estimation method of Chatterjee [11] perform similarly with respect to type I error whereas the two-stage meta-regression method of Wang et al. [14] is overly conservative and the stratified logistic regression method of Rosner et al. [13] is anti-conservative. When differences in type I error are calibrated, all methods achieve similar power.

In this article we have focused on subtypes formed by cross-classification of tumor markers, and on the distinct influences of the individual tumor markers. In breast cancer research, subtypes based on immunohistochemical staining of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) are commonly formed. Each of these tumor markers can be either positive (+) or negative (–) and the disease subtypes are defined as luminal A (ER+ or PR+, HER2–), luminal B (ER+ or PR+, HER2+), HER2–type (ER–, PR–, HER2+), and triple negative (ER–, PR–, HER2–). This configuration is not congruent with the second stage models described in equations 3–4 on which the methods proposed by Wang et al. [14], Chatterjee [11] and Rosner et al. [13] are based. Of the methods compared here, only polytomous logistic regression can address whether a risk factor effect differs across subtypes that are not formed by cross-classification of the individual tumor markers. This is important for epidemiologic researchers, who must carefully consider whether the individual tumor markers are of interest, or if it is truly a more complex aggregation of those tumor markers that is expected to demonstrate a differential association with risk factors.

The methods of Chatterjee et al. [11] and Rosner et al. [13] were clearly designed with the goal of studying multiple tumor markers in a flexible modeling framework. Thus one can envision a study with a number of tumor markers where the dimension is reduced by eliminating selected, or all, interactions, and thereby permitting an analysis that would not be possible in the context of polytomous logistic regression. Further exploration is needed into the performance of each method under an increasing number of subtypes and risk factors.

We have limited this investigation to methods that require pre-specification of subtypes. With increasing use of genomic profiling, often it will be of interest to first identify disease subtypes based on a large number of either binary or continuous tumor markers. Begg et al. [18] proposed an approach to address this challenge by introducing a scalar measure of heterogeneity that allows an investigator to compare different subtyping configurations based on, for example, gene expression data. The ultimate investigation of risk factor associations with the resulting subtypes in this approach relies on polytomous logistic regression. Future research should address whether other methods compared here could be combined advantageously with methods designed to identify subtypes.

In conclusion, the study of etiologic heterogeneity will become increasingly common in the age of genomic profiling and personalized medicine, and statistical methods are needed to

reliably address these questions. The results of this investigation can serve to guide selection of a method that will favorably balance statistical and practical considerations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research was supported by the National Cancer Institute, awards CA163251 and CA167237. We are grateful to Nilanjan Chatterjee and Haoyu Zhang for courteously supplying us with their code for running the Chatterjee [11] method.

References

1. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. Molecular portraits of human breast tumours. *Nature*. 2000; 406(6797):747–52. DOI: 10.1038/35021093 [PubMed: 10963602]
2. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*. 2001; 98(19):10 869–74. DOI: 10.1073/pnas.191367098
3. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A*. 2003; 100(18):10 393–8. DOI: 10.1073/pnas.1732912100 [PubMed: 12506193]
4. Ahrendt SA, Decker PA, Alawi EA, Zhu Yr YR, Sanchez-Cespedes M, Yang SC, Haasler GB, Kajdacsy-Balla A, Demeure MJ, Sidransky D. Cigarette smoking is strongly associated with mutation of the k-ras gene in patients with primary adenocarcinoma of the lung. *Cancer*. 2001; 92(6):1525–30. [PubMed: 11745231]
5. Riely GJ, Kris MG, Rosenbaum D, Marks J, Li A, Chitale DA, Nafa K, Riedel ER, Hsu M, Pao W, et al. Frequency and distinctive spectrum of kras mutations in never smokers with lung adenocarcinoma. *Clin Cancer Res*. 2008; 14(18):5731–4. DOI: 10.1158/1078-0432.ccr-08-0646 [PubMed: 18794081]
6. Ogino S, Chan AT, Fuchs CS, Giovannucci E. Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field. *Gut*. 2011; 60(3):397–411. DOI: 10.1136/gut.2010.217182 [PubMed: 21036793]
7. Brinton LA, Felix AS, McMeekin DS, Creasman WT, Sherman ME, Mutch D, Cohn DE, Walker JL, Moore RG, Downs LS, et al. Etiologic heterogeneity in endometrial cancer: evidence from a gynecologic oncology group trial. *Gynecol Oncol*. 2013; 129(2):277–84. DOI: 10.1016/j.ygyno.2013.02.023 [PubMed: 23485770]
8. Schildkraut JM, Iversen ES, Akushevich L, Whitaker R, Bentley RC, Berchuck A, Marks JR. Molecular signatures of epithelial ovarian cancer: analysis of associations with tumor characteristics and epidemiologic risk factors. *Cancer Epidemiol Biomarkers Prev*. 2013; 22(10):1709–21. DOI: 10.1158/1055-9965.epi-13-0192 [PubMed: 23917454]
9. Begg CB, Zabor EC. Detecting and exploiting etiologic heterogeneity in epidemiologic studies. *Am J Epidemiol*. 2012; 176(6):512–8. DOI: 10.1093/aje/kws128 [PubMed: 22922440]
10. Dubin N, Pasternack BS. Risk assessment for case-control subgroups by polychotomous logistic regression. *Am J Epidemiol*. 1986; 123(6):1101–17. [PubMed: 3706280]
11. Chatterjee N. A two-stage regression model for epidemiological studies with multivariate disease classification data. *Journal of the American Statistical Association*. 2004; 99(465):127–138. DOI: 10.1198/016214504000000124

