Article

Modeling Secondary Phenotypes Conditional on Genotypes in Case–Control Studies

Naomi C. Brownstein [1,*], Jianwen Cai [2], Shad Smith [3], Luda Diatchenko [4,5], Gary D. Slade [6] and Eric Bair [2,6]

[1] Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL 33612, USA
[2] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420, USA; cai@bios.unc.edu (J.C.); ericdavidbair@aim.com (E.B.)
[3] Center for Translational Pain Medicine, Department of Anesthesiology, Duke University Medical Center, Durham, NC 27710, USA; shad.smith@duke.edu
[4] Faculty of Dental Medicine and Oral Health Sciences, McGill University, Montréal, QC H3A 0G1, Canada; luda.diatchenko@mcgill.ca
[5] Department of Anesthesia, Faculty of Medicine, McGill University, Montréal, QC H3A 0G1, Canada
[6] School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7450, USA; gary_slade@unc.edu
* Correspondence: naomi.brownstein@moffitt.org

Abstract: Traditional case–control genetic association studies examine relationships between case–control status and one or more covariates. It is becoming increasingly common to study secondary phenotypes and their association with the original covariates. The Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) project, a study of temporomandibular disorders (TMD), motivates this work. Numerous measures of interest are collected at enrollment, such as the number of comorbid pain conditions from which a participant suffers. Examining the potential genetic basis of these measures is of secondary interest. Assessing these associations is statistically challenging, as participants do not form a random sample from the population of interest. Standard methods may be biased and lack coverage and power. We propose a general method for the analysis of arbitrary phenotypes utilizing inverse probability weighting and bootstrapping for standard error estimation. The method may be applied to the complicated association tests used in next-generation sequencing studies, such as analyses of haplotypes with ambiguous phase. Simulation studies show that our method performs as well as competing methods when they are applicable and yield promising results for outcome types, such as time-to-event, to which other methods may not apply. The method is applied to the OPPERA baseline case–control genetic study.

Keywords: bootstrap; case–control studies; inverse-probability weighting; secondary analysis

check for updates

Citation: Brownstein, N.C.; Cai, J.; Smith, S.; Diatchenko, L.; Slade, G.D.; Bair, E. Modeling Secondary Phenotypes Conditional on Genotypes in Case–Control Studies. *Stats* **2022**, *5*, 203–214. https://doi.org/10.3390/stats5010014

Academic Editor: Eddy Kwessi

Received: 31 December 2021
Accepted: 3 February 2022
Published: 22 February 2022

1. Introduction

Prospective studies are more straightforward and less prone to confounding than other study designs. However, they may require either extremely long follow-up periods or large sample sizes, and lack power. For rare diseases in particular, the sample sizes required in a prospective cohort study to have adequate statistical power to test hypotheses of interest may be prohibitively large. This can be especially problematic in genetic association studies, which may cost thousands of dollars per participant just to extract their genetic profiles. Retrospective case–control studies are more cost effective. The number of case–control studies focusing on the relationship between genetics and disease outcomes has grown astronomically in recent years.

It is well-known that when modeling the probability of case status in a case–control design, logistic regression may be used to model the primary outcome as if the study were prospective [1]. However, researchers may design studies based on one outcome and study secondary outcomes simultaneously or subsequently. Without proper care, the analysis of secondary phenotypes in case–control studies may be problematic. Standard unadjusted methods, such as linear and logistic regression, may be biased, inefficient, or lead to misleading inference. The standard method of unweighted regression on the full case–control sample and the method of adjusting for case status with an indicator variable have inflated type I error when the disease is not rare and when the disease is related to the secondary phenotype [2]. The popular practices of restricting to either cases only or controls only reduce efficiency and may be subject to bias.

This work arose in consideration with data from the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study [3,4]. The OPPERA study was primarily designed to identify risk factors for temporomandibular disorders (TMD). In addition to the cohort of initially TMD-free adults enrolled in the prospective cohort study, people with examiner-verified chronic TMD were enrolled to create an unmatched case–control study. A large number of putative risk factors were collected at enrollment [4]. Investigators sought to explain relationships between TMD and other chronic pain conditions. One putative risk factor of interest in its own right is the (ordinal) number of comorbid pain conditions a subject experiences. The genetic information collected may be predictive of comorbid conditions as well as of TMD.

Some methods have been proposed for analyzing secondary phenotypes in case–control studies. Subjects from a nested prospective case–control study may be weighted by the reciprocal of their probability of selection [5]. This stratum-weighted logistic regression method, also called inverse probability weighting (IPW), achieves the nominal type I error rate, but it can be less efficient than the standard unadjusted method or the method of adjusting for case status [2]. (Yet, in light of the fact that the method of adjusting for case status may have inflated type I error, the lower power of IPW is less alarming.) The IPW estimator [5] may merit a correction factor for the standard error. Another IPW-based method was developed for continuous outcomes [6,7] by fitting estimating equations separately to cases and controls and then combining the estimates to have minimal variance.

Likelihood-based methods have been developed [8,9] that are more powerful than IPW and can be used for both continuous and binary outcomes. However, the results may be biased and have inflated size when there is significant interaction between the genotypes and the original outcome [10,11]. Robust likelihood estimation approaches are provided for continuous outcomes using bootstrapping under the assumption that the secondary phenotype can be modeled with homoscedastic regression or the disease is rare [12], as well as under heteroscedasticity or higher prevalence [13]. Robust sandwich estimators have been proposed for the variance based on generalized estimating equations (GEE), which applies to both continuous [2,14] and binary outcomes [2]. A general conditional likelihood-based method [15] exists for outcomes that utilize standard link functions (identity, logit, and log), i.e., binary, continuous, and count outcomes.

