Published in final edited form as:

Stat Med. 2010 April 30; 29(9): 1004–1013. doi:10.1002/sim.3833.

A polytomous conditional likelihood approach for combining matched and unmatched case-control studies

Mulugeta Gebregziabher^{1,*}, Paulo Guimaraes¹, Wendy Cozen², and David V. Conti²

¹Medical University of South Carolina Department of Biostatistics, Bioinformatics and Epidemiology 135 Cannon St., Charleston Suite 303, SC 29425, Fax: 843-876-1126, Tel: 843-876-1112, gebregz@musc.edu

²University of Southern California Department of Preventive Medicine, Los Angeles, CA 90089-9175

SUMMARY

In genetic association studies it is becoming increasingly imperative to have large sample sizes to identify and replicate genetic effects. To achieve these sample sizes, many research initiatives are encouraging the collaboration and combination of several existing matched and unmatched case-control studies. Thus, it is becoming more common to compare multiple sets of controls with the same case group or multiple case groups to validate or confirm a positive or negative finding. Usually, a naive approach of fitting separate models for each case-control comparison is used to make inference about disease-exposure association. But, this approach does not make use of all the observed data and hence could lead to inconsistent results. The problem is compounded when a common case group is used in each case-control comparison. An alternative to fitting separate models is to use a polytomous logistic model but, this model does not combine matched and unmatched case-control data. Thus, we propose a polytomous logistic regression approach based on a latent group indicator and a conditional likelihood to do a combined analysis of matched and unmatched case-control data. We use simulation studies to evaluate the performance of the proposed method and a case-control study of multiple myeloma and Inter-Leukin-6 as an example. Our results indicate that the proposed method leads to a more efficient homogeneity test and a pooled estimate with smaller standard error.

Keywords

case-control; latent group indicator; polytomous conditional logistic; pooled estimate

1. INTRODUCTION

Case-control studies are the most widely used study design in chronic disease epidemiology [1]. Especially, in situations where experimental studies are not feasible, both matched and unmatched case-control study designs have proven to provide crucial causal links between disease and exposure [2]. In fact, with modest or even small genetic effects, it is becoming increasingly imperative to have large sample sizes for discovery and replication. Current trends favor large collaborative research programs (consortiums) as a way to achieve these sample sizes in order to detect small genetic effects, to confirm a positive or negative finding, to assess selection bias, or to compare relative risk estimators for the same risk

factors studied in different phases of a disease. These consortiums often consist of both matched and unmatched studies and include comparisons of two or more different sets of controls to either a single set or multiple sets of cases. A good example is the case-combined-control design [3,4]. The analysis usually involves fitting separate models for each case-control comparison, testing the homogeneity of the parameter estimates and, if appropriate, obtaining a pooled estimate. However, to obtain a pooled estimate there are two problems that need to be addressed: 1) how to account for the correlation (ρ) between study-specific estimates when testing for homogeneity and 2) how to obtain an efficient pooled estimate from a combined analysis of matched and unmatched studies?

For example, in a case-control study of the risk of multiple myeloma and several polymorphisms of the Inter-Leukin-6 (IL-6) genes, two groups of controls (family controls and population controls) were compared to a single set of cases [5]. Separate analysis were made for each comparison. While conceptually simple and attractive, fitting separate models tends to lead to estimates with underestimated variance (if $\rho > 0$) and hence a homogeneity test with inflated type-I-error rate or overestimated variance (if $\rho < 0$) and hence a homogeneity test with inflated type-II-error rate. As we will see, a more sound conclusion is obtained from a combined analysis.

One approach to combining matched and unmatched case-control data sets is the adaptation in [6] of the Mantel-Haenszel method leading to a pooled estimate obtained as a weighted average of the parameter estimates from the matched and unmatched case-control studies. However, this approach assumes homogeneity of parameters and can only be used when the number of covariate patterns is small [7]. A related method based on the theory of estimating the common mean of two populations was recently suggested [8]. Moreno et al [9] proposed a likelihood approach for estimating a pooled common parameter from a direct product of two likelihoods, one for the matched dataset and another for the unmatched dataset. Subsequently, Huberman and Langholz [10] showed that this approach can be easily implemented with standard software by rearranging the data from the unmatched study to look like a matched study by creating “virtual partners” for the cases and controls. With appropriately generated “virtual partners”, a conditional logistic likelihood analysis is performed on the complete data set for valid inference.

In situations where a comparison of one or more case-sets to one or more control groups is of interest, one approach is to use a single multinomial or polytomous logistic regression model [11,12,13]. This approach has the advantage of using all available data for inference and estimation purposes. However, it is restricted to situations where the cases and controls arise from either a matched or unmatched design. To our knowledge there are no available methods that can handle a combined analysis of matched and unmatched case-control studies in a polytomous logistic regression framework.

Here, we propose a polytomous conditional likelihood method that allows for comparison of one or more case-sets with one or more control sets in a pooled study of several case-control studies. This proposed method extends and generalizes the virtual partner idea in [10] to a polytomous logistic regression setting so that it is possible to make comparison of multiple control groups to a single case group. This allows for simple and efficient pooled estimation and corresponding homogeneity tests. We demonstrate the application of this proposed method by using simulated as well as real life data. Since pooling studies can lead to a biased pooled estimate when the assumption of homogeneity is incorrect, we emphasize a power comparison of different homogeneity tests - especially when analyzing small sample sizes [11]. Furthermore, we investigate via simulation the relative performance of the different approaches in terms of bias and efficiency.

