Published in final edited form as:
Stat Methods Med Res. 2017 June ; 26(3): 1416–1428. doi:10.1177/0962280215581112.

A statistical method for studying correlated rare events and their risk factors

Xiaonan Xue, **Mimi Y Kim**, **Tao Wang**, **Mark H Kuniholm**, and **Howard D Strickler**
Division of Biostatistics, Department Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA

Abstract

Longitudinal studies of rare events such as cervical high-grade lesions or colorectal polyps that can recur often involve correlated binary data. Risk factor for these events cannot be reliably examined using conventional statistical methods. For example, logistic regression models that incorporate generalized estimating equations often fail to converge or provide inaccurate results when analyzing data of this type. Although exact methods have been reported, they are complex and computationally difficult. The current paper proposes a mathematically straightforward and easy-to-use two-step approach involving (i) an additive model to measure associations between a rare or uncommon correlated binary event and potential risk factors and (ii) a permutation test to estimate the statistical significance of these associations. Simulation studies showed that the proposed method reliably tests and accurately estimates the associations of exposure with correlated binary rare events. This method was then applied to a longitudinal study of human leukocyte antigen (HLA) genotype and risk of cervical high grade squamous intraepithelial lesions (HSIL) among HIV-infected and HIV-uninfected women. Results showed statistically significant associations of two HLA alleles among HIV-negative but not HIV-positive women, suggesting that immune status may modify the HLA and cervical HSIL association. Overall, the proposed method avoids model nonconvergence problems and provides a computationally simple, accurate, and powerful approach for the analysis of risk factor associations with rare/uncommon correlated binary events.

Keywords

correlated data; rare events; permutation; generalized estimating equation; exact method

1 Introduction

Longitudinal cohort studies often involve repeated observations related to conditions (events) that can recur over time, but are nonetheless uncommon or even rare. Examples include repeated occurrences of AIDS-defining illnesses, adverse pregnancy outcomes (e.g.

Reprints and permissions: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Xiaonan Xue, Division of Biostatistics, Department of Epidemiology and Population Health, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, USA. xiaonan.xue@einstein.yu.edu.

Conflict of interest
None declared.

pre-eclampsia, prematurity, and fetal abnormalities) and the repeated development of cervical neoplasia, polyps of the colon, benign breast disease, etc. Because repeated events involving the same subjects over time are often correlated, statistical methods that take into account these inherent intra-subject correlations have been developed, such as logistic regression models that incorporate generalized estimating equations (GEE).[1] However, if the events are rare, either complete separation or quasi-separation[2] can occur so that the estimate for at least one coefficient in the regression model and their standard error will be infinite, leading to non-convergence of the model. Complete separation, using an example with a single binary exposure variable, corresponds to both "empty cells" in the off-diagonal of the exposure by event table; quasi-separation corresponds to only one empty cell in the off-diagonal cells. Further, even if the model did converge, the conventional assumption of asymptotic consistency and normality for the parameter estimates may not be applicable; i.e. the effect estimates may be inaccurate and statistical inferences may be invalid.[3–9] These concerns hold true even for a large size of the study and, furthermore, may grow as the use of genetic/epi-genetic assays and other new technologies increasingly involve the analysis of exposures (e.g. genotypes) that are also rare, leading to either complete or quasi-separation and therefore nonconvergence.

Exact conditional logistic regression models have long been used to study binary events in cross-sectional and case-control studies (i.e. single endpoint) with sparse data (e.g. due to small sample size, multiple exposure strata, few events, etc.).[10–14] However, there are few exact methods for correlated binary event data. Tang et al.[15] considered exact and approximate unconditional methods for testing the equality of successful surgery rates for both eyes between two groups of patients. But the method is only applicable when each cluster contains exactly two individual observations. Hunsberger et al.[16] proposed a simulation-based method for testing logistic regression coefficients with cluster samples when there are few positive outcomes. In their proposed simulation-based method, the approximate distribution of the generalized-score test statistic under the null hypothesis was generated from simulation. To account for the correlation between binary data, the intra-class correlation parameter was estimated first from the original data set and then was used when generating the distribution of the test statistics. This proposed method therefore lessens the reliance on asymptotic distribution assumptions; however, it does rely on the fact that the logistic regression model to be converged. For rare or uncommon events, we often encounter nonconvergence of the logistic regression model. An exact trend test on binary correlated data was proposed[17] based on a quadratic exponential model for multivariate binary outcomes.[18,19] In addition to conditioning on the sufficient statistics for baseline parameters, their exact inference further conditioned on the sufficient statistics for the correlation parameter in order to eliminate the nuisance correlation parameter. The method can be used for a logistic regression model with a single binary or ordinal scale variable. However, this additional condition on the sufficient statistics of the correlation parameter imposes more constraint on the data space so that the computation of the exact *p*-value requires using a complicated algorithm involving a network approach[20] which currently cannot be implemented directly with standard statistical software. A Bayesian approach was proposed to deal with clustered binary data with complete or quasi-complete separation through the use of a weakly informative prior distribution[21] and another approach was recently proposed to extract information about the prior distribution from part of the data and use this estimated prior distribution for the remaining part of the data.[22] These methods provide a plausible solution to rare event clustered binary data. However, the Bayesian methods assume a mixed effects model rather than a marginal model in which the intra-class correlation is treated as a nuisance parameter. Furthermore, the Bayesian methods also require extensive and complicated computations. Methods to analyze correlated binary data for rare or uncommon events with few assumptions regarding the correlation structure of the data, and are mathematically straightforward and easy to use, are needed.