Additional bootstrap estimates have been proposed for binary outcomes [16]. They derive non-linear equations relating the logistic regression coefficients to known sample sizes and prevalence estimates for the primary disease and secondary phenotype, calculate the unadjusted parameter estimate and standard error, resample parameter estimates from a normal distribution with that mean and standard error, refine the estimate for each replication by solving the aforementioned non-linear equations, and use bootstrapping for confidence intervals. They also adapt their bias correction method to the frequency-matched case–control study design and extend the IPW method to retrospective case–control studies using an estimate of primary disease prevalence to calculate weights. These methods have greater efficiency than the IPW method, but they assume that the secondary phenotype and covariates are binary and there is no gene–environment interaction [17].

When there is gene–environment interaction, estimation with adaptive weighting motivated by a Bayesian shrinkage estimator is recommended for when the disease is rare [10] or common [18]. Joint modeling based on Gaussian copulas for the analysis of multiple secondary phenotypes in the exponential family has a controlled type I error and is more powerful than the IPW method [11].

Existing methods may not be appropriate for all applications. First, the IPW method [5] may require an adjustment to the standard error and was originally only designed to apply to binary outcomes. Additional variable types are commonly analyzed in practice. More recent methods have only been explicitly derived and tested for continuous [2,6–9,12,13,15,19], binary [2,7–9,15,16,19], or count outcomes [7,15]. As far as we know, there are no methods that have been developed for and applied to secondary phenotypes outside the exponential family. Time-to-event outcomes, for example, may be of clinical interest, and may be present in large studies, such as OPPERA. Sequencing studies also utilize more complicated test statistics outside of the exponential family. Additionally, there may be a lack of user-friendly software for implementation.

In this paper, we propose a method for analyzing secondary phenotypes of general form in case–control genetic association studies. We advocate using the IPW method [5] for parameter estimation, but estimating the standard error via bootstrapping. This maintains the simplicity and intuitiveness of IPW and generalizes it to a wider variety of situations than previously applied, while providing a valid estimate of the standard error. Our method can handle arbitrary types of analyses, including time-to-event and non-parametric methods, as well as logistic regression and linear models, as described in the literature. Moreover, our method can be easily generalized to outcomes for which no existing method applies. We describe our methodology in detail in Section 2. Simulations are presented in Section 3. The method is applied to the OPPERA study in Section 4. We conclude with a discussion.

2. Proposed Method

Consider a case–control study consisting of n cases and m controls. Let Z_i, a $p \times 1$ vector, denote covariate information, D_i denote the case–control status (1 = case, 0 = control), and Y_i denote the secondary phenotype for $i = 1, \ldots, n+m$. For example, in the OPPERA study, Z_i denotes the number of copies of the minor allele, D_i is an indicator of whether participant i is a chronic case of TMD, and Y_i is the ordinal number (0, 1, 2+) of comorbid pain conditions for participant i. (In general, Y_i can take other forms, as described below.) If one were to ignore the case–control study design and consider the data as a random sample from the population, then one would use standard methodology to study the relationship between $Y = (Y_1, \ldots, Y_{n+m})'$ and $Z = (Z_1, \ldots, Z_{n+m})'$. The log-likelihood is denoted under the assumption of random sampling as $l(\theta|Y_i, Z_i)$ where θ is a $q \times 1$ vector of model parameters of interest.

In our proposed method, parameter estimates ($\hat{\theta}$) are generated using the IPW method of [5]. IPW simply weights standard analyses appropriately to account for the oversampling of cases. Specifically, in a prospective (nested) case–control study, if we denote f_{ca} as the sampling fraction for cases and f_{co} as the sampling fraction for controls, we use $w_i = 1$ as the weight for cases and $w_i = \frac{f_{ca}}{f_{co}}$ as the weight for controls. For retrospective case–control studies, the weights may be estimated as in [16] by $w_i = 1$ for cases and $w_i = \frac{n(1-p_e)}{mp_e}$ for controls, where p_e is the estimated prevalence of cases in the population. We may write $w_i(D_i) = D_i + (1 - D_i)w_i$.

The weighted log-likelihood is the weighted sum of the log-likelihood of each observation

$$l_W(\theta|Y, Z, D) = \sum_{i=1}^{m+n} w_i(D_i) l(\theta|Y_i, Z_i), \tag{1}$$

where $l(\theta|Y_i, Z_i) = \log[f(Y_i|Z_i)]$ if Y is continuous with pdf $f(Y_i|Z_i)$,
or $l(\theta|Y_i, Z_i) = \log[P(Y_i|Z_i)]$ if Y is discrete with pmf $P(Y_i|Z_i)$.

As a simple example, suppose Y_i is continuous, and let $\tilde{Z}_i = (1, Z_i')'$ add an column of ones to the covariate information in Z_i. Then, in the setting of random sampling, $Y_i = \alpha'\tilde{Z}_i + \epsilon_i$, $E(Y_i) = \alpha'\tilde{Z}_i$, and $Var(Y_i) = \epsilon_i \sim N(0, \sigma^2)$ would comprise the underlying unweighted linear model with parameter $\theta = (\alpha, \sigma^2)$, where α is a $(p+1) \times 1$ vector and $\sigma^2 \geq 0$ is a non-zero constant scalar.