2. DATA EXAMPLE

The proposed method is motivated by a real data example from a case-control study of multiple myeloma risk and several genetic polymorphisms within the IL-6 region. This study has two different sets of controls, relative controls and population controls. The details of the study are reported elsewhere [5]. Briefly, cases were 150 residents of Los Angeles County diagnosed with primary multiple myeloma or plasmacytoma (ICD-03 9731–9734) from October 1, 1999 through December 31, 2002, and were ascertained by the University of Southern California Cancer Surveillance Program (USC-CSP) and the population-based cancer registry for Los Angeles County. Two groups of controls were recruited. The first group consisted of 117 relatives of cases (family controls) and the second group consisted of 126 population-controls frequency matched to an expected race-, age-, and sex-distribution. DNA was extracted from case and relative (family) control samples and at least one single nucleotide polymorphism (SNP) successfully genotyped from 150/150 cases, 112/117 family controls, and 126/126 population controls. Five SNPs (–174 GC, IL-6-*ra*, –572GC, –597GA and –373AnTn) were genotyped. We focus on the IL-6-572 SNP which has a genotype distribution of GG:101,85,90; GC:29,17,15; CC:8,5,15 for case, relative control and population control respectively.

In this study, a SNP in the promoter gene (IL-6-572) at position –572 showed association with the risk of multiple myeloma in the comparison of cases to population controls ($\hat{\eta}_2$ and 95% CI = (1.2,4.7)), but a non-significant positive association for the comparison of cases to relative controls [$\hat{\eta}_1$ and 95% CI = (0.7,4.7)], where η denotes the odds ratio. Our objective is to efficiently use all the data from both the matched and the unmatched controls to test whether to pool the data and to obtain a common pooled estimate with its corresponding variance. We also want the test to be able to account for the correlation in estimates due to the use of the same case set. In the next section, we review and demonstrate this using existing methods and propose a unified and more efficient method.

3. METHODS

For data sets with multiple case sets matched and unmatched to multiple control sets, we explore four different analysis options: (1) a naive approach that analyzes each case-control comparison independently, (2) a break the matching approach (BMM) that uses a unconditional polytomous regression framework, (3) a binary latent group indicator approach (BLGIM) that combines matched and unmatched case-control studies, and (4) polytomous latent group indicator (PLGIM) that combines matched and unmatched correlated case-control comparisons. Each approach yields separate comparison-specific estimates, a pooled estimate (exception is the naive analysis), and a corresponding test of heterogeneity. Both the naive and BLGIM tests of heterogeneity are based on the assumption of independence between $\hat{\beta}_1$ and $\hat{\beta}_2$ which are the natural logs of $\hat{\eta}_1$ and $\hat{\eta}_2$, respectively. The tests corresponding to BMM and PLGIM are tests of an interaction term for X and a study indicator W ($W = 1$ if matched study, $W = 0$ if unmatched study) and they can account for any covariation in $\hat{\beta}_1$ and $\hat{\beta}_2$.

3.1. Naive Approach

In situations where there is a common case group (as in our data example) and two control groups (one matched or related and the other unmatched or unrelated), the conventional approach involves analyzing the data twice, using unconditional logistic for the case-unmatched control comparison and conditional logistic for the case-matched control comparison. This is often followed by a test of homogeneity (Wald test of $H_0: \eta_1 = \eta_2$) that assumes independence between the two estimates. While there are relatively straightforward

ad hoc approaches to obtain a pooled estimate, such as an inverse-variance weighted average of the two estimates, its variance is usually either under-estimated or over-estimated due to the naive assumption of independence between $\hat{\eta}_1$ and $\hat{\eta}_2$.

3.2. Break the Matching Method

Another approach for combining matched and unmatched case-control studies is to break the matching (BMM) and consider the response as a polytomous outcome. This results in a polytomous logistic regression analysis of the data (with a case set and two control groups) using an unconditional likelihood. The likelihood [11,13] for $J \geq 2$ groups is given by,

$$L(\alpha_1, \dots, \alpha_J, \beta_1, \dots, \beta_J) = \prod_{i=1}^N \prod_{j=0}^J \pi_j(\mathbf{X}_i)^{D_{ij}}, \quad (1)$$

where, N is the total number of subjects, $\mathbf{X}$ is a covariate vector and D_{ij} is case-control indicator for the i th subject in the $j = 0, \dots, J$ th group ($j = 0$ is the referent group) which takes a value of one if a subject is from group j and 0 otherwise. And

$\pi_j(\mathbf{X}_i) = \exp(\alpha_j + \beta_j' \mathbf{X}_i) / (1 + \exp(\alpha_j + \beta_j' \mathbf{X}_i))$, where $\alpha_0 = \beta_0 = 0$. However, this approach ignores the matching and hence could lead to biased estimates in the analysis of individually matched case-control data [16]. Partial control of the bias is possible by adjusting for the matching factors in the analysis. Though plausible, this method doesn't always work, even though it often gives reasonable answers [17]. In this approach homogeneity is usually assessed using a Wald test that is based on a variance estimate that could possibly be biased by ignoring the matching. A pooled estimate (which could be biased) can be obtained by analyzing the case-combined-control data set using unconditional logistic regression [3].