To address these concerns, we propose a two-step approach, involving an additive model to measure associations between potential risk factors and rare or uncommon events that are subject to recurrence, followed by a permutation test, to estimate the statistical significance of these associations. Additive models, such as linear regression models, are much more likely to converge than multiplicative models, such as logistic regression models. However, statistical inference based on normal approximation may no longer be appropriate when events are rare or uncommon. Instead, permutation tests provide a mathematically straightforward and computationally simple approach that avoids any parametric assumption of the parameter estimates.

In this paper, we first present the two-step approach, and then use simulation studies to evaluate the performance of the new method and compare it with conventional statistical models in section 3. In section 4, we apply the two-step approach to a real data set from a longitudinal study of human leukocyte antigen (HLA) genotype and risk of high grade squamous intraepithelial lesions (HSIL), an uncommon event that can recur, among HIV-infected and HIV-uninfected women. Finally, we present our conclusions and discussion in section 5.

2 A two-step approach

Consider a binary disease outcome Y_{ij} for subject i at jth visit, and x_i is a binary exposure variable of interest (e.g. a genetic variable such as the presence or absence HLA genotype Drb*15:01), where $i = 1, \ldots, n$ and $j = 1, \ldots, J_i$. Here, we focus on risk factors that are also binary, since the motivating example for this paper is a study of genetic variation and cervical HSIL.

When the outcome is rare so few $Y_{ij} = 1$ and the remainder are zero, a logistic regression of Y_{ij} on X_i may either not converge or fail to yield a reliable statistical inference. An alternative is to use an additive model (i.e. linear regression model), which uses an “identity” link function. Linear regression models provide estimates of difference in event risk between exposure groups: when events are rare, the difference in risk is close to 0, not near the boundary of its parameter space, thus achieve model convergence much more readily than logistic regression models. However, the distribution of the risk difference tends to be skewed when events are rare and, therefore, the use of a normal approximation may lead to low statistical power. Instead, we propose using a permutation test to empirically examine the statistical significance of the main and interactive effects of the exposure.

2.1 Assessing the exposure effect

A linear regression model is defined as follows

$$P(Y_{ij}=1)=\beta_0+\beta_1 x_i \quad (1)$$

where β_1 is interpreted as the difference in proportions of events between the two exposure groups. This difference can be interpreted as the increased risk due to the exposure. Denote X_i be the covariate matrix for subject i, $X_i = (1, x_i)^T II$ and $\beta = (\beta_0, \beta_1)^T$ where II is a vector of 1's. Model (1) is then estimated using the following estimating equation

$$\sum_{i=1}^{N} X_i V_i^{-1}(Y_i - X_i^T \beta)=0$$

where $V_i=A_i^{1/2}R_i(\alpha)A_i^{1/2}$ and $Cov(Y_i) = A_i\phi$ and $R_i(\alpha)$ is the working correlation between repeated events from the same subject and ϕ is the dispersion parameter. An appropriate working correlation can be assumed and a robust variance is used. However, when events are rare, although less frequently as compared to the logistic regression GEE model, the linear regression GEE model can occasionally have a nonconvergence problem, related to the estimation of the robust variance. Further, the validity of the statistical inference on β_1 under rare events is unknown.

Therefore, we propose to first estimate β_1 assuming independence between subject's repeated observations and then obtain an empirical p-value for β_1 using a permutation test. With an independent working correlation and a binary x, it can be easily shown that $\hat{\beta}_1 = \hat{P}_1 - \hat{P}_0$ where $\hat{P}_1$ and $\hat{P}_0$ are observed proportion of events across all visits in the nonexposed and exposed groups, respectively.

In a permutation test, an empirical p-value[23–25] for the hypothesis of H_0:$\beta_1 = 0$ is obtained. Here the permutation is conducted at the level of the subjects but not at the level of the repeated observations so that the correlation structure within the subject is maintained: sicker patients who had relatively more events remain to be sicker patients and healthier patients who had little events remain to be healthier patients in the permuted sample. The distribution of Y under the null hypothesis and, therefore, the variation in the estimated event rate across subjects do not change after permutation. The variation of the parameter estimates which depends on subjects' variation in the permuted sample consequently remains unbiased. The strength of a permutation test is that it requires no assumptions regarding the distribution of $\hat{\beta}_1$. The procedure of the permutation test is described in the following steps (Algorithm 1):

1. Compute for $\hat{\beta}_1 = \hat{P}_1 - \hat{P}_0$;
2. Permute subject id;
3. Calculate β_1 based on the permutated data: $\tilde{\hat{\beta}}_1$;
4. Repeat steps 2 and 3 for N times, where N is a pre-specified large number.

The empirical two-sided $p-\text{value}=\max[P(|\widehat{\tilde{\beta}}_1| \geq |\hat{\beta}_1|), 2P(\widehat{\tilde{\beta}}_1 \geq \hat{\beta}_1|\hat{\beta}_1 \geq 0), 2P(\widehat{\tilde{\beta}}_1 \leq \hat{\beta}_1|\hat{\beta}_1<0)]$, i.e. the larger value between (1) the proportion of times the magnitude of $\widehat{\tilde{\beta}}_1$ is greater than or equal to the magnitude of $\hat{\beta}_1$ and (2) twice the proportion of times that $\widehat{\tilde{\beta}}_1$ is more extreme than $\hat{\beta}_1$. When the event rate is very low (e.g. <1%) the empirical distribution of $\hat{\beta}_1$ may be highly discrete. To address this, the mid *p*-value is sometimes used as an alternative to the traditional empirical *p*-value, in order to better approximate that for a continuous distribution and to be less conservative.[23–25] The mid $p-\text{value}=P(|\widehat{\tilde{\beta}}_1|>|\hat{\beta}_1|)+0.5P(|\widehat{\tilde{\beta}}_1|=|\hat{\beta}_1|)$, i.e. the proportion of times that the magnitude of $\widehat{\tilde{\beta}}_1$ is greater than the magnitude of $\hat{\beta}_1$ plus half the proportion of times the magnitude of $\widehat{\tilde{\beta}}_1$ is equal to the magnitude of $\hat{\beta}_1$.