In the case–control setting discussed in this paper, a weighted linear model could be utilized to estimate $\theta = (\alpha, \sigma^2)$ with log-likelihood according to Equation (2):

$$l_W(\theta|Y, Z, D) = \sum_{i=1}^{m+n} \log[f(Y_i|Z_i, D_i)] = \sum_{i=1}^{m+n} w_i(D_i)\left[-\frac{\log(2\pi\sigma^2)}{2} - \frac{(Y_i - \alpha'\tilde{Z}_i)^2}{2\sigma^2}\right]. \quad (2)$$

If Y_i is binary with covariate $\tilde{Z}_i = (1, Z_i')'$, then under random sampling, one could use a standard logistic regression model, $\log[\frac{P(Y_i=1)}{1-P(Y_i=1)}] = \alpha'\tilde{Z}_i$, where α has length $p+1$. For the case–control setting, one would typically use weighted logistic regression to estimate $\theta = \alpha$, with

$$l_W(\theta|Y, Z, D) = \sum_{i=1}^{m+n} \log[P(Y_i = 1|Z_i, D_i)] = \sum_{i=1}^{m+n} w_i(D_i)\{Y_i\alpha'\tilde{Z}_i - \log[1 + \exp(\alpha'\tilde{Z}_i)]\}. \quad (3)$$

If Y_i is ordinal with K levels, denoted $(1, 2, \ldots, K)$, then under random sampling, one may use a proportional odds model with cumulative logits [20] to estimate $\theta = (\zeta, \beta)$, where $\zeta = (\zeta_1, \ldots, \zeta_{K-1})$ denotes intercepts, and

$$\log\left[\frac{P(Y_i \le k)}{1 - P(Y_i \le k)}\right] = \zeta_k + \beta' Z_i = \beta' Z_i + \sum_{j=1}^{K-1} \zeta_j I(j = k), \quad (4)$$

for $k = 1, \ldots, K-1$, leading to cumulative probabilities,

$$P(Y_i \le k|Z_i) = \frac{\exp(\zeta_k + \beta' Z_i)}{1 + \exp(\zeta_k + \beta' Z_i)} = \frac{\exp[\beta' Z_i + \sum_{j=1}^{K-1} \zeta_j I(j = k)]}{1 + \exp[\beta' Z_i + \sum_{j=1}^{K-1} \zeta_j I(j = k)]}. \quad (5)$$

The cumulative probabilities from (5) may be used to calculate individual probabilities into the likelihood, where

$P(Y_i = 1|Z_i) = P(Y_i \le 1|Z_i)$,
$P(Y_i = k|Z_i) = P(Y_i \le k|Z_i) - P(Y_i \le k-1|Z_i)$ for $k = 2, \ldots, K-1$, and
$P(Y_i = K|Z_i) = 1 - \sum_{k=1}^{K-1} P(Y_i = k|Z_i)$.

Then the weighted likelihood under case–control sampling is:

$$l_W(\theta|Y, Z, D) = \sum_{i=1}^{m+n} \log[P(Y_i|Z_i, D_i)] = \sum_{i=1}^{m+n} w_i(D_i) \log[P(Y_i = k|Z_i)]. \quad (6)$$

If (Y_i, Δ_i) is a (possibly censored) time-to-event outcome with failure time T_i and censoring time C_i, $Y_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i < C_i)$, then one may use a Cox proportional hazards model [21] for the estimation of $\theta = \beta$

$$\lambda_{T_i}(t|Z_i) = \lambda_0(t)\exp(\beta' Z_i) \quad (7)$$

with weighted log-partial-likelihood given by

$$l_W(Y|Z, D) = \sum_{i=1}^{m+n} w_i(D_i)\Delta_i\{\beta' Z_i - \log[\sum_{l=1}^{m+n} w_l(D_l) I(Y_i < Y_l)\exp(\beta' Z_l)]\}. \quad (8)$$

We propose the use of bootstrapping to estimate the standard error of the estimate of interest, $se(\hat{\theta})$, in any of the aforementioned scenarios. We select R samples from the empirical distribution of the original data. For each bootstrap replication, we apply the IPW method described above. Specifically, let (D, Y, Z) denote the data with empirical cdf f, and let (D_r^*, Y_r^*, Z_r^*) denote bootstrap replications for $r = 1, \ldots, R$. The first step is to fit a model to (D_r^*, Y_r^*, Z_r^*) using the weighted log likelihood (1) for each replication. The variance of the parameter estimate $\hat{\theta}$ is given by the estimated variance of the R bootstrap parameter estimates of θ. Confidence intervals may be generated by the percentile method, bias-corrected and accelerated (BCa) method, approximate bootstrap confidence interval (ABC) method, bootstrap-t, or a normal approximation [22,23]. Standard software, such as the boot package in R [24,25], can easily generate these estimates.

3. Simulation Study

3.1. General Setup

We simulated data under the framework in [8]. In order to have n cases and m controls, we generated $n_s = \frac{3\max(m,n)}{p_e}$ observations. This ensured there would be enough cases and controls in each dataset. For all subjects in this superset, $i = 1, \ldots, n_s$, we assume that the relationship between case–control status, D_i, the number of copies of the minor allele, Z_i, and the secondary phenotype, Y_i, is given by a logistic regression model based on the genetic profile and secondary phenotype. The distribution of each type of outcome and corresponding specific form of the logistic regression model are given in Sections 3.2–3.4 for binary, ordinal, and time-to-event outcomes, respectively. Each equation specifies that the probability of being a case of the primary disease (rather than a control) depends on the status of the genetic profile and the secondary phenotype. The inheritance model is assumed to be additive. We assumed a minor allele frequency of 30% and an estimated prevalence of D of 10% to approximate the 8.6% prevalence of TMD in the OPPERA prospective cohort study [26].