3.3. Binary Latent Group Indicator Method

The binary latent group indicator method (BLGIM) [9,10] was originally proposed for combining separate (matched and unmatched) case-control studies (i.e., with different cases in the different studies). However, as we will see, it can be modified to accommodate the situation in our data example. This approach is based on a likelihood that is obtained as the product of a conditional likelihood for matched data and an unconditional likelihood for unmatched data. Huberman and Langholz (1999) proposed a computational trick that makes use of conditional logistic likelihood to implement the approach suggested by Moreno et al (1996). The idea consists of adding a “virtual partner” of the opposite status in the unmatched study with covariates of the “virtual partners” set to zero while those of the “real partners” remain at the original values. Moreover, a latent group indicator variable that shows whether the subject is virtual partner or not is created ($M = 1$ for virtual and $M = 0$ for real partner). While Huberman and Langholz developed this for two independent case-control sets, we adapt it here for a scenario where the case group is common to both control groups (as in our data example). In our adaptation, separate estimates for each comparison can be obtained by introducing an interaction term of X with W where W is an indicator of whether the control group arises from a matched or unmatched study ($W=1$ if matched, $W=0$ if unmatched). The corresponding test of homogeneity is simply a Wald test based on the naive assumption of independence between the two estimates.

The BLGIM likelihood in this context where the case group contributes twice to the likelihood is given by,

$$L(\beta, \gamma) = \prod_{k=1}^K \frac{\exp(\mathbf{X}'_{1k}\beta + M_{1k}\gamma)}{\exp(\mathbf{X}'_{1k}\beta + M_{1k}\gamma) + \exp(\mathbf{X}'_{0k}\beta + M_{0k}\gamma)}, \quad (2)$$

where $\mathbf{X}_{jk}$, $j = 0, 1$; $k = 1, \dots, K$, is the vector of covariates including the interactions with W ($j = 0$ for control and $j = 1$ for case and K is the total number of strata), M_{jk} is latent group indicator, β is the vector of parameters of interest and γ is a parameter associated with M .

3.4. Polytomous Latent Group Indicator Method

We propose the Polytomous Latent Group Indicator Method (PLGIM) an extension of the BLGIM to polytomous logistic regression model. The parameters of this model will be estimated using polytomous conditional likelihood [12,19,20,21]. Suppose there are $k = 1, \dots, K$ strata (or matched sets). With a slight abuse of notation, for the k th matched set and for $J + 1$ groups, let $D_k = (D_{k0}, \dots, D_{kJ})$, where $D_{kj}=1$ denotes whether a subject in stratum k belongs to group $j \in \{0, \dots, J\}$. Let $\mathbf{X}_{ki} = (X_{ki1}, \dots, X_{kip})$ for $i = 1, \dots, J + 1$ denote a covariate vector of dimension p . Let the matrix $\beta = (\beta_0, \dots, \beta_J)$. Assume that $\pi_j(\mathbf{X}_{ki}) = \Pr(D_{kj}=1)$ follows a linear logistic model with log odds ratio parameter β_j and an unspecified nuisance parameter α_{kj} given by,

$$\pi_j(\mathbf{X}_{ki}) = \frac{\exp(\alpha_{kj} + \mathbf{X}'_{ki}\beta_j)}{\sum_{i=1}^{J+1} \exp(\alpha_{ki} + \mathbf{X}'_{ki}\beta_i)}. \quad (3)$$

A different reparametrization of (3) needs to be considered to permit identifiability of the model without affecting its generalizability. In our case we assume $\beta_0 = 0$ and $\alpha_{k0} = 0$.

To make the presentation clearer, we first introduce the PLGIM in the simplest scenario of a single case set and two control sets ($J = 2$) both arising from a 1:1 individually matched case-control design as matched triplets. We assume a univariate X which takes values x_{ki} , ($i = 1, 2, 3$; $k = 1, \dots, K$). For the general case of $J > 2$ and for a scenario of $1:m$ matching we refer the reader to Levin's work [19,21]. Let an indicator variable $D_{kj} = 1$ if subject belongs to group j and 0 otherwise (where $\sum_{j=0}^2 D_{kj} = 1$). Given the covariate values for the matched triplet from the k th strata (X_{k1}, X_{k2}, X_{k3}), let $(Z_{k0}(t), Z_{k1}(t), Z_{k2}(t))$ equal one of the $t = 1, \dots, 3/(1!1!1!)$ different permutations of the triplet (X_{k1}, X_{k2}, X_{k3}) . Then, the polytomous conditional likelihood is defined as the product of the probability that $Z_{k0}(1)$ comes from the case and $Z_{k1}(1)$ comes from the first control for $k = 1, \dots, K$ and is given by,

$$\begin{aligned} L(\beta_1, \beta_2) &= \prod_{k=1}^K \frac{\pi_0(Z_{k0}(1))\pi_1(Z_{k1}(1))\pi_2(Z_{k2}(1))}{\sum_{t=1}^6 \pi_0(Z_{k0}(t))\pi_1(Z_{k1}(t))\pi_2(Z_{k2}(t))} \\ &= \prod_{k=1}^K \frac{e^{X'_{k2}\beta_1 + X'_{k3}\beta_2}}{e^{X'_{k1}\beta_1} \left[e^{X'_{k2}\beta_2 + e^{X'_{k3}\beta_2}} + e^{X'_{k2}\beta_1} \left[e^{X'_{k3}\beta_2 + e^{X'_{k1}\beta_2}} + e^{X'_{k3}\beta_1} \left[e^{X'_{k1}\beta_2 + e^{X'_{k2}\beta_2}} \right] \right] \right]}, \end{aligned} \quad (4)$$

where $X'_{ki} = -X_{ki}$. Replacing each component of Equation 4 by the formula in Equation 3, leads to the needed likelihood. It has been shown that the polytomous conditional logistic likelihood gives consistent and asymptotically efficient estimates [19,20,21].

