# Analyses of Case–Control Data for Additional Outcomes

*David B. Richardson,* Peter Rzehak,† Jochen Klenk,† and Stephan K. Weiland†*

**Abstract:** Consider a case–control study in which prevalent cases of a given disease define the index series and members of the base population without the disease are sampled to provide the referent series. Information on a set of explanatory variables (eg, genotypes) is collected at great cost for cases and controls. The objective of the study is to evaluate the relationship between case status and the explanatory variables. Subsequently, an investigator notes that the prevalence of a second disease was measured for the members of the index and referent series. The investigator wishes to make efficient use of the available data by assessing the relationship between this second disease and the set of explanatory variables. In this paper, we discuss 2 analytic approaches that might be used to assess associations between the explanatory variables and an outcome other than the original disease. One is through the inclusion of a design variable for original disease status as a covariate; and, the second is through weighted logistic regression using the inverse of the sampling fractions as the weights. The latter approach allows the investigator to derive an estimate of association between the explanatory variables and the second disease without adjustment for the first disease. Weighted logistic regression methods are readily implemented using available statistical packages.

(*Epidemiology* 2007;18: 441–445)

The conduct of epidemiologic research can be expensive. When possible, it is advantageous to develop analytic methods that maximize the efficiency of such studies, and permit investigation of additional study questions using existing data. The context for the analytical method discussed in this paper is as follows. A large multicenter cross-sectional study of the prevalence of childhood asthma was conducted with a focus on environmental and genetic risk factors.[1] Subsequently, a nested case–control study was conducted in which cases were sampled from children in the study base who reported wheezing during the last 12 months. Controls were sampled from children in the study base who had not reported wheezing during the last 12 months. Biologic measurements, such as bronchial responsiveness and serum immunoglobulin E levels (IgE), were made on the members of this case–control study. These data were used to assess the relationship between predictors, such as IgE, and the prevalence odds of self-reported wheezing.

A variety of allergy-related health endpoints other than self-reported wheezing were also assessed in the study described above. The question arose of how data from the nested case–control study could be used to analyze associations between the explanatory variables and other health endpoints. This is a generic question in case–control analyses. The problem can be viewed as analysis of data derived from a case–control study using disproportionate stratified subsamples of the study base (ie, samples of those with and without some characteristic assessed in the full study base).[2,3] Ignoring the sampling structure can lead to severely biased results. A solution to this problem is simply to derive estimates of prevalence odds ratios that are adjusted for the matching factor (case status) that defined the original index case series for the nested case–control study. Prentice and Pyke[4] have shown that when an unconditional logistic regression model is fit to data derived from a cumulative case–control study design (in which an investigator selects controls from among those who remain disease-free at the end of the study), only the model intercept is biased. However, the effect measure that is obtained is the estimate of the change in log prevalence odds of disease per unit change in exposure adjusted for the case status that defined the original case series. This result is not necessarily an estimate of the effect measure of interest and, in some instances, it may not be a valid effect estimate at all. For example, if the disease that defined the original case series (wheezing) is an intermediate in the causal relationship between exposure and another outcome of interest (eg, clinically diagnosed asthma) then adjustment for wheezing in analyses of associations between IgE and clinically-diagnosed asthma will result in a biased effect estimate. An alternative, simply to restrict the analysis to the noncases in the original study (ie, the controls), has similar implications. The effect measure that is obtained is conditioned on the case status that defined the case series for the original study; furthermore, discarding information on the original case series seems inefficient. Therefore, we were interested in exploring other analytic approaches to this problem.

Submitted 18 October 2006; accepted 2 February 2007; posted 30 April 2007.
From the *Department of Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, NC; and †Institute of Epidemiology, Ulm University, Germany.
This work has been funded by the 5th framework program (FP5) of the European Union (contract: QLK4-CT-1999-01288). Dr. Richardson was supported by a grant from the Alfred Krupp Kolleg Greifswald Foundation and Alfred Krupp von Bohlen und Halbach Foundation.

***Editor's note:*** *Stephan K. Weiland, head of the Institute of Epidemiology, died unexpectedly on 19 March 2007. A Remembrance will appear in the September 2007 issue of* EPIDEMIOLOGY.

Correspondence: David Richardson, Department of Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, NC 27599-7435. E-mail: david.richardson@unc.edu.

