Published in final edited form as:
Epidemiology. 2007 July ; 18(4): 461–468. doi:10.1097/EDE.0b013e31806462d3.

An efficient sampling and inference procedure for studies with a continuous outcome

Haibo Zhou[1], **Jianwei Chen**[2], **Tiina H. Rissanen**[3], **Susan A. Korrick**[4], **Howard Hu**[4], **Jukka T Salonen**[3], and **Matthew P. Longnecker**[5]

[1]Department of Biostatistics, University of North Carolina, Chapel Hill, NC [2]Department of Biostatistics and Computational Biology University of Rochester Medical Center, Rochester, NY [3]Research Institute of Public Health, University of Kuopio, Kuopio, Finland. Department of Public Health and Clinical Nutrition, University of Finland, Finland [4]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massaschusetts [5]Epidemiology Branch, National Institute of Environmental Health Sciences, National Institute of Health, Department of Health and Human Services, RTP, NC

Abstract

To characterize the relation between an exposure and a continuous outcome, the sampling of subjects can be done much as it is in a case-control study, such that the sample is enriched with subjects who are especially informative. In an outcome dependent sampling (ODS) design, observations made on a judiciously chosen subset of the base population can provide nearly the same statistical efficiency as observing the entire base population. Reaping the benefits of such sampling, however, requires use of an analysis that accounts for the ODS design. In this report, the authors examined the statistical efficiency of a plain random sample analyzed with standard methods, compared with that of data collected with an ODS design and analyzed by either of two appropriate methods. In addition, three real datasets were analyzed using an ODS approach. The results demonstrate the improved statistical efficiency obtained by using an ODS approach and its applicability in a wide range of settings. An ODS design, coupled with an appropriate analysis, can offer a cost-efficient approach to studying the determinants of a continuous outcome.

Keywords

biased sampling; continuous outcome; empirical likelihood; epidemiologic methods; epidemiologic research design; semiparametric

A simple and well known outcome dependent sampling design is the case-control study. The case-control study design is logistically and economically appealing because observations made on a judiciously chosen subset of the population base provide nearly the statistical efficiency of observing the entire population base [1]. The principal idea of the case-control design and its subsequent extensions such as case-cohort and two-stage designs[e.g.2,3,4,5,6,7], is to concentrate resources on observations carrying the greatest amount of information. Related ideas of response based sampling have also be developed in economics and survey sampling[8,9,10].

An unique property of the case-control design is that under the logistic regression model with a binary Y, the estimated regression parameter (the odds ratio) for the exposure is the

Reprint requests to Dr. Haibo Zhou (zhou@bios.unc.edu, 919-966-3885, 919-966-3804 (fax)).

same under retrospective sampling as it is under a cohort study, with the effect of sampling confined to the intercept[2,4]. i.e., analyzing the a data set from case-control study as if it were from a cohort study will only affect the intercept. This property does not hold for a continuous response variable or if the regression model is not logistic. While designs for studying dichotomous outcomes have continued to develop, for studies of continuous outcomes analogous work has lagged. As the scope of epidemiologic inquiry grows, so does the need for efficient approaches to studying the determinants of a continuous outcome's level. This need is especially clear when the measurement of exposure is expensive.

As an example, the authors wanted to study background-level in utero exposure to the neurodevelopmental toxicant polychlorinated biphenyls (PCBs) in relation to performance on the Bayley Scale of Infant Development (BSID). Maternal pregnancy serum was available from a previously completed cohort study in which BSID had been measured, and PCB concentration in the maternal serum could provide a good surrogate measure of in utero exposure.

In studies of the relation between a relatively expensive exposure measure and the level of a continuous outcome, one approach has been to dichotomize the outcome and conduct a nested case-control study[11,12]. Dichotomizing a continuous outcome, however, could cause the estimand to be different and usually will result a loss of information as a lower order scale for the response is used [13].

To reap the benefits of a reduced sample size, one can employ case-control-like sampling adapted for continuous outcomes, outcome dependent sampling (ODS). Zhou et al.[14]. proposed an new ODS scheme that allows for two components. First, an overall random sample (SRS) of the population base is taken. Second, one or more supplementary random samples are taken, in which (as with case-control sampling) the probability of selection depends on the level of the outcome variable. For example, supplementary random samples might be taken only from the tails of the distribution of outcome. In most settings, by oversampling subjects in the tails, greater efficiency can be achieved than with a plain random sample of the same total sample size as the ODS sample, because the "tail" observations can provide greater influence on the parameter under study. Unlike the case-control study, failure to account for the biased sampling scheme with a continuous response in the statistical analysis will led to biased estimates for the regression parameters. In oder to consistently estimate the same estimand as the cohort study, the sampling scheme must be accounted for properly in the analysis. One commonly used method that can be adapted to our setting is to conduct a weighted analysis (weighted estimating equation approach), with weights inversely proportional to the probability of being sampled (IPW) [e.g.,15]. Another alternative is the weighted pseudo-likelihood method[10], which requires that one correctly specify all the underlying distributions. Misspecification of these distributions will lead to biased and erroneous conclusions. These methods, however, are all accounting the sampling scheme in an approximate way. A more efficient analysis that based on a likelihood function that truly reflect the biased sampling design[14] has been recently developed. The likelihood function used in (14) reflects all observed data and characteristics of the ODS sampling design. No additional distribution assumptions about the exposure variable are needed, nor is enumeration of the base population required (as is the case with the IPW method).