Following the development of polytomous conditional likelihood for $J = 2$, we propose the PLGIM to handle combining matched and unmatched case-control data for the more general model,

$$\log \frac{\pi_{kj}(\mathbf{X}_{ki})}{\pi_{k0}(\mathbf{X}_{k1})} = \alpha_{kj} + \mathbf{X}_{ki}'\beta_j, \quad (5)$$

where $j = 0, \dots, J$. Without loss of generality, assume that there is a common case group and the J control groups come from an individually matched (with controls related to the case) or from an unmatched design or from a case-combined-control design [3]. To do a combined analysis of matched and unmatched case-control data in a polytomous logistic framework, we create “virtual partners” for each group member in the unmatched data with their covariate values assigned the baseline category of X (usually zero). Then, a latent group indicator variable M_{ki} that takes a value of 1 when the i th subject is virtual and 0 when it is real is computed. It should be noted that use of the virtual subjects does not inflate the sample size in any informative way nor result in decreased standard errors such as is found when missing covariate data are singly imputed in a missing data problem which entails some effort to obtain valid standard errors. Rather, virtual subjects merely allow for a convenient computational trick to construct a valid likelihood from the matched and unmatched data sets that are being combined.

Let $D_{k0}, \dots, D_{kJ}$ denote group membership indicators, where D_{kj} takes a value of one if a subject in the k th stratum is from group j and 0 otherwise. Let W be an indicator variable of which study a subject came from. Given the data from the k th stratum, let $(Z_{k0}, \dots, Z_{kJ})$ be one of the $t = 1, \dots, C_k = n_k!/(n_{k0}! \dots n_{kJ}!)$ different permutations of the X 's. The polytomous conditional logistic likelihood is defined as the product of the probability that $Z_{k0}(1)$ comes from the case and $Z_{k1}(1)$ comes from the first control group, ..., and Z_{kJ-1} comes from the $(J - 1)$ th control group for $k = 1, \dots, K$ and is given by,

$$L(\beta_1, \dots, \beta_J, \gamma_1, \dots, \gamma_J) = \prod_{k=1}^K \frac{\exp(S_{kj}(1) + M_j \gamma_j)}{\sum_{t=1}^{C_k} \exp(S_{kj}(t) + M_j \gamma_j)}, \quad (6)$$

where, $S_{kj}(t) = \sum_{i=1}^{J+1} \beta_j \mathbf{X}_{ki}' \mathbf{I}\{Z_{kj}(t) = \mathbf{X}_{ki}\}$, γ_j is a parameter corresponding to M_j for $j = 0, \dots, J$, $\beta_0 = 0$, $\gamma_0 = 0$ and $n_k = n_{k0} + \dots + n_{kJ}$ (where n_{kj} is the number of subjects of type j) is the total number of subjects in stratum k . A special case of this is when each group has only one subject in each stratum such that $C_k = (J + 1)!$. Whether pooling is possible can be checked by testing if the interaction between X and W is significant or not via Wald or Likelihood ratio tests. These tests account for correlation in the estimated parameter.

4. DATA EXAMPLE RESULTS

We demonstrate the implementation of PLGIM using our data example. The important problem is estimating the association between a genetic exposure X and multiple myeloma in a case-control study with a single case-set and two sets of controls (individually matched/related and unmatched/unrelated controls). For this data set, the polytomous logistic regression will involve the assignment of the triplets denoted by: cases (D_{k0}), related controls (D_{k1}) and unrelated controls (D_{k2}) into matched triplets with a “real” or “virtual” partners. This assignment results in three varieties of matched triplets: (*case, related control, virtual unrelated control*), (*virtual case, virtual related control, unrelated control*) and (*case, virtual related control, virtual unrelated control*). And, these matched triplets have six

possible permutations of their covariate values: (X_{k1}, X_{k2}, X_{k3}) , (X_{k2}, X_{k3}, X_{k1}) , (X_{k3}, X_{k1}, X_{k2}) , (X_{k1}, X_{k3}, X_{k2}) , (X_{k2}, X_{k1}, X_{k3}) and (X_{k3}, X_{k2}, X_{k1}) .

After creating “virtual” indicator variables and re-arranging the data in the form described above, analysis was made using polytomous conditional logistic regression. For comparison purposes, analysis was also made using the naive method, BMM and BLGIM.

The results from these analyses are displayed in Table I. Tests of homogeneity for all the methods (BMM, BLGIM and PLGIM) indicated that the odds ratios from the matched and unmatched analysis were homogenous and thus we estimated the pooled odds ratio, except for the naive scenario. As we expected, the smaller standard error and consequently the tighter confidence interval for the PLGIM suggests a pooled common odds ratio that is estimated with more precision and efficiency. Using this approach, the variant allele of the IL-6 promoter SNP-572 was associated with a two fold increased risk of plasma cell neoplasm when cases were compared with controls (family and population controls).

5. SIMULATION STUDY

5.1. Data Generation and Simulation Parameters

We performed several simulations with the objective of evaluating the relative performance of the different estimators as well as the power and type-I error rate of the homogeneity tests.