ISSN: 1044-3983/07/1804-0441
DOI: 10.1097/EDE.0b013e318060d25c

## METHODS

Consider a case–control study in which prevalent cases of disease $D_1$ define a case series. Subjects who were free of disease $D_1$ at the time of the prevalence survey were randomly sampled to provide a referent series. Information on a vector of explanatory variables, $X$, was collected for members of the index and referent series. These data were subsequently used to investigate associations between $X$ and an outcome other than $D_1$, denoted $D_2$.

An analysis of the association between $X$ and $D_2$ can be viewed as an analysis of case–control data with biased sampling. While the literature on this topic is substantial, biased sampling is usually a term used to discuss matching, or sampling, with respect to a covariate of interest. In such instances, the data analyst intends to control for the factor that defines sampling strata in the data; therefore, a commonly-advocated approach is to stratify on the covariate. If the sampling fractions are known, the data analyst can use the results of a logistic regression model fitting not only to estimate odds ratios but, by adjusting the model intercept for the stratum-specific sampling fractions of cases and controls, to estimate odds or predicted risks for specific covariate patterns. The literature on biased sampling also includes discussions in which biased sampling strategies are employed to ensure that sampling probabilities are known, so that a data analyst can assess effect modification between an exposure of interest and factors that define sampling strata on scales other than the multiplicative.[3,5] Langholz et al[6] have proposed over sampling with respect to the primary exposure of interest (termed counter-matching). The counter-matching approach may increase the statistical precision of effect measures if the exposure distribution is skewed.

In contrast, we are considering study data that have been sampled with respect to a factor $D_1$ that is not of interest as an exposure, nor necessarily as a covariate. If we do not know the sampling fraction of cases of disease $D_2$, then we have a "convenience sample" of cases of $D_2$ observed within an index series and referent series defined by disease $D_1$. In the sections below we discuss 2 approaches to deriving estimates of the prevalence odds ratio for disease $D_2$, contrasting subjects with covariate patterns defined by explanatory variables $X$.

### Adjustment for $D_1$

One approach to analysis of the relationship between disease $D_2$ and an explanatory variable $X$ (in a case–control study where the index series was defined by disease $D_1$) is via a logistic regression model where a design variable is included to adjust for disease status $D_1$.[4,7] This is a standard unconditional logistic regression model in which one of the covariates in the model is a binary indicator variable, *stratum*, for the disease $D_1$ that defines the index case series,

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i + \beta_2 stratum_i$$

where $\pi$ is the probability of $D_2 = 1$, given the model covariates, and

$$Ln\hat{L}(\beta;X) = \sum_{i=1}^{n}(y_i \ln\pi_i + [1-y_i]\ln[1-\pi_i])$$

is the log-likelihood-function from which the estimates of the logistic regression coefficients $\beta$ are derived, given that $y$ denotes the value of the outcome variable.[7] The logistic model fitting will provide an estimate of the association between $X$ and $D_2$ in a model that conditions on a design variable for $D_1$, which determined the study's sampling scheme.

### Stratum-Weighted Logistic Regression

An alternative analytical approach is to replace the log likelihood above by a weighted sum using weights $w_h$ given by the reciprocal of the selection probability for the respective stratum,

$$Ln\hat{L}(\beta;X) = \sum_{h=1}^{H} w_h \sum_{i=1}^{n_h}(y_{hi}\ln\pi_{hi} + [1+y_{hi}]\ln[1-\pi_{hi}])$$

where $h$ indexes case–control ($D_1$) status.[2,8,9] For example, for a pair-matched case–control study in which all cases in the study base are ascertained, the weight (inverse selection probability) for cases is 1 and the weight for controls is the number of noncases in the study base divided by the number of controls selected.

Rather than adjusting for the design variable for the stratum, an estimator accounting for the stratified sampling is obtained with stratum-specific sampling weights that are part of the log-likelihood-function. An advantage of this latter approach is that the data analyst can derive an estimate of association between $X$ and $D_2$ that is not adjusted for $D_1$. As would be observed in logistic model fittings to data for the entire study base, the results obtained via these 2 approaches will diverge if $D_1$ behaves like a confounder or an intermediate variable in the association between $X$ and $D_2$.

### Empirical Example

An empirical example is used to illustrate 2 points. First, stratum-weighted logistic regression allows a data analyst to derive an unbiased estimate of association between $X$ and $D_2$ that is not adjusted for $D_1$. Second, the estimate of association derived from stratum-weighted logistic regression may differ from that obtained via a standard unconditional logistic regression analysis in a model that adjusts for $D_1$, via analyses that ignore the original sampling structure, and via analyses that use data for the referent series only.