The present report builds on our previous, technically-oriented piece[14] in several ways. First, we provide a more intuitive explanation for why our ODS estimator is more statistically efficient than the alternatives. Our simulation study shows this efficiency under a wide range of exposure distributions, and translates the gain in efficiency into the reduction in sample sizes that yield equivalent statistical power. The simulation also explores the impact of various options in sampling and expresses results in terms of

statistical power. The real data examples demonstrate the wide applicability and special advantages of the proposed ODS design and estimator.

MATERIALS AND METHODS

A Semiparametric Inference Procedure for ODS with a Continuous Outcome

In this section, we give a brief overview of the ODS design Zhou *et al.*[14] proposed and their method for statistical inference under such a design. Let Y denote the continuous outcome variable and X denote the exposure. Assume that each Y falls into one of three mutually exclusive intervals: a lower tail strata, a middle section strata, and a upper tail strata. The general structure of the proposed design consists of two components: an overall random sample (SRS), and a supplement random sample from each of the three strata of Y. Let C_k, $k = 1, 2, 3$, denote the strata in Y. The observed data structure in the above ODS design is as follows: one observes the supplement random samples conditional on Y being in strata C_k, i.e., $\{Y_{ki}, X_{ki} | Y \in C_k\}$, where $i = 1, 2, \ldots, n_k$; One also observes an overall SRS sample whose individuals are denoted by $\{Y_{0i}, X_{0i}\}$ where $i = 1, 2, \ldots, n_0$. The total sample size in the ODS sample is therefore $n = n_0 + n_1 + n_2 + n_3$, where any of the n_k, $k = 0, 1, 2, 3$ can be zero. The above general sampling strategy encompasses several special cases, e.g., when $n_1 = n_2 = n_3 = 0$, then the ODS design reduces to the simple random sample design or cohort design; when Y is binary and $n_0 = 0$, the ODS design reduces to the usual case-control design.

Denote by $f_\beta(Y|X)$ the conditional density function for the population, where β is the vector of regression coefficients that links exposure X and the outcome Y. Let G and g denote the cumulative distribution and density functions of X, respectively. The joint likelihood function, $L(\beta)$, of the observed ODS data is

$$= \left[\prod_{i=1}^{n_0} f_\beta(Y_{0i}, X_{0i})\right] \left[\prod_{k=1}^{3}\prod_{i=1}^{n_k} f_\beta(Y_{ki}, X_{ki} | Y_{ki} \in C_k)\right] = \left[\prod_{i=1}^{n_0} f_\beta(Y_{0i}|X_{0i}) g(X_{0i})\right] \left[\prod_{k=1}^{3}\prod_{i=1}^{n_k} f_\beta(Y_{ki}, X_{ki} | Y_{ki} \in C_k)\right]. \quad (1)$$

The component in the first bracket is data contribution to the likelihood from the SRS sample, the second bracket is the contribution from each of the supplement samples. An important feature of this likelihood is easier to appreciate if it is re-expressed. From Bayes formula,

$$f_\beta(Y_{ki}, X_{ki} | Y_{ki} \in C_k) = I[Y_{ki} \in C_k] \frac{f_\beta(Y_{ki}|X_{ki}) g(X_{ki})}{Pr(Y_{ki} \in C_k)}, \quad (2)$$

where I is an indicator function for stratum membership, $Pr(Y_{ki} \in C_k)$ involves both g and β through $Pr(Y_{ki} \in C_k) = \int f_\beta(Y_{ki}|x) g(x) dx$. Plugging equation (2) into equation (1), the likelihood function we began with, denoted as $L(\beta, G)$ now to reflect the dependence on the unknown distribution of X, can be rewritten as

$$L(\beta, G) = \left[\prod_{k=0}^{3}\prod_{i=1}^{n_k} f_\beta(Y_{ki}|X_{ki})\right] \left[\prod_{k=0}^{3}\prod_{i=1}^{n_k} g(X_{ki})\right] \left[\prod_{k=1}^{K} P(Y_{ki} \in C_k)^{-n_i}\right]. \quad (3)$$