Note that the above algorithm does not provide a confidence interval for β_1. An apparent option is to obtain a confidence interval based on bootstrapping samples of the original data. However, the bootstrap method does not apply to the rare events case related to the discreteness in the empirical distribution of the parameter estimate. Consider an extreme example, if the exposure group did not experience any events but the nonexposure group did so that the estimate of β_1 is negative. A bootstrapping sample by resampling the subjects will always give a negative estimate of β_1 (i.e. the estimate is bounded by 0) so that the confidence interval based on bootstrap percentiles always excludes 0, leading to a significant test regardless the magnitude of β_1 and the size of the study. More research on methods to obtain an appropriate confidence interval estimate when events are rare is warranted.

2.2 Assessing interactions

To assess statistical interactions and detect differences in exposure–disease associations between two or more strata (e.g. HIV-positive versus HIV-negative women), a linear regression model is defined as follows

$$P(Y_{ij}=1)=\beta_0+\beta_1 x_i+\beta_2 w_i+\beta_3 x_i w_i \quad (2)$$

where w_i indicates the stratum of the *i*th person. Similarly as model (1), this model can be estimated using GEE as described above with a covariate matrix $X_i = (1, x_i, w_i, x_i w_i)^T II_4$ and a parameter vector $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$. Same as model (1), this model can also sometimes run into non-convergence problem related to the estimation of the robust variance also if converged the validity of the statistical inference based on the robust variance estimates has not been examined. Therefore, we propose to use a two-step approach of first obtaining a parameter estimate by ignoring the correlation between repeated observation and then using permutations to obtain an empirical *p*-value. With an independent working correlation and binary x and binary strata w, it can be shown that $\hat{\beta}_3 = (\hat{P}_{11} - \hat{P}_{10}) - (\hat{P}_{01} - \hat{P}_{00})$ where $\hat{P}_{kl}$ is the observed proportion of events across all observations for the *k*th stratum and *l*th exposure group, $k, l = 0,1$.

The permutation test for the interaction effects has rarely been used because it is in general more complex than that for the main effect. One major complication is that it is uncertain whether or not to remove the main effect from the model and how to remove the main effect. Several approaches have been proposed for independent continuous outcome. These include randomization of the residuals with adjustment of main effects,[26] or with adjustment of both the main and interaction effects in the model.[27] An exact permutation test (not to be confused with standard exact methods) was proposed to permute uncorrelated residuals obtained from the transformation of the correlated residuals.[28] A comparison between these methods and the method of simply permuting the observations themselves determined that the exact method performs slightly better when sample sizes are small.[28] Because the exact permutation test requires obtaining a transformation matrix using decomposition of the idempotent matrix, i.e. I-H, where I is an identity matrix and H is the hat matrix, the method is not applicable here as this transformation matrix may not exist under rare events data. Because the method of removal of main effects was shown to be asymptotically exact,[29] in this paper we used the approach of permuting the residuals with appropriate adjustment of the main effects. The procedure is described in the following steps (Algorithm 2):

1. Compute for $\hat{\beta}_3 = (\hat{P}_{11} - \hat{P}_{10}) - (\hat{P}_{01} - \hat{P}_{00})$;
2. Compute for the residual $\hat{e}_{ij}$ for model with only main effects, i.e. $P(Y_{ij} = 1) = \beta_0 + \beta_1 x_i + \beta_2 w_i$ for all i and j;
3. Permute the subject id to obtain permuted residuals;
4. Calculate β_3 based on the permutated residuals: $\hat{\tilde{\beta}}_3 = (\hat{r}_{11} - \hat{r}_{10}) - (\hat{r}_{01} - \hat{r}_{100})$ where $\hat{r}_{kl}$ is the average of permuted residuals for the kth stratum, $k = 0,1$ and lth exposure group, $l = 0,1$;
5. Repeat steps 3 and 4 for N times.

The empirical two-sided p-value is then calculated using the same method for Algorithm 1. Either algorithm, Algorithm 1 or 2, can be easily implemented using existing statistical software. A program written using the R software package is available upon request. In the following section, we use simulations to evaluate the performance of our proposed approach for studying correlated binary data involving uncommon events and compare it with logistic and linear regression GEE models.

3 Simulations

3.1 Main effect

First, we generated a dataset in which there were $n = 400$ subjects with a 30% prevalence of binary exposure variable $X = 1$ and each subject had 10 repeated observations. The sample size and parameters were set to be similar to the data in our example (shown below) as a starting point. The correlated binary data were generated using a beta binomial distribution so that binary outcomes from the same subject share the same rate of events over time. Let P_1 and P_0, denote the overall event rates for $X = 1$ and $X = 0$, respectively, and ρ be the correlation between repeated observations. We set $P_0 = 0.001$ and let P_1 vary from 0.001 to 0.01, representing a rare to uncommon event rate and let ρ vary from 0.2 and 0.5, representing a low to moderate level of correlation. For each simulated data set, we calculated the p-values for the exposure-disease association based on the proposed two-step method as well as the logistic regression GEE model and the linear regression GEE model. For both GEE models, we used the independence working correlation because this assumption leads to a slightly better convergence rate than other working correlation assumptions as it does not require an extra step to estimate the working correlation parameters. The simulation was repeated 1000 times.