12. Chatterjee N, Sinha S, Diver WR, Feigelson HS. Analysis of cohort studies with multivariate and partially observed disease classification data. *Biometrika*. 2010; 97(3):683–698. DOI: 10.1093/biomet/asq036 [PubMed: 22822252]
13. Rosner B, Glynn RJ, Tamimi RM, Chen WY, Colditz GA, Willett WC, Hankinson SE. Breast cancer risk prediction with heterogeneous risk profiles according to breast cancer tumor markers. *Am J Epidemiol*. 2013; 178(2):296–308. DOI: 10.1093/aje/kws457 [PubMed: 23645624]
14. Wang M, Kuchiba A, Ogino S. A meta-regression method for studying etiological heterogeneity across disease subtypes classified by multiple biomarkers. *Am J Epidemiol*. 2015; 182(3):263–70. DOI: 10.1093/aje/kwv040 [PubMed: 26116215]
15. Begg CB, Zhang ZF. Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiol Biomarkers Prev*. 1994; 3(2):173–5. [PubMed: 8049640]
16. Gail MH, Lubin JH, Rubinstein LV. Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika*. 1981; 68(3):703–707. DOI: 10.1093/biomet/68.3.703
17. Breslow NE, Day NE. Statistical methods in cancer research. volume i - the analysis of case-control studies. IARC scientific publications. 1980; (32):5–338. [PubMed: 7216345]
18. Begg CB, Zabor EC, Bernstein JL, Bernstein L, Press MF, Seshan VE. A conceptual and methodological framework for investigating etiologic heterogeneity. *Stat Med*. 2013; 32(29):5039–52. DOI: 10.1002/sim.5902 [PubMed: 23857589]

Figure 1.

Table 1

Interpretation of model parameters

Does the risk factor effect differ with respect to subtypes?	
Parameter	Interpretation
β_{11}	log odds ratio for subtype $m = 1$ vs controls
β_{12}	log odds ratio for subtype $m = 2$ vs controls
β_{13}	log odds ratio for subtype $m = 3$ vs controls
β_{14}	log odds ratio for subtype $m = 4$ vs controls
Does the risk factor effect differ with respect to tumor markers?	
Parameter	Interpretation
γ_{11}	average of differences in log odds ratios when tumor marker $k = 1$ is + vs – and $k = 2$ is fixed
γ_{12}	average of differences in log odds ratios when tumor marker $k = 2$ is + vs – and $k = 1$ is fixed