Our data was generated using a conceptual framework that relies on failure time considerations. The data generation process is as follows. At a first stage we generate data from a cohort of size 50,000 which represents a large population of individuals such that the disease incidence rate for the i th individual in the k th stratum is given by $\lambda_{ik} = \exp(\alpha_k z_k + \beta x_{ik})$. The variable Z is the matching variable and is generated from a discrete uniform distribution taking integer values between 1 and 100. The parameter α_k was defined to take values $k/100$ so that the risk of becoming a case varies by strata. We define X as a continuous exposure variable that is normally distributed with mean and variance selected to ensure that the linear correlation between X and Z equals pre-specified values (0, 0.2, 0.4, 0.8). A binary x with $\Pr(x = 1) = 0.4$ is also considered. We set the parameter β to different values corresponding to odds ratios of 1.0, 1.2, 1.5 and 2.0. After evaluating λ_{ik} we sequentially sample a fixed number of cases, where the probability of being a case is taken to be proportional to the individual disease incidence rate, for those subjects that are still at risk. We first randomly sampled 100 and 200 cases-control pairs from each stratum defined by Z . Then, we considered a randomly selected set of unmatched controls from the remainder of the population who are neither a case nor matched control. A second control group randomly selected from a population with a different β was also considered. For all the simulations we set the number of unmatched controls to be twice that of the number of cases. In each simulation, 2000 data sets were generated.

5.2. Simulation Results

In Tables II, III and IV we show the summary results of the simulations. We considered a total of 64 different scenarios (2 sample size scenarios, 4 (X, Z) correlation, 4 different values for β s, X normal and binary). We report the scenario for normal X with a sample size of 100 cases, 100 matched and 200 unmatched controls to provide overall patterns of results. The additional scenarios with larger sample size and binary X have similar patterns while demonstrating an increased ability to detect heterogeneity and an increase in statistical efficiency. The column labeled “method” depicts the four different methods: (1) Naive for separate fitting of the matched and unmatched data and a Wald homogeneity test that assumes independence in the estimates; (2) BMM for the break the match method and a

Wald homogeneity test; (3) BLGIM for the binary latent group method and a Wald homogeneity test that assumes independence in the estimates; and (4) PLGIM for the polytomous latent group method and a likelihood ratio homogeneity test that accounts for covariation in the estimates. Summary statistics are based on 2000 replications.

Table II shows the performance of the different methods in terms of bias and efficiency under the assumption of true homogeneity. With the exception of BMM all three methods seem to provide unbiased estimates. More specifically, in Table II, as the correlation increases from 0 to 0.8, the bias in BMM is quite small for the pooled estimate and $\beta = 0$, with only a slight increase from 0.003 to 0.006. However, for $\beta = 0.69$, the bias becomes substantial, with the mean dropping from 0.703 to 0.552 as the correlation between X and Z increases from 0 to 0.8, respectively. More interesting is the pattern observed in the standard errors. Here, we are able to see a substantial gain of efficiency when using the PLGIM method compared to the other methods. The gain is more pronounced when estimating the common pooled β and for high levels of correlation between X and Z . This efficiency gain is also seen for the estimation of β_1 , the estimate obtained from the matched comparison. For example, as shown in Table II, when no correlation exists between X and Z and $\beta = 0$ the standard error for β_1 is 0.101 for the PLGIM method as compared to 0.124 for the naive, BMM and BLGIM approaches. The gain in efficiency is higher for the pooled estimator. While the standard error for PLGIM is 0.092, it is 0.116 and 0.094 for the BMM and BLGIM approaches, respectively. Of particular interest is the observation that the PLGIM approach gives consistently better performance than the other approaches as the correlation between X and Z increases (where the correlation is measuring the level of confounding by Z in the association between X and D).

Table III shows the mean and standard error of the parameter estimates under simulation scenarios in which $\beta_1 \neq \beta_2$. Here, the pattern in the performance of the methods is consistent with the previous results. While BMM resulted in biased estimates with high correlation, all other methods appear unbiased with PLGIM demonstrating substantial gains in terms of efficiency.

Table IV reports the power and test size performance across various scenarios. We note that since the Naive test (based on a simple Wald statistic) and the BLGIM approach (based on a test of interaction between X and the study indicator, W) both assume independence between the estimates, they have equivalent Type I error and are reported jointly in a single line. Whenever the type-I error rates substantially deviate from the nominal 5%, we do not report the power estimates. Clearly, the results show the superior performance of the tests based on the PLGIM approach. The PLGIM is the only test with correct Type I error for all scenarios considered in the simulation. As expected, power increases with the difference between the true values of the coefficients but decreases when the confounding effect of Z is higher. In additional investigations (results not tabulated), we have confirmed that all the tests considered have the expected Chi-square distribution at the null and a non-central chi-square distribution away from the null, and hence power calculations are apparently made under the correct assumption.

6. DISCUSSION

In this paper we have proposed a novel approach to combine matched and unmatched case-control data using a polytomous/multinomial conditional logistic likelihood with “virtual partners”. In a real data example, we demonstrated the practical implications of this novel approach. Here, although the two estimates from the original unmatched and matched case-control comparisons provided similar values, the individual statistical tests were inconclusive with only one providing statistical significance. In contrast, the PLGIM test of

homogeneity indicated that a combined estimate was appropriate and a single statistically significant estimate was obtained. Via simulations we have demonstrated that in comparison to the more commonly used naive approach of keeping the two estimates independent, PLGIM provides an unbiased and more statistically efficient pooled estimate and has a more powerful test of homogeneity while still maintaining the correct Type I error rate.