$L(\beta, G)$ now has three components: the specified regression model $f_\beta(Y|X)$ in the first bracket, the unspecified $g(X)$ in the second bracket, and the ODS sampling induced probability $P(Y_{ki} \in C_k)$ that ties $f_\beta(Y|X)$ and $g(X)$ together in the third bracket. The first component would be the usual likelihood function for observed data, had the sampling been

simple random sampling. The last component reflects the biased sampling nature of the ODS design and ignoring it in analysis would result in biased β estimates. Hence $g(X)$, or G, in the second bracket cannot be simply factored out as would be the case with a simple random sample design. Statistical inference about β using the standard maximum likelihood estimation method will depend on a known or a parameterized G. In practice, however, G is rarely known. Misspecification of the distribution could lead to an erroneous conclusion and bias the parameter estimation. Consequently, statistical approaches that do not rely on the extra parameterization of G are desirable.

To estimate β without specifying $G(X)$, Zhou et al. [14] developed a maximum likelihood based approach that maximizes $L(\beta, G)$ by modeling G nonparametrically. They used the profile likelihood idea where it (a) fixes β in equation (3) and solve for an empirical likelihood estimate $\hat{G}(\beta)$ from a constrained likelihood function, constraints placed on $\hat{G}$ that reflect its properties of being a discrete distribution function, using the Lagrange multiplier technique. An explicit solution for $\hat{G}(\beta)$ can be obtained. (b) Plugging $\hat{G}(\beta)$ into equation (3), the Zhou et al. estimator $\hat{\beta}_Z$ can be obtained, using the Newton-Raphson procedure, by maximizing the resulting likelihood. An explicit standard error formula based on an asymptotic distribution is given in Zhou *et al.*. The statistical program for this analysis can be obtained from the web page (www.bios.unc.edu zhou) or from the authors.

The inverse probability weighted approach of Horvitz and Thompson[15] can also be adapted in this situation by crudely treating all observed data, including the SRS sample, as if it were sampled from three strata, each with a given selection probability. Like the Zhou *et al.*. estimator, the IPW approach also yields a consistent estimate for β. The IPW method is commonly used with data from a two-stage study[e.g. 16,17] If all N individuals were fully observed in the entire population, the log likelihood function would be $\sum_{i=1}^{N} \log P(y_i|x_i;\beta)$. An estimate of this quantity is obtained if we use the completely observed individuals and weight their contributions inversely according to their selection probability into the second stage. Th IPW estimator $\hat{\beta}_{IPW}$ is the solution to following weighted score equation

$$\frac{1}{N}\sum_{k}\sum_{i\in C_k}\frac{1}{p_k}\frac{\frac{\partial}{\partial\beta}P_\beta(y_i|x_i)}{P_\beta(y_i|x_i)}=0.$$

where p_k can be estimated by $\frac{n_k}{N_k}$ if there is a complete information. Note that the IPW actually need more information than the likelihood approach employed by the Zhou et al. method since it requires the sampling probabilities to be known. However, since the IPW approach is based on crudely accounting for the sampling scheme and is based on estimating equations approach, it may not be as efficient as a likelihood based estimator. It has been shown that when the number of bins to group Y is not fine enough (when the number of categories of Y is small), the method is not efficient[18]. The realistic settings for the ODS design we considered had k between 2 to 4.

A SIMULATION STUDY

We designed a simulation to study the efficiency of different methods under a variety of conditions that mimic situations one might face in real applications. The basic simulation setting is modeled after a real study by Daniels *et al.*[19], of prenatal exposure to low levels of PCB in relation to mental and motor development, where an ODS design was used in the data collection. The data were generated according to the following model,

$$Y=\beta_0+\beta_1 X^p+\beta_2 Z_1+\beta_3 Z_2+\beta_4 Z_3+e,$$

where X is the exposure variable that takes on several distributions, $p = 1$ indicates a linear dose-response relationship and $p = 2$ represents a nonlinear relationship for $E[Y|X]$, the Zs are independent covariate variables, and e is a standard normal random error. We generate X from several distributions that includes normal, exponential, lognormal, and Bernoulli. These selections reflect possible real situations where X could be a rare-binary variable, a continuous variable, or a skewed variable. We generated Z_1 from a binary distribution (Bernoulli(0.45)), Z_2 from a log-normal distribution (LN(0, 0.25)), and Z_3 from a three level polynomial distribution (P(n, 0.3, 0.7)).