For each regression model, we calculated the convergence rate, and determined the proportion of *p*-values below 5% among the simulated datasets in which the model converged. Briefly, the proportion of *p*-values below 5% when $P_1 = P_0 = 0.001$ (i.e. $\beta_1 = P_1 - P_0 = 0$) provides the "*empirical level of significance*" where the nominal level is 5%, i.e. considered as an indication of validity of the test; conversely, when $P_1 \neq P_0$ (i.e. $\beta_1 \neq 0$) the proportion below 5% indicates the "*empirical power*". We also determined the proportion of *p*-value below 5% among all datasets treating nonconvergence as a failure to demonstrate a significant exposure effect. This latter definition is useful, since in practice if a model fails to converge there is often no further investigation, leading to a possible false negative result.

We also calculated and compared the bias in effect estimates for the logistic regression GEE models and the linear regression GEE models that converged and the bias in $\hat{\beta}_1$ using the proposed method. Bias is expressed as the percentage change from the known value in the simulation. Note that although the linear regression GEE model under independent working correlation gives the same effect estimates as the proposed method, the linear regression GEE model can sometimes run into non-convergence problem related to the estimation of the robust standard error when $\hat{\beta}_1 = 0$ in the simulated data sets. When the portion of nonconvergence is nonignorable, the omission of data sets with $\hat{\beta}_1 = 0$ can bias the average estimates when the true $\beta_1 \neq 0$. Therefore, there are some differences in average bias estimates between the linear regression GEE models that converged and the proposed method in Tables 1 and 2.

As shown in Table 1, the linear regression GEE model had a much higher convergence rate than the logistic regression GEE model. While both types of regression models, as well as the proposed two-step approach, were each shown to be valid; the regression methods were too conservative, particularly according to the empirical level of significance among all simulated data sets. The permutation test had, as expected, the greatest empirical power as well as lowest bias under a range of effect sizes that the simulations incorporated. For example, when $\rho = 0.2$ and $OR = 4.0$ for logistic regression ($\beta_1 = 0.003$ for linear regression), the empirical statistical power was 4.8% for the linear regression model, 14.8% (20.2% among converged) for the logistic regression model and 31.0% for the permutation test; the relative bias was −12.6% for the logistic regression model and 1.9% for the linear regression model and the proposed method; when $OR = 1$, the logistic regression model had a bias of 185% while the proposed method almost had zero bias. This finding that the two-step approach achieves the highest statistical power and has the least bias was true regardless of level of correlations ($\rho = 0.2$ or 0.5).

3.2 Interaction effect

Next, we evaluated the procedure for assessing interaction by a binary stratum variable (w). We assumed 200 subjects with $w = 0$ and $P(X = 1) = 40\%$; 400 subjects with $w = 1$ and $P(X = 1) = 30\%$. Again, these numbers were chosen to be similar to those in the example. The event rates were set to be: $P_{00} = 0.0005$, $P_{01} = 0.001$, $P_{10} = 0.001$ and P_{11} varying from 0.0015 to 0.0205, where P_{kl} is the proportion of events for the kth stratum and lth exposure group, k, $l = 0,1$ so that the difference in exposure effects between w = 1 and w = 0, i.e. $\beta_3 = (P_{11} - P_{10}) - (P_{01} - P_{00})$ varies from 0 to 0.019. The correlation between repeated observations ρ was set to be 0.2 and 0.5. Model convergence, empirical significance/power and bias were evaluated. Table 2 shows that there was little convergence of the logistic regression GEE models, whereas most linear regression models converged. Similar to the first set of simulations, bias was low for the proposed method in estimating the interaction effects (<7%), and much higher for logistic regression. Further, the permutation test was the most powerful test.

3.3 A larger sample size

To examine if the findings above persist with a larger sample size, we increased the sample size to be $n = 1000$ subjects for the main effect and $n = 1200$ for the interactive effect while the other parameters remained the same. Table 3 indicates that with a larger n, the issue of lack of convergence, inaccuracy and low statistical power remained particularly in the logistic GEE models but the severity of the issue reduces sometimes significantly as the number of events increases. This observation is especially true for the model with a main effect. For example, when $P_1 = P_0 = 0.001$ so that the expected number of events is 10, there were about 25% of the data sets for which the logistic regression GEE models failed to achieve convergence and the bias was large. But when $P_0 = 0.001$ and let $P_1 = 0.004$ so that the expected number of events is about 20, almost all the logistic regression GEE models converged and the bias was low and the statistical power was comparable to what observed for the proposed two-step method. For the model with an interactive effect, although the convergence and accuracy was also greatly improved with a larger number of subjects, it is yet far from satisfactory. For example, even when the ratio of two odds ratios becomes 10.4 and the expected number of events is 15, there was about 60% of the logistic regression model failed to converge and the bias was still greater than 5% and the statistical power continued to be much lower than that from the proposed two-step approach. The performance of the linear regression GEE model, on the other hand, is close to that of the proposed two-step approach as the number of events increases. Whether or not a “threshold” on the number of total events exists in order for the logistic regression GEE approach to achieve acceptable performance needs to be further investigated.

It should be emphasized that the proposed method outperforms the logistic regression GEE and the linear regression GEE models when the events are rare, as we demonstrated in our simulations. But when the events are not rare, such superiority of the proposed method no longer exists. We did another set of simulations (result not shown) which indicated that when events are not rare, the logistic regression GEE models and the proposed method perform similarly while the linear regression GEE models tend to have a slightly lower statistical power.