Finally, for each simulation run, we extracted the first n cases and m controls from the superset of n_s to comprise the simulated case–control dataset of size $m + n$. For each method and scenario, we estimated the value of the parameter relating to the secondary phenotype and the number of copies of the minor allele. (This was the log-odds ratio or log-hazard ratio depending on the type of outcome.) Average bias, empirical coverage, and average confidence interval width were compared between our method, and the naive methods that restrict to cases, restrict to controls, or adjust for case–control status using an indicator variable. We also compared the performance of the IPW with the GEE method of [2] when applicable, i.e., for continuous outcomes. All simulation types included 1000 datasets each with $n = 1000$ cases and $m = 1000$ controls. In each case, the average bias was estimated using the difference between the mean of the 1000 parameter estimates and the true parameter value. Analogously, the average confidence interval width produced by each method is equal to the average difference between the upper and lower confidence limits among the 1000 scenarios. Similarly, we calculated the empirical coverage probability produced by each method using the proportion of the 1000 simulations for which the true parameter value is contained in the confidence interval.

3.2. Continuous Phenotypes

For continuous secondary phenotypes, we assume a standard linear model with normally distributed errors, $\epsilon_i \sim N(0, \sigma^2)$,

$$Y_i = \beta_0 + \beta_1 Z_i + \epsilon_i, \tag{9}$$

where Z_i is defined in Section 3.1, and case status is defined by

$$logit[P(D_i = 1|Y_i, Z_i)] = \gamma_0 + \gamma_1 Z_i + \gamma_2 Y_i. \tag{10}$$

Our simulations included the parameters $\beta_0 = 0$, $\beta_1 = -0.12, -0.5, -1, -2$ and $\gamma_1 = \log(2), \log(3), \log(5), \log(10)$, $\gamma_2 = \log(2)$ and $\sigma^2 = 1$.

In order to keep the prevalence approximately constant, we set the value of γ_0 separately for each simulation, according to

$$\hat{\gamma_0} = \log\left[\frac{p_e \exp(-\tilde{X})}{1 - p_e}\right], \tag{11}$$

where

$$\tilde{X} = \gamma_1 \bar{Z} + \gamma_2 \bar{Y}, \tag{12}$$

$p_e = 0.1$, and $\bar{Z} = \frac{1}{n_s}\sum_{i=1}^{n_s} Z_i$ and $\bar{Y} = \frac{1}{n_s}\sum_{i=1}^{n_s} Y_i$ are the averages of the $i = 1, \ldots, n_s$. Z_i and Y_i values.

The parameter of interest was β_1. Considering continuous outcomes facilitated the comparison of the IPW with the GEE method of [2].

Simulations with continuous outcomes yielded the following results. In all scenarios, our method had negligible bias and coverage rate near 95%, as desired. Performance in terms of bias, coverage, and confidence interval width was comparable to that of [2]. The bootstrapping IPW method had comparable bias to the method of [2] and less bias than all other methods. Details are found in Table 1. This shows that when competing methods are applicable, our method does at least as well as, if not better than the competitors.

3.3. Ordinal Phenotypes

We tested four scenarios for ordinal phenotypes with 3 levels. For simplicity, we will denote these as $Y_i = 0$, $Y_i = 1$ and $Y_i = 2$. In general, we generated the ordinal outcomes with the following probabilities

$$p_0 = P(Y_i = 0) = \exp(\zeta_0 + \beta Z_i)/(1 + \exp(\zeta_0 + \beta Z_i)),$$

$$p_1 = P(Y_i = 1) = \exp(\zeta_1 + \beta Z_i)/(1 + \exp(\zeta_1 + \beta Z_i)) - p_0,$$

$$p_2 = P(Y_i = 2) = 1 - p_1 - p_0.$$

For all subjects, $i = 1, \ldots, m+n$, we assume that the relationship between case–control status, D_i, the number of minor allele copies, Z_i, and the secondary phenotype, Y_i, is given by the following logistic regression model

$$logit[P(D_i = 1|Y_i, Z_i)] = \gamma_0 + \gamma_1 Z_i + \gamma_{2a} I(Y_i = 1) + \gamma_{2b} I(Y_i = 2), \tag{13}$$

where

$$\bar{Y} = \gamma_{2a}\left(\frac{\sum_{i=1}^{m+n} I(Y_i = 1)}{m+n}\right) + \gamma_{2b}\left(\frac{\sum_{i=1}^{m+n} I(Y_i = 2)}{m+n}\right) \tag{14}$$

is used to define the average outcome in Equation (12) and thus the value of γ_0 in (11).

For the four scenarios, we used the following parameters:

1. $\beta = 0.5$, $\zeta_0 = 1.5$, $\zeta_1 = 2.5$, and $\gamma_1 = \gamma_{2a} = \gamma_{2b} = \log(2)$
2. $\beta = 1$, $\zeta_0 = 0$, $\zeta_1 = 1$, and $\gamma_1 = \gamma_{2a} = \gamma_{2b} = \log(2)$
3. $\beta = 0.75$, $\zeta_0 = 1$, $\zeta_1 = 2$, and $\gamma_1 = \gamma_{2a} = \gamma_{2b} = \log(2)$
4. $\beta = 0.5$, $\zeta_0 = 1.5$, $\zeta_1 = 2.5$, $\gamma_1 = \gamma_{2a} = \log(2)$, and $\gamma_{2b} = \log(3)$.