Study data on symptoms of respiratory and atopic diseases as well as on various risk factors in children (ages 9–11 years) were collected by parental questionnaires conducted in 1994–1995 in Munich, Germany. Details and results of this study are published elsewhere.[1,10] In Munich, measurements of blood samples (eg, specific serum IgE) and physical examinations (skin prick testing or bronchial challenges) were collected in an unstratified random sample of the population. We analyzed the data from Munich for all children with valid information on specific IgE-levels, self-reported hay fever, and self-reported wheezing (n = 1630). The study was part of a worldwide collaboration.[1] The protocol allowed a study center to perform IgE analyses only

in a nested case–control study that included disproportionate stratified subsamples of children based upon disease status; most study centers have chosen the latter approach. As study instruments and methods were the same for all centers worldwide, we used the data from Munich to illustrate the effect of different approaches to analyze disproportionate stratified subsamples.

For illustrative purposes we present a study in which self-reported hay fever is the disease that defined the index case series. We use these case–control data to assess the association between allergen-specific IgE levels (categorized as $<0.7$ vs. $\geq 0.7$ kU/L) and another outcome of interest (self-reported wheezing). A standard unconditional logistic regression model was used to derive a parameter estimate for the IgE-wheezing association ($\hat{\beta}_{ul}$) using data for the full study sample. This unconditional logistic regression model included a single parameter for IgE, defined as a dichotomous variable. Next, we drew a random sample of 100 children who reported hay fever (cases) and 200 children without hay fever (controls) from these data. Figure 1 illustrates the structure of the nested case–control study. The SAS v. 9.1 statistical package (PROC SURVEYLOGISTIC; SAS Institute, Cary, NC) was used to derive a parameter estimate for the IgE-wheezing association $\hat{\beta}_{swl}$ via a stratum-weighted logistic regression model applied to the nested case–control data. The regression model included a single parameter for IgE, defined as a dichotomous variable, and the outcome variable was a binary indicator of wheezing. The process of selecting cases and controls defined by hay fever status, and deriving an estimate of the IgE-wheezing association, was repeated over 1000 iterations, after which the average parameter estimate, $E(\hat{\beta}_{swl})$, was calculated. Standard logistic regression models were also fitted to the data for the nested case–control samples to obtain parameter estimates that were adjusted for hay fever status. Lastly, to illustrate the effect of ignoring the sampling scheme, standard logistic regression models that included a single parameter for IgE, defined as a dichotomous variable, were fitted to the data for the nested case–control samples; and, to illustrate the effect of analyzing the control series only, standard logistic regression models that included a single parameter for IgE, defined as a dichotomous variable, were fitted to the data for the control series of the nested case–control samples.

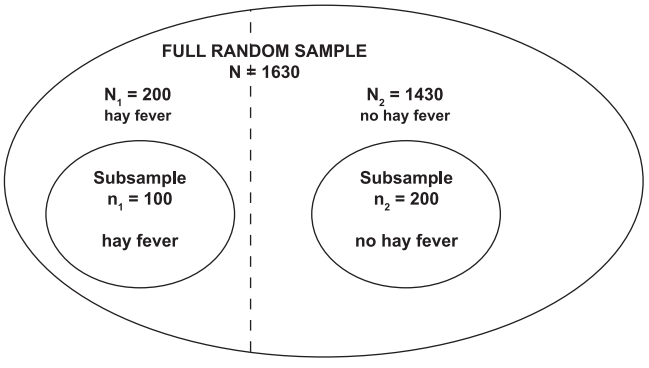


**FIGURE 1.** Nested case–control study of childhood hay fever where cases and controls were sampled from a study of 1630 German children.

## RESULTS

The estimate of the prevalence odds ratio of wheezing (when comparing high IgE to low IgE study members) was derived via stratum-weighted logistic regression analyses of nested case–control data originally collected to study hay fever. Figure 2 provides a histogram of the parameter estimates for the IgE-wheezing association $\hat{\beta}_{swl}$ obtained via 1000 iterations of this procedure. As illustrated, the parameter estimates from weighted logistic regression were symmetrically distributed around a median value of 1.35. There was no evidence of skewness in this distribution and the average parameter estimate ($E(\hat{\beta}_{swl}) = 1.38$) closely approximated the result obtained via standard unconditional logistic regression using the full study sample ($\hat{\beta}_{ul} = 1.34$, $\text{se}[\hat{\beta}_{ul}] = 0.18$). The averaged standard error of the parameter estimates obtained via stratum-weighted logistic regression for the IgE-wheezing association was 0.47.