4 Motivating example

We applied the proposed method to examine the association between cervical HSIL and HLA genotype in HIV-positive and HIV-negative women. HLA genes are among the most variable in the human genome and the encoded HLA proteins play a central role in the adaptive T-cell responses to viral infections. It was therefore hypothesized that an association between HLA and HSIL would be observed in HIV-negative but not as much in HIV-positive women, since immunogenetic factors would have less of a biologic impact in broadly immunocompromised individuals. Consequently, we were also interested in examining the possible interaction between HLA genes and HIV status. The study was based in a longitudinal cohort called the Women's Interagency HIV Study (WIHS), which enrolled 2793 HIV-positive and 975 HIV-negative women during two enrollment periods, in 1994 and again in 2001.[30–33] At each semi-annual visit, DNA of human papillomavirus (HPV), the virus that causes cervical cancer, was detected using a well-established and highly sensitive polymerase chain reaction (PCR) assay, and Pap tests were conducted. High-resolution HLA class I and II genotyping[34] was conducted in a stratified random sample of 830 women in the WIHS cohort based on HIV status and CD4 levels. At the time of the study, the women had completed 15 semi-annual visits. Overall, there were 512 HIV-positive and 285 HIV-negative women with a total of 3682 visits and 2400 visits, respectively. Here we focus on two HLA alleles that were previously examined to illustrate the method: DRB*15:01 (denoted by Drb1501) and Bw4;[33] the former was previously reported increase and the latter decrease risk of cervical HSIL. The event rate of any HSIL was 10%. However, HSIL containing specific HPV genotype was uncommon. HPV 16 is the most important cancer-related HPV type and HPV 18 is the second most important cancer-related HPV type. In this dataset, the event rate of HSIL containing HPV16 (HPV16HSIL) was less than 3% while that for the HSIL containing HPV18 (HPV18HSIL) was less than 1%. In this paper, we focused on HPV18HSIL as the endpoint because in our original analysis when the conventional statistical methods were applied to examine the association of HLA genotypes with the occurrence of HSIL containing either HPV16 or HPV 18, the analysis on HPV18HSIL had the most issue with model convergence.

First, we estimated the level of correlation between repeated observations based on a beta-binomial model for the binary outcomes, under the null hypothesis of no exposure and disease association. We used the method of moment to obtain an estimate of correlation, specifically,

$$\frac{\sum_{i}\sum_{j<k} Y_{ij}Y_{ik}/\sum_{i}\sum_{j}\frac{J_i!}{2!(J_i-2)!} - (\sum_{i}\sum_{j}Y_{ij}/\sum_{i}J_i)^2}{\sum_{i}\sum_{j}Y_{ij}/\sum_{i}J_i(1-\sum_{i}\sum_{j}Y_{ij}/\sum_{i}J_i)}$$

so that the correlation between repeated observations from the same subject was estimated to be 0.2, a small but nonignorable level of correlation between repeated events in this dataset. Next, we conducted analyses of HLA genotype and its relation with HPV18HSIL using (i) logistic regression GEE model (i.e. the traditional approach), (ii) linear regression GEE model, and (iii) the proposed two-step approach. We also assessed whether these associations differed by HIV status.

As shown in Table 4, the logistic regression GEE models for DRB1501 and its association with HPV18HSIL failed to converge for either HIV-positive or HIV-negative women or the interaction model. In contrast, the linear regression GEE model converged and showed a decreased risk of HPV18HSIL of 0.36% related to DRB1501 among HIV-positives, but an increased risk of 0.67% among HIV-negatives. While neither of these associations were statistically significant based on the linear regression GEE model, the permutation test showed the association in HIV-negative women to be statistically significant ($p = 0.047$), whereas the result in HIV-positive women did not approximate significance ($p = 0.882$). The interaction term was of borderline significance in the permutation test ($p = 0.071$).

For allele Bw4, the logistic regression GEE model did not converge in HIV-negative women, nor in the analysis of the interaction effect, whereas all models using linear regression converged. Furthermore, the proposed permutation test but not the linear regression GEE model showed that there was an inverse association of Bw4 with HPV18HSIL that was significant among HIV-negative women ($p = 0.025$).

For both the alleles, the two-step approach but not the two GEE models suggested a stronger association with HPV18HSIL in HIV-negative women than HIV-positive women, supporting our hypothesis that HLA genotypes have less of a biologic impact in immunocompromised individuals (HIV-positive women).

It is worth noting that even though in the simulations the empirical statistical power is uniformly higher for the two-step method, it does not imply that the p-values obtained in any data set such as in this example will be invariably smaller for the two-step method. Instead, the finding from the simulations suggests that when there is a real effect, the two-step method on the average is more likely to achieve statistical significance than the other two methods.

5 Conclusion and discussion

This paper describes a statistical approach for measuring associations between risk factors and rare or uncommon events that are subject to recurrence—often correlated endpoints. As we demonstrated in the simulations, traditional methods such as GEE models in particular the logistic regression GEE models in the case of rare events data can no longer provide valid statistical inference because they either fail to converge or their parameter estimates lose consistency as well as asymptotical normality. The proposed statistical approach involves two steps: (i) a linear (additive) regression model to estimate the strength of the association, and (ii) a permutation test to estimate the statistical significance of the association. The permutation was conducted at the level of the subjects but not at the level of the repeated observations so that the correlation structure within the subject is maintained. Because the permutation test, unlike multiplicative or additive GEE models, does not need to estimate a robust variance, our two-step approach avoids the problem of model convergence. Further, the permutation test does not require a normality assumption for parameter estimates which may violate under rare or uncommon events and therefore achieves a much higher statistical power than is possible using either logistic or linear regression GEE models. The proposed approach is also easy to implement using widely available statistical software.