Our weighted bootstrap method has less bias than all other methods. Using controls only yielded close but lower coverage in scenarios 1, 2, and 4, and identical coverage in scenario 3. However, this finding may be an artifact of the relatively low simulated prevalence of cases and may not be replicated for higher population case rates. None of the other methods have adequate coverage for these ordinal simulations. Results are given in Table 2.

Table 1. Results of simulations for intermediate continuous phenotypes.

Parameters	$\beta_1 = -0.12$ $\gamma_1 = \log(2)$		$\beta_1 = -0.12$ $\gamma_1 = \log(3)$		$\beta_1 = -0.12$ $\gamma_1 = \log(5)$		$\beta_1 = -0.12$ $\gamma_1 = \log(10)$		$\beta_1 = -0.5$ $\gamma_1 = \log(2)$		$\beta_1 = -1$ $\gamma_1 = \log(2)$		$\beta_1 = -2$ $\gamma_1 = \log(2)$	
	Bias	**Cover**	**Bias**	**Cover**	**Bias**	**Cover**	**Bias**	**Cover**	**Bias**	**Cover**	**Bias**	**Cover**	**Bias**	**Cover**
LM	−0.044	0.728	0.059	0.577	0.061	0.532	0.048	0.666	0.028	0.883	0	0.944	−0.058	0.643
LM, controls only	−0.039	0.875	−0.073	0.698	−0.108	0.444	−0.151	0.206	−0.022	0.926	−0.002	0.939	0.037	0.892
LM, cases only	−0.056	0.761	−0.099	0.394	−0.166	0.06	−0.256	0	−0.028	0.911	0.002	0.949	0.061	0.794
LM adjusted for case status	0.048	0.696	−0.088	0.235	−0.14	0.021	−0.208	0	−0.025	0.877	0	0.945	0.047	0.736
Monsees	−0.002	0.950	−0.001	0.951	0.001	0.948	0.001	0.949	0.001	0.952	−0.001	0.938	−0.003	0.949
Bootstrap	−0.002	0.950	−0.001	0.956	0.001	0.944	0.001	0.946	0.001	0.950	−0.001	0.937	−0.003	0.944
	CI Width (Valid Methods Only)													
Monsees	0.164		0.160		0.155		0.148		0.166		0.167		0.168	
Bootstrap	0.162		0.160		0.154		0.147		0.165		0.167		0.168	

LM = Linear Model.

Table 2. Results of simulations for intermediate ordinal phenotypes.

Method	**Result**							
	Scenario 1		**Scenario 2**		**Scenario 3**		**Scenario 4**	
	Bias	**Coverage**	**Bias**	**Coverage**	**Bias**	**Coverage**	**Bias**	**Coverage**
Naive	−0.054	0.909	−0.068	0.836	−0.053	0.910	−0.069	0.871
Controls only	0.075	0.937	0.051	0.937	0.063	0.948	0.093	0.933
Cases only	0.805	0	−0.815	0	0.226	0	0.903	0
Adjusted for case status	0.057	0.904	0.027	0.937	0.054	0.913	0.101	0.824
Bootstrap	0.020	0.943	0.006	0.944	0.015	0.948	0.015	0.951
	CI Width (Valid Methods Only)							
Bootstrap	0.519		0.399		0.477		0.512	

3.4. Time-to-Event Phenotypes

Survival outcomes were generated as in [27] with exponential failure and censoring times. The failure time T_i satisfies Equation (7) where $\lambda_0(t) = 1$ for all t, $\beta = -1$ and Z_i is the number of copies of the minor allele. The censoring time was exponential with shape parameter 2. The parameter of interest was β and the outcome of interest was (Y_i, Δ_i) where $Y_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i < C_i)$. This yielded about 84% censoring. High censoring was used in simulations to parallel findings in the application of interest, the OPPERA study [26].

The case status was similar to Equation (10) for continuous outcomes, but instead depended on the true failure time rather than the observed time as follows

$$logit[P(D_i = 1|Z_i, T_i)] = \gamma_0 + \gamma_1 Z_i + \gamma_2 T_i. \tag{15}$$

The value of γ_0 was set by Equation (11) with $\tilde{X}$ defined by Equation (12) and $\bar{X}$ and $\bar{Y}$ defined as in Section 3.2. We used $\gamma_1 = \gamma_2 = \log(2)$.

For time-to-event outcomes, our method retained empirical coverage around 95% and had less bias than all other methods. None of the other methods have adequate coverage, except the method that adjusts for case status. However, the latter method was overly conservative. See Table 3 for details. Other methods do not apply for this type of outcome. Consequently, no comparison was made.

Table 3. Results of simulations for intermediate time-to-event phenotypes.

Method	**Bias**	**Coverage**
Naive	−0.457	0.006
Controls only	0.272	0.396
Cases only	−0.800	0.225
Adjusted for case status	0.091	1.000
Bootstrap	−0.017	0.944
	CI Width (Valid Methods Only)	
Bootstrap	0.439	

4. Data Application

We applied the method to the baseline case–control genetic study within OPPERA. The prospective cohort study consisted of 3263 healthy TMD-free volunteers and 186 volunteers determined at baseline to have TMD. All 186 cases were retained and 1633 controls were randomly selected for the baseline case–control study.