Methods for combining different case-control comparisons are increasing in their importance as genetic association studies rely more heavily on large sample sizes to identify and replicate genetic effects. To achieve these sample sizes, many research initiatives are encouraging the collaboration and combination of several existing matched and unmatched case-control studies. Thus, it is becoming more common to compare multiple sets of controls with the same case group or multiple case groups to identify, validate or confirm a positive or negative finding. For example the US National Cancer Institute (NCI) has initiated the Cohort Consortium to bring together nine existing cohorts with nested case-control sampling to assess the effects of common genetic variants in cancer risk [22]. Similarly, responding to the interest in studying the gene-environment ($G \times E$) interactions, Andrieu and Goldstien (2004) have proposed a case-combined-control design to increase the power to detect the $G \times E$. PLGIM can be applied in these settings. In addition to a straightforward comparison of study specific estimates and a common pooled estimate, the method can be used to compare relative risk estimates of the same risk factors studied in different phases of a disease in order to explore factors that may be more important in one phase as compared to another [9]. By examining the comparison-specific estimates and a combined estimate, the approach can also be used to validate or confirm results of a case-control study using another set of controls or to assess selection bias in the case or control population [5]. The latent group approach is also related to a missing data indicator method for missing covariate data presented in [25]. Moreover, this method applies to the analysis of missing family relative data especially in situations where incomplete matched triplet data could lead to missing data bias, for example, if refusal to consent or failure to collect DNA were associated with the disease and risk factors.

In summary, the PLGIM is a unified approach that can be used to do a combined analysis of matched and unmatched binary as well as multinomial response case-control data accounting for any interdependence in the parameter estimates from the different studies. For instance, when studies share the same case group, the assumption of independence usually made when testing the equality of parameters leads to an inflated type-I or type-II error rates depending on the magnitude and direction of ρ .

This work can be extended to a case of ordered categories which requires adaptation of the recently suggested conditional logistic methods for ordinal response data [11,26,27]. Similarly, the PLGIM could be extended to a $1:m$ matched design. Finally, we should note that the PLGIM method can be implemented using standard statistical software after only minor and simple reorganization of the data. Program codes for implementing this method in SAS, R and Stata softwares are available at the first author's web site <http://people.musc.edu/gebregz/StatisticalPrograms>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is partially supported by NIH-NIA grants P30 AG021677-05, NSF (SC EPSCoR/IDeA EPS-0447660) and MUSC office of the Provost. The use of the cancer registry to identify multiple myeloma patients was supported in part by the California Department of Health Services as part of the statewide cancer reporting program

mandated by the California Health and Safety Code Section 103885. The multiple myeloma case-control study has been funded in part with Federal funds from the National Cancer Institute, NIH, Department of Health and Human Services under contract no. N01-PC-35139, 5P30-CA-14089-30, and IP50-CA-10070. Dr. Conti is supported by grants P50 CA084735, U01 DA020830 P30, ES 007048, U01CA122839, U01 ES015090. We also extend our thanks to the reviewers.

REFERENCES

1. Breslow NE. Statistics in Epidemiology: the case-control study. *Journal of the American Statistical Association*. 1996; 91:14–28. [PubMed: 12155399]
2. Schlesselman, JJ. *Case-Control Studies: Design, Conduct and Analysis*. Oxford University Press; New York: 1982.
3. Andrieu N, Goldstein AM. The case-combined-control design was efficient in detecting gene-environment interactions. *J Clin Epidemiol*. 2004; 57:662–671. [PubMed: 15358394]
4. Goldstein AM, Dondon MG, Andrieu N. Unconditional analyses can increase efficiency in assessing gene-environment interaction of the case-combined-control design. *Int J Epidemiol*. 2006; 35:1067–1073. [PubMed: 16556643]
5. Cozen W, Gebregziabher M, Conti D, et al. Interleukin-6 Related Genotypes, Bodymass Index And Risk Of Multiple Myeloma And Plasmacytoma. *Cancer Epidemiology, Biomarkers and cancer Prevention*. 2006; 15:1–7.
6. Duffy SW, Rohan TE, Altman DG. A Method For Combining Matched And Unmatched Binary Data: Application To Randomized, Controlled Trials Of Photocoagulation In The Treatment Of Diabetic Retinopathy. *American Journal of Epidemiology*. 1989; 130:371–378. [PubMed: 2750732]
7. Greenland S. RE: A Method For Combining Matched And Unmatched Binary Data: Application To Randomized, Controlled Trials Of Photocoagulation In The Treatment Of Diabetic Retinopathy. *American Journal of Epidemiology*. 1990; 132:197–198. [PubMed: 2192550]
8. Le Cessie S, Nagelkerke N, Rosendaal FR, et al. Combining matched and unmatched control groups in case-control studies. *Am J Epidemiol*. 2008; 168:1204–10. [PubMed: 18836151]
9. Moreno V, Martin ML, Bosch FX, et al. Combined Analysis Of Matched And Unmatched Case-Control Studies: Comparison Of Risk Estimates From Different Studies. *American Journal of Epidemiology*. 1996; 143(3):293–300. [PubMed: 8561164]
10. Huberman M, Langholz B. RE: Combined Analysis of Matched and Unmatched Case-Control Studies: Comparison of Risk Estimates from Different Studies. *American Journal of Epidemiology*. 1999; 150:219–220. [PubMed: 10412970]
11. Agresti, A. *Categorical Data Analysis*. 2nd ed. Wiley; Hoboken: 2002.
12. Liang KY, Stewart WF. Polychotomous Logistic Regression Methods For Matched Case-Control Studies With Multiple Case Or Control Groups. *American Journal of Epidemiology*. 1987; 125:720–730. [PubMed: 3826049]
13. Hosmer, DW.; Lemeshow, S. *Applied Logistic Regression*. Wiley; New York: 2000.
14. Mantel N, Haenszel W. Statistical Aspects Of The Analysis Of Data From Retrospective Studies Of Disease. *J Natl Cancer Inst*. 1959; 22:719–748. [PubMed: 13655060]
15. Breslow, NE.; Day, NE. *Statistical Methods in Cancer Research II: The Analysis of Case-Control Studies*. IARC Scientific Publications; Lyon, France: 1980.
16. Breslow, NE.; Day, NE. *Statistical Methods in Cancer Research I: The Analysis of Case-Control Studies*. IARC Scientific Publications; Lyon, France: 1980.
17. Levin B, Paik MC. The Unreasonable Effectiveness of a Biased Logistic Regression Procedure in the Analysis of Pair-Matched Case-Control Studies. *Journal of Statistical Planning and Inference*. 2001; 96:371–385.
18. Langholz B, Goldstein L. Conditional Logistic Analysis Of Case-Control Studies With Complex Sampling. *Biostatistics*. 2001; 0:1–22.
19. Levin B. The Saddlepoint Correction In Conditional Logistic Likelihood Analysis. *Biometrika*. 1990; 77:275–285.
20. Levin B. Conditional Likelihood Analysis in Stratum-Matched Retrospective Studies with Polytomous Disease States. *Communications in Statistics B*. 1988; 16:699–718.