Overall, the proposed statistical method represents an important advance in measuring risk factor associations with rare or uncommon but correlated binary outcomes. There is a misconception in the literature that rare event problems only occur when sample sizes are small or when there are too many strata and there is a tendency in practice to dismiss rare events data because it is not informative. In the simulation study, we considered up to over 1200 subjects with over 10,000 repeated visits and in our example, we have about 800 subjects with 6000 repeated visits and only two strata (i.e. HIV positive and negative) yet rare event problems occurred. Therefore, we hope this paper emphasized that rare event problems do occur even in a large data set without many strata and rare or uncommon events do not imply lack of information. In fact, important exposure and disease associations can be identified from rare/uncommon events data as we have illustrated in our example.

In the current paper we used a study of incident HSIL and a genetic risk factor as the primary example of this type of data, but the considerations raised in this paper potentially affect the analysis of many longitudinal cohorts, and in the Introduction we discussed other examples. The application of the proposed method will grow as the use of genetic/epi-genetic assays and other new technologies increasingly involve the analysis of exposures (e.g. genotypes) that are also binary and rare or uncommon, also leading to separation problems. For example, an important application of the proposed method to genetic/epi-genetic data is to examine the risk of a single binary outcome in association with several binary genetic variables of interest that may be rare/uncommon and correlated with one another. The proposed method can be potentially extended to this situation through the use of the correlated exposures as the outcome variable and the single binary outcome as the exposure variable. Additional research is also warranted to extend the method to incorporate continuous exposure variables and multiple confounders, as well as to develop methods to estimate confidence intervals for each risk estimate.

Acknowledgments

Funding

This work was supported by National Cancer Institute [5R01AI057006-04 and 5P30-CA013330].

References

1. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13–22.
2. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika. 1984; 71:1–10.
3. Schaefer RL. Bias correction in maximum likelihood logistic regression. Stat Med. 1983; 2:71–78. [PubMed: 6648121]
4. Cordeiro GM, McCullagh P. Bias correction in generalized linear models. J Roy Stat Soc Ser B. 1991; 53:629–643.

5. Bull SB, Greenwood CMT, Hauck WW. Jackknife bias reduction for polychotomous logistic regression. Stat Med. 1997; 16:545–560. [PubMed: 9089962]

6. Cordeiro GM, Cribari-Neto F. On bias reduction in exponential and non-exponential family regression models. Commun Stat. 1998; 27:485–500.

7. Leung DH-Y, Wang YG. Bias reduction using stochastic approximation. Aust NZ J Stat. 1998; 40:43–52.

8. Anderson JA, Blair V. Penalized maximum likelihood estimation in logistic regression and discrimination. Biometrika. 1982; 69:123–136.

9. Self SG, Liang K-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc. 1987; 82:605–610.

10. Agresti A. Exact inference for categorical data: recent advances and continuing controversies. Stat Med. 2001; 20:2709–2722. [PubMed: 11523078]

11. Potter DM. A permutation test for inference in logistic regression with small- and moderate sized data sets. Stat Med. 2005; 24:693–708. [PubMed: 15515134]

12. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat Med. 2002; 21:2409–2419. [PubMed: 12210625]

13. Firth, D. Generalized linear models and Jeffreys priors: An iterative weighted least-squares approach. In: Dodge, Y., Whittaker, J., editors. Computational statistics. Heidelberg: Physica-Verlag; 1992. p. 553-557.

14. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993; 80:27–38.

15. Tang M-L, Tang N-S, Rosner B. Statistical inference for correlated data in ophthalmologic studies. Stat Med. 2006; 25:2772–2783.

16. Hungsberger S, Graubard BI, Korn EL. Testing logistic regression coefficients with clustered data and few positive outcomes. Stat Med. 2008; 27:1305–1324. [PubMed: 17705348]

17. Corcoran C, Ryan L, Senchaudhuri P, et al. An exact trend test for correlated binary data. Biometrics. 2001; 57:941–948. [PubMed: 11550948]

18. Molenberghs G, Ryan LM. An exponential family model for clustered multivariate binary data. Environmetrics. 1999; 10:279–300.

19. Ryan, L., Molenberghs, G. Statistical methods for developmental toxicity: Analysis of clustered multivariate binary data. In: Bailer, AJ.Maltoni, C., Bailar, JC., editors. Uncertainty in the risk assessment of environmental and occupational hazards: An international workshop (Annals of the New York Academy of Sciences). New York: New York Academy of Sciences; 1999. p. 196-211.

20. Mehta CR, Patel NR, Senchaudhuri P. Exact stratified linear rank tests for ordered categorical and binary data. J Comput Graph Stat. 1992; 1:21–40.

21. Gelman A, Jakulin A, Pittau MG, et al. A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat. 2008; 2:1360–1383.

22. Abrahantes JC, Aerts M. A solution to separation for clustered binary data. Stat Model. 2012; 12:3–27.

23. Lancaster HO. Significance tests in discrete distributions. J Am Stat Assoc. 1961; 56:223–234.

24. Bernard GA. Must clinical trials be large? The interpretation of p-values and the combination of test results. Stat Med. 1990; 9:601–614. [PubMed: 2218164]

25. Hirji KF, Tan S-J, Elashoff RM. A quasi-exact test for comparing two binomial proportions. Stat Med. 1990; 9:601–614. [PubMed: 2218164]

26. Still AW, White AP. The approximate randomization test as an alternative to the F-test in the analysis of variance. Br J Math Stat Psychol. 1981; 34:243–252.

27. Jockel, KH., editor. Bootstrapping and related techniques. Berlin: Springer; 1992. Ter Braak CJF. Permutation versus bootstrap significance tests in multiple regression and ANOVA; p. 79-86.

28. Jung BC, Jhun M, Song SH. A new random permutation test in ANOVA models. Stat Papers. 2007; 48:47–62.

29. Good P. Extension of the concept of exchangeability and their applications. J Modern Appl Stat Meth. 2002; 1:243–247.