Table 2

Results of data example

Does the risk factor effect differ with respect to subtypes?				
Method	Subtype	Parameter	Estimate	p-value
Polytomous ¹	ER-/PR-	β_{11}	0.31	0.042
	ER+/PR-	β_{12}	-0.11	
	ER-/PR+	β_{13}	0.22	
	ER+/PR+	β_{14}	0.03	
Wang ²	ER-/PR-	β_{11}	0.31	0.042
	ER+/PR-	β_{12}	-0.11	
	ER-/PR+	β_{13}	0.22	
	ER+/PR+	β_{14}	0.03	
Chatterjee ³	ER-/PR-	β_{11}	0.31	0.042
	ER+/PR-	β_{12}	-0.11	
	ER-/PR+	β_{13}	0.21	
	ER+/PR+	β_{14}	0.03	
Rosner ⁴	ER-/PR-	β_{11}	0.29	0.029
	ER+/PR-	β_{12}	-0.15	
	ER-/PR+	β_{13}	0.18	
	ER+/PR+	β_{14}	0.00	
Does the risk factor effect differ with respect to tumor markers?				
Method	Tumor marker	Parameter	Estimate	p-value
Polytomous ¹	ER	γ_{11}	-0.30	0.046
	PR	γ_{12}	0.02	0.887
Wang ²	ER	γ_{11}	-0.33	0.031
	PR	γ_{12}	0.05	0.731
Chatterjee ³	ER	γ_{11}	-0.33	0.028
	PR	γ_{12}	0.05	0.719
Rosner ⁴	ER	γ_{11}	-0.34	0.024
	PR	γ_{12}	0.05	0.739

¹ Polytomous logistic regression² Two-stage meta-regression [14]³ Two-stage regression with simultaneous estimation [11]⁴ Stratified logistic regression [13]

Table 3

Results of simulation study

Does the risk factor effect differ with respect to subtypes?						
Null hypothesis				Alternative hypothesis		
Method	Parameter	Truth	Bias	Type I error	Truth	Power
Polytomous ¹	β_{11}	0.1	-0.000	0.051	0.2	0.822
	β_{12}	0.1	-0.004		0.3	-0.006
	β_{13}	0.1	-0.001		0.3	-0.002
	β_{14}	0.1	-0.007		0.8	-0.007
Wang ²	β_{11}	0.1	-0.000	0.051	0.2	0.822
	β_{12}	0.1	-0.004		0.3	-0.006
	β_{13}	0.1	-0.001		0.3	-0.002
	β_{14}	0.1	-0.007		0.8	-0.007
Chatterjee ³	β_{11}	0.1	-0.000	0.051	0.2	0.822
	β_{12}	0.1	-0.004		0.3	-0.006
	β_{13}	0.1	-0.001		0.3	-0.002
	β_{14}	0.1	-0.007		0.8	-0.007
Rosner ⁴	β_{11}	0.1	0.047	0.089	0.2	0.190
	β_{12}	0.1	0.043		0.3	0.178
	β_{13}	0.1	0.046		0.3	0.182
	β_{14}	0.1	0.040		0.8	0.146
Does the risk factor effect differ with respect to tumor markers?						
Null hypothesis				Alternative hypothesis		
Method	Parameter	Truth	Bias	Type I error	Truth	Power
Polytomous ¹	γ_{11}	0.0	-0.005	0.063	0.3	0.589

Does the risk factor effect differ with respect to subtypes?									
Null hypothesis					Alternative hypothesis				
Method	Parameter	Truth	Bias	Type I error	Truth	Bias	Power		
Wang ²	γ_{12}	0.0	-0.001	0.051	0.3	-0.001	0.599		
	γ_{11}	0.0	-0.005	0.037	0.3	0.005	0.483		
	γ_{12}	0.0	-0.001	0.031	0.3	0.010	0.475		
Chatterjee ³	γ_{11}	0.0	-0.005	0.063	0.3	0.006	0.560		
	γ_{12}	0.0	-0.001	0.050	0.3	0.010	0.572		
Rosner ⁴	γ_{11}	0.0	-0.005	0.077	0.3	-0.012	0.605		
	γ_{12}	0.0	-0.001	0.073	0.3	-0.008	0.624		

¹ Polytomous logistic regression
² Two-stage meta-regression [14]
³ Two-stage regression with simultaneous estimation [11]
⁴ Stratified logistic regression [13]