21. Levin B. Polychotomous Logistic Regression Methods For Matched Case-Control Studies With Multiple Case Or Control Groups. *American Journal of Epidemiology*. 1988; 128:445–446.
22. Hunter DJ, Riboli E, Haiman CA, et al. A Candidate Gene Approach To Searching For Low-Penetrance Breast And Prostate Cancer Genes. *Nature Reviews Cancer*. 2005; 5:977–985.
23. Rothman, KJ.; Greenland, S. *Modern Epidemiology*. Lippincott Williams and Wilkins; Philadelphia: 1998.
24. Huberman M, Langholz B. Application of the Missing-Indicator Method in Matched Case-Control Studies with Incomplete Data. *American Journal of Epidemiology*. 1999; 150:1340–1345. [PubMed: 10604777]
25. Gebregziabher M, Langholz B. A Semiparametric Missing-Data-Induced Intensity Method for Missing Covariate Data in Individually Matched Case-Control Studies. *Biometrics*. 2009:0000–00.
26. Mukherjee B, Liu I, Sinha S. Analysis Of Matched Case-Control Data With Ordinal Disease States: Possible Choices And Comparisons. *Statistics in Medicine*. 2007; 26:3240–3257. [PubMed: 17206600]
27. Liu I, Agresti A. The Analysis Of Ordered Categorical Data: An Overview And Survey Of Recent Developments. *Test*. 2005; 14:1–73.

Table I

Odds ratio[†] and standard error (SE) estimates for the association of IL-6-572 and the risk of multiple myeloma, Los-Angeles County, 1999–2002.

Method	$\hat{\eta}_1$ (SE)	$\hat{\eta}_2$ (SE)	$\hat{\eta}$ (SE)	95% CI of $\hat{\eta}$
Naive	1.8 (0.89)	2.4 (0.80)	NA	NA
BMM	2.1 (0.57)	2.0 (0.67)	2.1 (0.57)	1.19–3.52
BLGIM	1.6 (0.76)	2.3 (0.78)	2.0 (0.57)	1.19–3.52
PLGIM	1.7 (0.56)	2.3 (0.79)	2.0 (0.54)	1.16–3.40

$\hat{\eta}_1$ for case-relative, $\hat{\eta}_2$ for case-population control, $\hat{\beta}$ pooled estimator

95% CI= 95% confidence interval, SE=asymptotic standard error

Naive=separate pairwise naive analysis; same as those reported in Cozen et al (2006)

BMM=break the match method

BLGIM=binary latent group indicator method

PLGIM=polytomous latent group indicator method

NA= not applicable

[†] adjusted for age, gender, education, BMI and race as in Cozen et al (2006)

Table II

Mean and standard error estimates of the log odds ratio when X is Normally Distributed, $n_1 = 100$, $n_2 = 200$, cohort size=50,000 and 2000 replications.

$\beta_1 = \beta_2 = 0$							$\beta_1 = \beta_2 = 0.69$							
Corr(X,Z)	Parameter	Method	Mean	ESE	Mean	ESE	Corr(X,Z)	Parameter	Method	Mean	ESE	Mean	ESE	
0	β	BMM	.003	.116	.703	.128								
		BLGIM	.003	.094	.706	.108								
		PLGIM	.003	.092	.704	.104								
	β_1	Naive	.002	.124	.707	.138								
		BMM	.002	.124	.704	.134								
		BLGIM	.002	.124	.705	.137								
β_2	PLGIM	.002	.101	.701	.112									
	Naive	.003	.145	.717	.179									
	BMM	.003	.143	.704	.152									
0.4	β	BMM	.003	.161	.669	.168								
		BLGIM	.005	.142	.708	.152								
		PLGIM	.005	.139	.709	.148								
	β_1	Naive	.009	.189	.712	.241								
		BMM	.009	.171	.711	-								
		BLGIM	.004	.171	.710	.241								
β_2	PLGIM	.008	.153	.709	.193									
	Naive	.000	.219	.715	.197									
	BMM	.001	.197	.592	-									
0.8	β	BLGIM	.000	.219	.715	.181								
		PLGIM	.001	.187	.714	.161								
	β_1	BMM	.006	.322	.552	-								
		BLGIM	.005	.433	.710	.437								
		PLGIM	.006	.427	.715	.430								
	Naive	.011	.574	.716	.579									