30. Barkan SE, Melnick SL, Preston-Martin S, et al. The Women's Interagency HIV Study. WIHS Collaborative Study Group. Epidemiology. 1998; 9:117–125. [PubMed: 9504278]

31. Strickler HD, Palefsky JM, Shah KV, et al. Human papillomavirus 16 and immune status in human immunodeficiency virus-seropositive women. J Natl Cancer Inst. 2003; 95:1062–1071. [PubMed: 12865452]

32. Strickler HD, Burk RD, Fazzari M, et al. HPV natural history and possible HPV reactivation in HIV-positive women. J Natl Cancer Inst. 2005; 97:577–586. [PubMed: 15840880]

33. Xue X, Gange SJ, Zhong Y, et al. Marginal and mixed effects models in the analysis of HPV natural history data. CEBP. 2010; 19:159–169.

34. Kuniholm MH, Gao X, Xue X, et al. The relation of HLA genotype to hepatitis C viral load and markers of liver fibrosis in HIV-infected and HIV-uninfected women. J Infect Dis. 2013; 12:1807–1814.

Table 1

Comparison of logistic regression generalized estimating equation (GEE) models and linear regression GEE models and the proposed two-step method in detection and estimation of the main allele effect and comparison of convergence rate between logistic and linear regression GEE models. Data were generated using different assumptions on effect size ($\beta_1 = P_1 - P_0$ from model (1) and its corresponding OR) and correlation among repeated events (ρ) as specified in the table where n = 400 and the event rate for non-exposed group $P_0 = 0.001$ and the proportion of exposure $P(X = 1) = 30\%$ based on 1000 simulations.

	Binary logistic regression GEE					Linear regression GEE					Proposed method	
	OR^a	**Prop[b] conv**	**Prop[c] sig among conv**	**Prop[c] sig among total**	**Bias[d] in log $\hat{OR}$**	**β_1**	**Prop conv**	**Prop sig among conv**	**Prop sig among total**	**Bias in $\hat{\beta}_1$[e]**	**Prop sig perm test**	**Bias in $\hat{\beta}_1$**
ρ = 0.2	1.0	39%	2.5%	1.0%	185%	0.000	88%	1.4%	1.2%	0.0%	3.8%	0.0%
	4.0	73%	20.2%	14.8%	−12.6%	0.003	99%	4.8%	4.8%	1.9%	31.0%	1.9%
	10.0	80%	63.4%	50.6%	−5.5%	0.009	100%	45.1%	45.0%	−3.8%	75.6%	−3.8%
ρ = 0.5	1.0	17%	6.0%	1.0%	289%	0.000	68%	0.3%	0.2%	−0.1%	1.8%	−0.1%
	4.0	44%	22.4%	9.8%	−22.9%	0.003	90%	0.9%	0.8%	7.6%	19.0%	−2.7%
	10.0	52%	45.0%	23.4%	−13.4%	0.009	99%	11.1%	11.0%	−1.4%	60.4%	−1.4%

[a] For the logistic regression model, the effect size is measured by odds ratio (OR) and for the linear regression model, the effect size is measured by the difference in proportion of events between two allele groups (β_1).

[b] For each regression model, we report the proportion of simulations which the model achieved convergence.

[c] For each regression model, we report the proportion of rejecting the null hypothesis of no allele and disease association at the nominal statistical significance level of 0.05 among the simulations which the model was converged as well as the proportion of rejection the null hypothesis of no allele and disease association among all simulations, while treating non-convergence as a failure to demonstrate a significant exposure effect; for the permutation test, we show the proportion of rejecting the null hypothesis among all simulations; note when the null hypothesis is true, the proportion of rejecting the null hypothesis reflects the empirical significance value and when the null hypothesis is not true, the proportion of rejecting the null hypothesis is the empirical power.

[d] For each regression model and the proposed approach, the bias of each parameter estimate was calculated as (estimated value-true value)/true value *100%; when the true effect is null, the bias = (exp(estimated value) − exp(true value))/exp(true value) *100%.

[e] Although $\hat{\beta}_1$ from the linear the linear regression GEE model with an independent working correlation equals to the estimate of β_1 from the proposed method, there is some difference in bias estimates because some linear regression GEE models failed to converge.

prop = proportion, conv = convergence.

Table 2

Comparison of logistic regression generalized estimating equation (GEE) models and linear regression GEE models and the proposed two-step method in detection and estimation of difference in allele effect over strata (the interaction effect) and comparison of convergence rate between logistic and linear regression GEE models. Data were generated using different assumptions on effect size ($\beta_3 = (P_{11} - P_{10}) - (P_{01} - P_{00})$) from model (2) and the corresponding OR_1/OR_0) and correlation among repeated events (ρ) as specified in the table where event rates for exposed and nonexposed are P_{10} = 0.001 and P_{11}, respectively for stratum 1 with n = 200 clusters and the proportion of exposure $P(X = 1)$ = 40%; event rates for exposed and nonexposed are P_{00} = 0.0005 and P_{01} = 0.001, respectively for stratum 0 with n = 400 clusters and $P(X = 1)$ = 30% based on 1000 simulations.