Fitting of a standard unconditional logistic regression model with a design variable for hay fever produces a parameter estimate for the association between IgE and self-reported wheezing that is adjusted for hay fever status. The average parameter estimate obtained via a standard logistic regression model that adjusts for hay fever status (1.13) differs from the result obtained via stratum-weighted logistic regression analyses. The average standard error of the param-

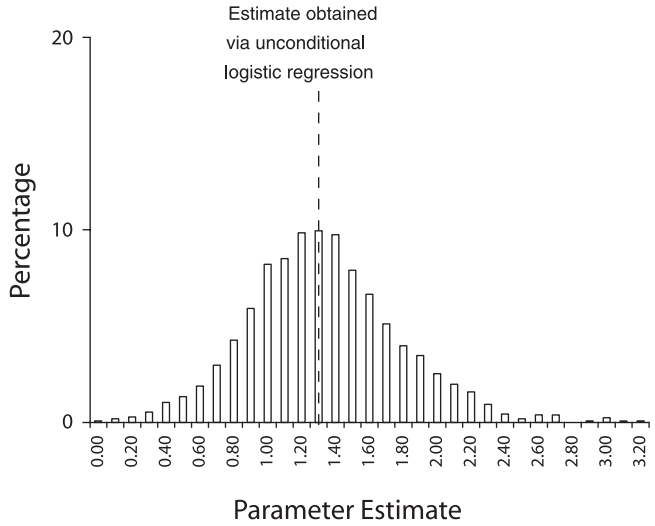


**FIGURE 2.** Histogram of the distribution of weighted logistic regression parameter estimates for the association between IgE level ($<0.7$ versus $\geq 0.7$) and self-reported wheezing (yes vs. no). Values obtained after 1000 replications of nested case–control analyses conducted by sampling 100 cases of children who reported hay fever and 200 children who did not report hay fever. Dashed line indicates the estimate of association between IgE level and self-reported wheezing obtained via standard unconditional logistic regression using data for the full study sample of 1630 German children.

eter estimates for the IgE-wheezing association obtained via fitting of a standard unconditional logistic regression model with a design variable for hay fever was 0.46. However, as discussed above, the scientific interpretation of an association between IgE and self-reported wheezing conditioning on hay fever status may be of little or no interest to an investigator.

In order to illustrate the effect of simply ignoring the disproportionate stratified sampling scheme, we examined the results obtained by fitting a standard logistic regression model without a design variable for hay fever status. The average parameter estimate for the association between IgE and self-reported wheezing was 1.51, with an average standard error of 0.42. Lastly, to illustrate the effect of analyzing the control series only, we fitted standard logistic regression models to the control series. The average parameter estimate for the association between IgE and self-reported wheezing was 1.13, and the average standard error of the parameter estimates was 0.95.

## DISCUSSION

There are at least 2 options when analyzing case–control data for an outcome other than the one that defined the index series. One is to adjust for the disease that defined the original case series. This approach produces estimates of association that are adjusted for a factor (the status of study members with respect to the disease that defined the index series) that is not necessarily of interest as a covariate in the association under study.

Another option is to use stratum-weighted logistic regression. This type of regression is readily-implemented via the SAS statistical package. This approach is generally applicable in analyses of 2-stage data when it is not desirable to adjust for the variable that defined the strata for the second-stage sampling. In the case–control context, stratum-weighted logistic regression may be particularly useful in situations where the disease that defined the original case series is a potential intermediate in the association between an exposure of interest and the disease under study (or situations in which the effects of predictor variables are confounded with the strata variables defining the index case series).

Stated differently, if nested case–control data are analyzed for an outcome other than the disease that defined the index case series, the data analyst will typically need to account for sample selection probabilities in a setting in which the selection probabilities may relate to values of the response variable. Conditioning on the values of the design variable (index case status) is often recommended as a plausible approach for dealing with the problem of unequal selection probabilities. However, inclusion of the design variables as covariates may undermine the scientific rationale for the regression analysis.[9] In such scenarios, this information about selection probabilities may be represented by sampling weights.

There are some issues to keep in mind when considering this approach. First, it is necessary to specify the sampling fractions employed in the original study. If the original case–control study was nested within an enumerated cohort, then sampling fractions can be calculated directly. In a population-based case–control study, these values may be estimable by reference to external information; the ratio of the sampling fraction of cases to the sampling fraction of controls may be easier to estimate in a population-based study than absolute sampling fractions.[11] This ratio can be used to derive weights for stratum-weighted logistic regression by assigning a weight of 1.0 to cases and a weight equal to the ratio of sampling fractions to the controls. In a matched case–control study, the stratum-specific sampling fractions would be employed in a similar manner. Second, this approach is best applied to large case–control series, using outcomes that are not extremely rare. If a disease is rare or the number of observations is small, then there may be insufficient data with which to investigate associations between explanatory variables and an outcome that differs from the disease that defined the original case series.