The covariates of interest were 2924 SNPs collected in a genetic association study of 3037 participants [28,29]. The outcome was the number of co-morbid conditions, categorized as either zero, one, or more than one co-morbid condition. Upon enrollment in

OPPERA, participants self-reported by checking experience with a list of 20 conditions on the Comprehensive Pain and Symptom Questionnaire (CPSQ). Examples of chronic pain conditions include arthritis, fibromyalgia, irritable bowel syndrome, include chronic pelvic pain, among others. All cases and 1626 controls filled out the CPSQ, [30]. Combining these yielded 166 cases and 1435 controls with information available on both their history of comorbid conditions and their genetic profiles. Recruited from 4 study sites and ranging from 18 to 44 years in age, these 1601 individuals comprise the proceeding analysis. For more details of the OPPERA study design, see [3,4,28,30].

For each SNP with less than 5% missing values, we fit a proportional odds model to the data, adjusting for study site, age, gender, and two racial eigenvectors calculated as described in [29].

We collected the p-values and created Q–Q plots of the negative logarithm of the p-values for the standard unweighted method and for our weighted bootstrapping method. The plots indicate that neither method found any SNPs that were significantly associated with comorbid pain conditions after adjusting for multiple comparisons. See Figures 1 and 2.

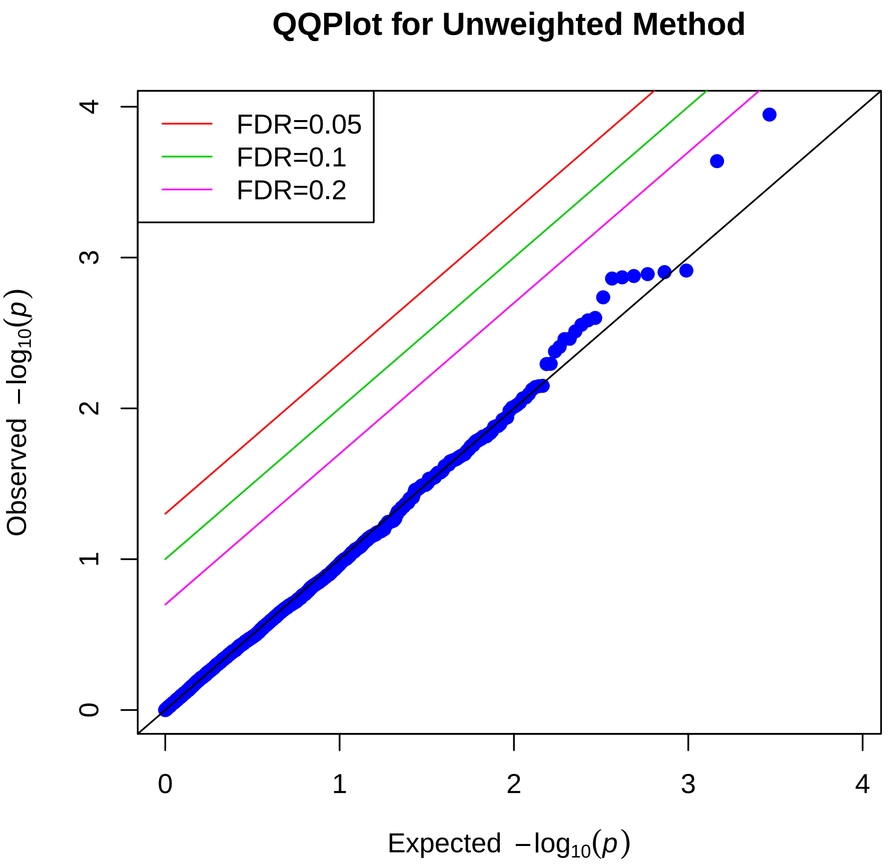

Figure 1. Q–Q plot for the unweighted method.

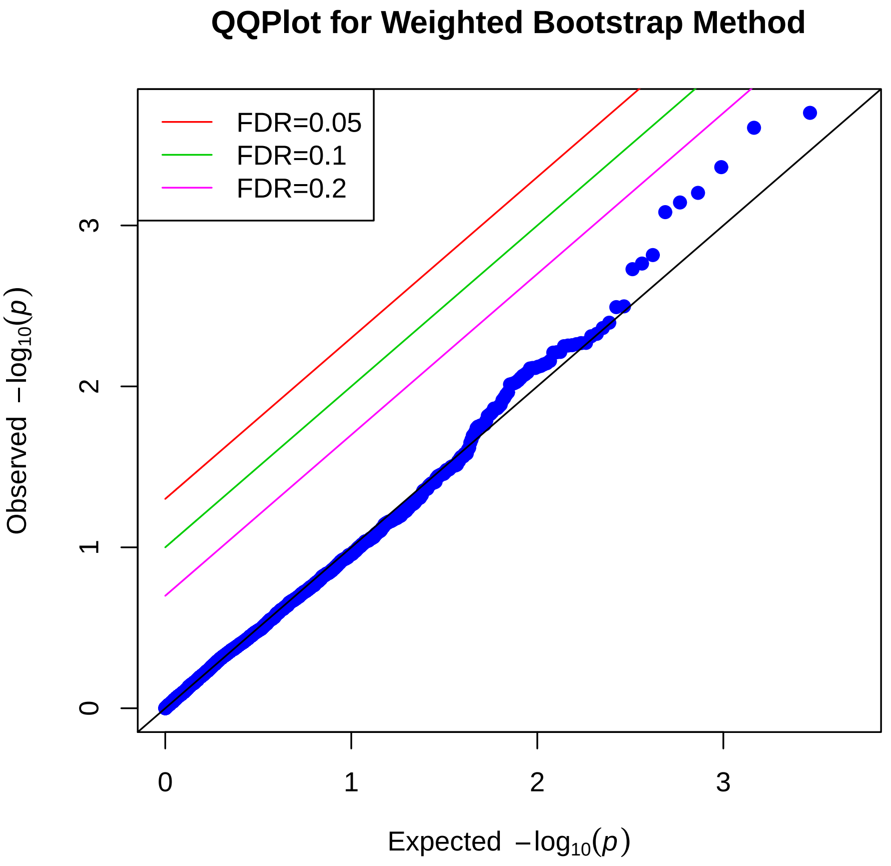

Figure 2. Q–Q plot for the weighted bootstrap method.