	Binary logistic regression GEE					Linear regression GEE					Proposed method	
	OR_1/OR_0[a]	Prop[b] conv	Prop[c] sig among conv	Prop sig among total[c]	Bias[d] in $\log(\hat{OR}_1/\hat{OR}_0)$	β_3	Prop conv	Prop sig among conv	Prop sig among total	Bias in $\hat{\beta}_3$	Prop sig perm test	Bias in $\hat{\beta}_3$
ρ = 0.2	0.75^5	7%	1.5%	0.1%	150%	0.000[e]	94%	0.5%	0.5%	0.1%	5.1%	0.1%
	2.26	10%	4.2%	0.4%	−107%	0.003	99%	4.7%	4.6%	4.7%	26.0%	4.7%
	5.81	11%	3.5%	0.4%	−36.0%	0.010	100%	27.9%	27.8%	−2.6%	64.6%	−2.6%
	10.4	13%	6.3%	0.8%	16.4%	0.019	100%	71.0%	71.0%	1.3%	91.2%	1.3%
ρ = 0.5	0.75	1%	0.0%	0.0%	317%	0.000	78%	0.0%	0.0%	0.0%	3.6%	0.0%
	2.26	2%	0.0%	0.0%	−108%	0.003	85%	0.2%	0.2%	14.1%	15.4%	−2.8%
	5.81	2%	0.0%	0.0%	−40.6%	0.010	98%	7.4%	7.2%	6.7%	53.1%	6.7%
	10.4	3%	5.9%	0.2%	−40.2%	0.019	100%	32.1%	32.0%	−1.4%	76.8%	−1.4%

[a] For the logistic regression model, the effect size is measured by ratio of odds ratio (OR_1/OR_0) and for the linear regression model, the effect size is measured by the difference in allele effect between two strata.

[b] For each regression model, we also show the proportion of simulated datasets for which the model achieved convergence.

[c] For each regression model, we show the proportion of rejecting the null hypothesis of no differential allele effects among the simulated datasets for which the model was converged as well as the proportion of rejection the null hypothesis of no differential allele effects at the nominal statistical significance level of 0.05 among all simulated datasets, while treating nonconvergence as a failure to demonstrate a significant interaction; for the proposed method, we show the proportion of rejecting the null hypothesis among all simulations; note when the null hypothesis is true, the proportion of rejecting the null hypothesis reflects the empirical significance value and when the null hypothesis is not true, the proportion of rejecting the null hypothesis is the empirical power.

[d] For each regression model and the propose approach, the bias of each parameter estimate was calculated as (estimated value-true value)/true value *100%; when the true effect is null, the bias = (exp(estimated value) − exp(true value))/exp(true value) *100%.

[e] The null interaction in the linear regression model does not correspond to a null interaction in the logistic regression model.

prop = proportion, conv =convergence.

Table 3

Under a larger sample size the comparison of logistic regression generalized estimating equation (GEE) models and linear regression GEE models and the proposed two-step method in detection and estimation of allele effect (n = 1000 clusters) and the difference in allele effect over strata (the interaction effect) (N=1200 clusters) and comparison of convergence rate between logistic and linear regression GEE models with a larger sample size. Data were generated using different assumptions on effect size ($\beta_1 = P_1 - P_0$ from model (1) and its corresponding OR and $\beta_3 = (P_{11} - P_{10}) - (P_{01} - P_{00})$) from model (2) and the corresponding ratio OR_1/OR_0) and correlation among repeated events (ρ) was set to be 0.2. For details in parameter values, see descriptions in Tables 1 and 2. The simulations were repeated 1000 times.

	Binary logistic regression GEE					Linear regression GEE					Proposed method	
	True parameter	Prop conv	Prop sig among conv	Prop sig among total	Bias	β_3	Prop conv	Prop sig among conv	Prop sig among total	Bias	Prop sig perm test	Bias
OR	1.0	76%	4.7%	3.6%	102%	0.000	100%	6.0%	6.0%	0.0%	3.0%	0.0%
	4.0	97%	46.4%	45.0%	4.4%	0.003	99%	28.4%	28.4%	1.0%	49.6%	1.0%
OR_1/OR_0	0.75	26%	3.8%	1.0%	153%	0.000	99%	1.8%	1.8%	0.0%	5.8%	0.0%
	5.81	42%	12.7%	5.4%	−17.4%	0.009	100%	65.0%	65.0%	−3.9%	83.6%	−3.9%
	10.4	43%	33.2%	14.2%	−6.0%	0.019	100%	96.2%	96.2%	2.2%	99.0%	2.2%

Table 4

The association between two alleles and HPV18HSIL and their interactions with HIV status using different methods for the example data set with 512 HIV-positive and 285 HIV-negative women with a total of 3682 visits and 2400 visits, respectively with complete information on HPV18HSIL.

		Logistic regression GEE		**Linear regression GEE**		**Proposed two-step method**	
Allele	**Stratum**	**$\hat{OR}$ (or OR_1/OR_0) & 95% CI[a]**	***p***	**$\hat{\beta}_1$ (or $\hat{\beta}_3$) and 95% CI**	***p***	**$\hat{\beta}_1$ (or $\hat{\beta}_3$)**	***p***
DRB1501	HIV−	–	–	0.0067 (−0.0061, 0.0194)	0.3042	0.0067	0.0465
	HIV+	–	–	−0.0036 (−0.0075, 0.0003)	0.0681	−0.0036	0.8820
	Interaction	–	–	−0.0103 (−0.0241, 0.0035)	0.1445	−0.0103	0.0710
BW4	HIV−	–	–	−0.0015 (−.0044, 0.0014)	0.3143	−0.0015	0.0245
	HIV+	1.771 (0.3277, 9.574)	0.5067	0.0025 (−0.0043, 0.0093)	0.4655	0.0025	0.6800
	Interaction	–	–	0.0040 (−0.0034, 0.0114)	0.2846	0.0040	0.4460

"–" indicates model was not converged.

[a] We provided confidence interval (CI) estimates for the parameter of interest when the GEE models were converged. However, one should be cautious in interpreting these CIs because the asymptotical consistency and normality may not be valid under rare events even if the models are converged, as we demonstrated in the simulations.