Simply ignoring the sampling scheme can lead to biased results, as illustrated via our empirical example. In contrast, adjustment for the disease that defined the original case series may be appropriate if the disease that defined the index series ($D_1$) is not associated with the outcome of interest ($D_2$), or if the disease that defined the index series is a true confounder of the association between the predictor variables and the disease of interest. For example, if $D_1$ is neither a confounder of, nor intermediate variable in, the association between exposure and $D_2$, then the estimate of association with $D_2$ obtained via stratum-weighted logistic regression will be identical to the estimate obtained conditioning on $D_1$. In such scenarios adjusting for $D_1$ in a standard logistic regression analysis will provide a valid (and possibly more precise) estimate of association. However, often the disease that defined the original case series is a factor that may not be desirable as a model covariate. In our example, the outcomes of interest ($D_1$ and $D_2$) are associated, and each is associated with the explanatory variable. Stratum-weighted regression offers an easily-implemented approach to estimate the association between predictor variables and $D_2$ without conditioning on $D_1$, and provides parameter estimates whose interpretation relates to the independent association between the exposure and outcome $D_2$.

Rather than use weighted logistic regression, a data analyst might consider restricting the analysis to some subset of the data (eg, controls only). If the controls had been selected using density sampling (such that the controls are a sample of the study base) then the association between the exposure and outcome $D_2$ restricted to the controls will provide a valid odds ratio estimate; however, restricting the analysis to the control series will result in a loss of information (most notably in a pair-matched case–control study). For a study of disease prevalence or cumulative incidence, restricting the analysis to the control series may undermine the scientific rationale for the regression analysis, similar to an analysis that adjusts for index case status.

A data analyst is not obliged to choose one of these analytical approaches. If an investigator suspects that $D_1$ is neither a confounder nor intermediate in the association between exposure and $D_2$, then this may be explored by comparing the magnitude and precision of results from stratum-weighted logistic regression and standard logistic regres-

sion with adjustment for $D_1$. The advantage of stratum-weighted logistic regression is that it offers a way to analyze case–control data for associations between predictor variables and health outcomes other than the one that defined the original index series, obtaining effect measures that are not adjusted for the disease that defined the index series. Although not a novel approach, stratum-weighted logistic regression is rarely applied by epidemiologists in this context. This approach allows a data analyst to make greater use of case–control data for investigations of etiologic associations and may be of particular value for analyses of data derived from large case–control studies.

## ACKNOWLEDGMENTS

*We are grateful to Lloyd Chambless for his contribution to this work and for his comments on earlier drafts of this manuscript.*

## REFERENCES

1. Weiland SK, Bjorksten B, Brunekreef B, et al. Phase II of the International Study of Asthma and Allergies in Childhood (ISAAC II): rationale and methods. *Eur Respir J*. 2004;24:406–412.
2. Fears TR, Brown CC. Logistic regression methods for retrospective case–control studies using complex sampling procedures. *Biometrics*. 1986;42:955–960.
3. Weinberg CR, Wacholder S. The design and analysis of case–control studies with biased sampling. *Biometrics*. 1990;46:963–975.
4. Prentice RL, Pyke R. Logistic disease incidence models and case–control studies. *Biometrika*. 1979;66:403–411.
5. Wacholder S, Weinberg CR. Flexible maximum likelihood methods for assessing joint effects in case–control studies with complex sampling. *Biometrics*. 1994;50:350–357.
6. Langholz B, Clayton D. Sampling strategies in nested case–control studies. *Environ Health Perspect*. 1994;102(suppl 8):47–51.
7. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. New York: John Wiley & Sons; 1989.
8. Chambless LE, Boyle KE. Likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. *Commun Statist Theor Meth*. 1985;14:1377–1392.
9. Pfeffermann D. The use of sampling weights for survey data analysis. *Stat Methods Med Res*. 1996;5:239–261.
10. Weiland SK, von Mutius E, Hirsch T, et al. Prevalence of respiratory and atopic disorders among children in the East and West of Germany five years after unification. *Eur Respir J*. 1999;14:862–870.
11. Greenland S. Introduction to regression modeling. In: Rothman K, Greenland S, eds. *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott, Williams, & Wilkins; 1998.