5. Discussion

Our proposed method for the analysis of intermediate phenotypes in case–control studies of genetic data is simple and easily implemented in standard software. The simulation results indicate that it is approximately unbiased, and has comparable coverage and confidence interval width to the method of IPW with GEE [2]. Under situations in which the retrospective likelihood-based method [8] is applicable, their method should be more powerful than our proposed method.

Our method is general enough to allow for the analysis of multiple outcomes simultaneously and of outcomes for which previous methodology may not be applicable. This generality allowed the analysis of the phenotype of interest from OPPERA applied in this manuscript, namely the number of comorbid pain conditions. Although two prior methods mention potential extensions to count data [7,15], such extensions have not been explicitly demonstrated, and thus their relative performance remains unknown. Currently, our method is the only known viable way to evaluate secondary time-to-event outcomes in case–control studies. In addition, multiple outcomes could be analyzed using standard multivariate methods but weighting each observation as described in this paper, and bootstrapping to estimate the standard error. More importantly, the method can be applied to complicated test statistics where there is no existing formula for the standard error, such as the many test statistics employed in sequencing studies. We expect the methods to perform well for different case to control ratios when the sample sizes for cases and controls are sufficient. We also expect the methods to perform well with multiple but a relatively small number of markers compared with the sample size when sample sizes for cases and controls are sufficient. In situations where the number of genetic markers exceeds the sample size, our methods may not be applicable. It is worth noting that our procedure is computationally non-trivial due to the use of bootstrapping, but the runtime is reasonable for modern computers. For 1000 runs of the survival scenario with 100 bootstrap replications, for instance, the output of proc.time was 773.288, or about 13 minutes of total elapsed time, with an average of 0.77 seconds per run using 100 bootstrap replications.

Author Contributions: Conceptualization, N.C.B., J.C. and E.B.; methodology, N.C.B., J.C. and E.B.; data curation, N.C.B., S.S., L.D., G.D.S. and E.B; writing—original draft preparation, N.C.B.; writing—review and editing, N.C.B., J.C., S.S., L.D., G.D.S. and E.B. All authors have read and agreed to the published version of the manuscript.

Funding: The OPPERA study was supported by NIH/NIDCR grant U01DE017018. Naomi Brownstein was supported by NIH/NIEHS T32ES007018 and NSF Graduate Research Fellowship Program grant 0646083. Jianwen Cai was supported by NIH/NCI grant P01CA142538 and NIH/NIEHS grant R01ES021900. Eric Bair was supported by NIH/NIDCR grant R03DE023592, NIH/NCATS grant UL1TR001111, and NIH/NIEHS grant P03ES010126.

Institutional Review Board Statement: The OPPERA study was conducted according to the guidelines of the Declaration of Helsinki and was reviewed and approved by the Institutional Review Boards of the OPPERA study sites.

Informed Consent Statement: OPPERA study participants provided informed, signed consent to participate in the OPPERA study.

Data Availability Statement: The data and code that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge and thank the principal investigators of the OPPERA study, namely, William Maixner, Bruce Weir, Richard Ohrbach, Roger Fillingim, Joel Greenspan, and Ronald Dubner.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

OPPERA	Orofacial Pain: Prospective Evaluation and Risk Assessment
TMD	Temporomandibular disorders
IPW	Inverse probability weighting
Q–Q	Quantile–quantile

References

1. Prentice, R.L.; Pyke, R. Logistic Disease Incidence Models and Case-Control Studies. *Biometrika* **1979**, *66*, 403–411. [CrossRef]
2. Monsees, G.M.; Tamimi, R.M.; Kraft, P. Genome-wide association scans for secondary traits using case-control samples. *Genet. Epidemiol.* **2009**, *33*, 717–728. [CrossRef] [PubMed]
3. Slade, G.D.; Bair, E.; By, K.; Mulkey, F.; Baraian, C.; Rothwell, R.; Reynolds, M.; Miller, V.; Gonzalez, Y.; Gordon, S.; et al. Study Methods, Recruitment, Sociodemographic Findings, and Demographic Representativeness in the OPPERA Study. *J. Pain* **2011**, *12*, T12–T26. [CrossRef]
4. Maixner, W.; Diatchenko, L.; Dubner, R.; Fillingim, R.B.; Greenspan, J.D.; Knott, C.; Ohrbach, R.; Weir, B.; Slade, G.D. Orofacial Pain Prospective Evaluation and Risk Assessment Study—The OPPERA Study. *J. Pain* **2011**, *12*, T4–T11. [CrossRef]
5. Richardson, D.B.; Rzehak, P.; Klenk, J.; Weiland, S.K. Analyses of case-control data for additional outcomes. *Epidemiology* **2007**, *18*, 441–445. [CrossRef] [PubMed]
6. Xing, C.; McCarthy, J.M.; Dupuis, J.; Adrienne Cupples, L.; Meigs, J.B.; Lin, X.; Allen, A.S. Robust analysis of secondary phenotypes in case-control genetic association studies. *Stat. Med.* **2016**, *35*, 4226–4237. [CrossRef]
7. Li, F.; Allen, A.S. Secondary analysis of case-control association studies: Insights on weighting-based inference motivate a new specification test. *Stat. Med.* **2020**, *39*, 2869–2882. [CrossRef]
8. Lin, D.Y.; Zeng, D. Proper analysis of secondary phenotype data in case-control association studies. *Genet. Epidemiol.* **2009**, *33*, 256–265. [CrossRef]
9. Ghosh, A.; Wright, F.A.; Zou, F. Unified analysis of secondary traits in case–control association studies. *J. Am. Stat. Assoc.* **2013**, *108*, 566–576. [CrossRef]