		$\beta_1 = \beta_2 = 0$		$\beta_1 = \beta_2 = 0.69$		
Corr(X/Z)	Parameter	Method	Mean	ESE	Mean	ESE
β_2		BMIM	.000	.342	.704	.347
		BLGIM	.010	.574	.703	.579
		PLGIM	.011	.468	.715	.473
	Naïve	Naïve	.003	.668	.718	.675
		BMIM	.010	-	.257	-
		BLGIM	.003	.668	.718	.675
	PLGIM	.001	.573	.720	.575	

ESE=empirical standard error

naïve=separate pairwise naïve analysis

BMM=Break the matching method

BLGIM=binary latent group indicator method

PLGIM=polytomous latent group indicator method

Corr(X,Z)=correlation or association between X and Z

- = value deleted if estimate is biased

Table III

Mean and standard error estimates of the log odds ratio when X is Normal, for different (β_1, β_2) combinations and $n_1 = 100$, $n_2 = 200$, cohort size=50,000 and 2000 replications.

$(\beta_1 = 0, \beta_2 = 0.18)$														$(\beta_1 = 0, \beta_2 = 0.41)$				$(\beta_1 = 0, \beta_2 = 0.69)$			
Corr(X,Z)	Parameter	Method	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE	Mean	ESE							
0	β	BMM	.120	.116	.271	.119	.437	.122													
		BLGIM	.104	.094	.240	.097	.401	.102													
		PLGIM	.125	.092	.284	.096	.470	.100													
	β_1	Naïve	.001	.125	.007	.131	.003	.143													
		BMM	.002	.125	.007	.131	.007	.139													
		BLGIM	.001	.125	.007	.131	.003	.142													
	β_2	PLGIM	.008	.102	.003	.108	.005	.118													
		Naïve	.184	.145	.417	.148	.707	.152													
		BMM	.183	.143	.417	.145	.704	.150													
0.4	β	BLGIM	.101	.142	.235	.144	.401	.148													
		PLGIM	.121	.140	.280	.141	.475	.145													
		Naïve	.001	.189	.002	.176	.001	.185													
	β_1	BMM	.001	.172	.002	.175	.001	.183													
		BLGIM	.001	.172	.002	.176	.001	.185													
		PLGIM	.008	.154	.002	.158	.004	.166													
	β_2	Naïve	.177	.219	.413	.221	.706	.224													
		BMM	.177	.198	.414	.202	.704	.203													
		BLGIM	.177	.219	.414	.221	.704	.224													
0.8	β	PLGIM	.178	.188	.416	.189	.719	.193													
		BMM	.122	.322	.277	.322	.457	.324													
		BLGIM	.110	.433	.240	.433	.396	.436													
		PLGIM	.130	.427	.280	.427	.473	.429													

$(\beta_1 = 0, \beta_2 = 0.18)$ $(\beta_1 = 0, \beta_2 = 0.41)$ $(\beta_1 = 0, \beta_2 = 0.69)$						
Corr(X,Z)	Parameter	Method	Mean	ESE	Mean	ESE
β_1		Naïve	.009	.343	.002	.344
		BMM	.004	.343	.000	.343
		BLGIM	.004	.343	.000	.344
		PLGIM	.009	.469	.002	.395
β_2		Naïve	.186	.669	.420	.668
		BMM	-	-	-	-
		BLGIM	.182	.669	.419	.668
		PLGIM	.185	.574	.421	.469

ESE=empirical standard error

naïve=separate pairwise naïve analysis

BMM=Break the matching method

BLGIM=binary latent group indicator method

PLGIM=polytomous latent group indicator method

Corr(X,Z)=correlation or association between X and Z

- = if pooled estimate not possible

Table IV

Type-I error rate and Power of Tests of $H_0 : \beta_1 = \beta_2$ when X is Normally Distributed and $n_1 = 100, n_2 = 200, cohortsize = 50, 000, 2000$ replications.

		Power			
		X is Normally distributed			
$Corr(X,Z)$	Method	Type-I Error	$(\beta_1 = 0, \beta_2 = 0.18)$	$(\beta_1 = 0, \beta_2 = 0.41)$	$(\beta_1 = 0, \beta_2 = 0.69)$
0	BMM	.043	.310	.891	1.00
	Naive/BLGIM	.004	-	-	-
	PLGIM	.040	.277	.862	1.00
0.4	BMM	.107	-	-	-
	Naive/BLGIM	.005	-	-	-
	PLGIM	.040	.132	.543	.935
0.8	BMM	.259	-	-	-
	Naive/BLGIM	.023	-	-	-
	PLGIM	.047	.046	.205	.991

naive=separate pairwise naive analysis

BMM=Break the match

BLGIM=binary latent group with interaction term

PLGIM=polytomous latent group with two latent group indicators

$Corr(X,Z)$ =correlation or association between X and Z

- = if not applicable or estimate is biased