10. Li, H.; Gail, M. H.and Berndt, S.; Chatterjee, N. Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genet. Epidemiol.* **2010**, *34*, 427–433. [CrossRef]
11. He, J.; Li, H.; Edmondson, A.C.; Rader, D.J.; Li, M. A gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics* **2011**. [CrossRef] [PubMed]
12. Wei, J.; Carroll, R.J.; Müller, U.U.; Keilegom, I.V.; Chatterjee, N. Robust estimation for homoscedastic regression in the secondary analysis of case–control data. *J. R. Stat. Soc. Ser. B* **2013**, *75*, 185–206. [CrossRef] [PubMed]
13. Ma, Y.; Carroll, R.J. Semiparametric estimation in the secondary analysis of case-control studies. *J. R. Stat. Soc. Ser. B* **2016**, *78*, 127. [CrossRef] [PubMed]
14. Schifano, E.D.; Li, L.; Christiani, D.C.; Lin, X. Genome-wide association analysis for multiple continuous secondary phenotypes. *Am. J. Hum. Genet.* **2013**, *92*, 744–759. [CrossRef]
15. Tchetgen Tchetgen, E.J. A general regression framework for a secondary outcome in case–control studies. *Biostatistics* **2014**, *15*, 117–128. [CrossRef]
16. Wang, J.; Shete, S. Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genet. Epidemiol.* **2011**, *35*, 190–200. [CrossRef]
17. Wang, J.; Shete, S. Power and type i error results for a bias-correction approach recently shown to provide accurate odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genet. Epidemiol.* **2011**, *35*, 739–743. [CrossRef]
18. Li, H.; Gail, M.H. Efficient adaptively weighted analysis of secondary phenotypes in case-control genome-wide association studies. *Hum. Hered.* **2012**, *73*, 159–173. [CrossRef]
19. Zhou, F.; Zhou, H.; Li, T.; Zhu, H. Analysis of secondary phenotypes in multigroup association studies. *Biometrics* **2020**, *76*, 606–618. [CrossRef]
20. Agresti, A. *Categorical Data Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2003; Volume 482.
21. Cox, D.R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B* **1972**, *34*, 187–202. [CrossRef]
22. DiCiccio, T.J.; Efron, B. Bootstrap confidence intervals. *Stat. Sci.* **1996**, *11*, 189–228. [CrossRef]
23. Davison, A.C.; Hinkley, D.V. *Bootstrap Methods and Their Application*; Number 1; Cambridge University Press: Cambridge, UK, 1997.
24. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018.
25. Canty, A.; Ripley, B.D. Boot: Bootstrap R (S-Plus) Functions, 2017. R Package Version 1.3-20. Available online: https://cran.r-project.org/web/packages/boot/index.html (accessed on 30 December 2021).
26. Bair, E.; Brownstein, N.C.; Ohrbach, R.; Greenspan, J.D.; Dubner, R.; Fillingim, R.B.; Maixner, W.; Smith, S.B.; Diatchenko, L.; Gonzalez, Y.; et al. Study protocol, sample characteristics, and loss to follow-up: the OPPERA prospective cohort study. *J. Pain* **2013**, *14*, T2–T19. [CrossRef] [PubMed]
27. Bender, R.; Augustin, T.; Blettner, M. Generating survival times to simulate Cox proportional hazards models. *Stat. Med.* **2005**, *24*, 1713–1723. [CrossRef] [PubMed]
28. Smith, S.B.; Maixner, D.W.; Greenspan, J.D.; Dubner, R.; Fillingim, R.B.; Ohrbach, R.; Knott, C.; Slade, G.D.; Bair, E.; Gibson, D.G.; et al. Potential Genetic Risk Factors for Chronic TMD: Genetic Associations from the OPPERA Case Control Study. *J. Pain* **2011**, *12*, T92–T101. [CrossRef] [PubMed]
29. Smith, S.B.; Mir, E.; Bair, E.; Slade, G.D.; Dubner, R.; Fillingim, R.B.; Greenspan, J.D.; Ohrbach, R.; Knott, C.; Weir, B.; et al. Genetic variants associated with development of TMD and its intermediate phenotypes: The genetic architecture of TMD in the OPPERA prospective cohort study. *J. Pain* **2013**, *14*, T91–T101. [CrossRef] [PubMed]
30. Ohrbach, R.; Fillingim, R.B.; Mulkey, F.; Gonzalez, Y.; Gordon, S.; Gremillion, H.; Lim, P.F.; Ribeiro-Dasilva, M.; Greenspan, J.D.; Knott, C.; et al. Clinical Findings and Pain Symptoms as Potential Risk Factors for Chronic TMD: Descriptive Data and Empirically Identified Domains from the OPPERA Case-Control Study. *J. Pain* **2011**, *12*, T27–T45. [CrossRef] [PubMed]