In our simulation, we first generated a cohort with 100, 000 individuals according to the above model and drew ODS samples from this cohort. We drew an overall random sample of size n_0, where we observed $\{Y,X,Z_1,Z_2,Z_3\}$. We also drew a supplement random sample of size n_1 from the lower tail of Y defined by $\{Y < \bar{Y} - a * \sigma_Y\}$ and a supplement random sample of size n_3 from the upper tail of Y defined by $\{Y > \bar{Y} + a * \sigma_Y\}$, where σ is the standard deviation of Y and a is a known constant. In addition to various configurations for the parameter values, we investigated the effect of varying the value of cut point a on the performance of the methods. We also investigated the impact on statistical power of varying the contribution to the total sample size from the overall random sample and the supplement random samples ($\rho = n_0/(n_0 + n_1 + n_3)$). Note $n_2 = 0$ here does not reduce the generality.

Under each setting, we compared the Zhou et al. estimator, denoted by $\hat{\beta}_Z$, with four other estimators: (i) the naive maximum likelihood estimator, $\hat{\beta}_N$, based on the observed ODS data but ignoring the sampling scheme; (ii) an inverse probability weighted estimator ($\hat{\beta}_{IPW}$); (iii) the maximum likelihood estimator based on a plain random sample of the same size as the ODS sample ($\hat{\beta}_P$); (iv) two logistic regression estimators ($\hat{\beta}_{Lk}$) based on dichotomizing a continuous Y by defining the outcome D as $D = 1$ if $Y > \text{mean}(Y) + k * \sigma_Y$ and $D = 0$ otherwise, where $k = 0, 1$. The weight used in calculating $\hat{\beta}_{IPW}$ is the inverse of the observed probability of being sampled in the respective strata of Y. The β_N and β_P estimates are the same as the ordinary least square estimates in our simulation setting. Each set of simulations generated 1000 data sets.

SIMULATION STUDY RESULTS

Table 1 shows the simulation results for $a = 1$, and $(n_0, n_1, n_3) = (200, 100, 100)$ (hence $\rho = 0.5$) for various exposure effects. The mean estimate given by $\hat{\beta}_N$ is biased for estimating β_1 in the simulation. Thus, ignoring the sampling scheme ($\hat{\beta}_N$) leads to a biased estimate for the exposure effect. It is also clear from Table 1 that different dichotomization of a continuous Y will lead to inconsistent, and different, estimates ($\hat{\beta}_{L0}$ and $\hat{\beta}_{L1}$) of the β_1. Perhaps more importantly, the logistic estimators can be less able to detect the true underlying relationship, as reflected by corresponding p values of 0.08 for $\hat{\beta}_{L1}$, compared with $p < 0.05$ for all other methods using continuous response. We do not present results for these two estimators in future comparisons. The other three methods all yielded consistent estimates of β_1. The actual coverage of their nominal 95 percent confidence intervals coverage (95% CI Coverage) are all close to 95 percent, indicating that a good approximation to the asymptotic normality is achieved with this sample size, and the estimated standard errors (SE) are close to the true standard deviations. Under the setting considered, $\hat{\beta}_Z$ has the smallest SE while b $\hat{\beta}_P$ has the largest SE. Because $\hat{\beta}_N$ is a biased estimator and its SE underestimates the true variation, we excluded it from the further studies of sample size and power below. The

above observations are consistent across different exposure effects listed in Table 1. Under the same linear model but when the X term is quadratic, $\hat{\beta}_Z$ is again more efficient.

Results in the lower panels of Table 1 provide the contrast for the normally distributed X to extreme X, namely a skewed exposure (Lognormal) and a rare binary exposure (Bernoulli(.05)). When compared to the Normal distributed exposure, the SE for β_Z is even smaller SE in the skewed exposure situations. For the rare binary exposure case, results in Table 1 demonstrate that $\hat{\beta}_Z$ is still the most efficient overall, though the sample size considered was not sufficiently large enough for any of them. This is reflected in the fact that the estimated standard errors are bigger than the estimates of the slope. This is not surprising because with the distribution of a rare binary X, there may not be enough information in the data set as $X = 1$ could be sparse. Future development of a modified ODS design for this situation is certainly warranted.

Figure 1 shows the power for testing $H_0 : \beta 1 = 0$ v.s. $H_1 : \beta_1 = \textit{true value}$, for $n = 400$ and type I error fixed at 5 percent. The points corresponding to the true value of $\beta_1 = 0$ shows the empirical type I error rate for each test. All three methods yield close to 5 percent type I error. For all three estimators, as β_1 increases, so does the statistical power, and the power of $\hat{\beta}_Z > \hat{\beta}_{IPW} > \hat{\beta}_P$.

Table 2 shows the sample sizes required to achieve a given statistical power for three values of β_1 according to type of estimator under the same settings used in the top panel of Table 1. Use of ODS with an appropriate estimator requires a smaller sample size. In this setting, the $\hat{\beta}_Z$ method on average needs about 60 percent of the subjects who would be needed if the study were conducted with a simple random sampling scheme and the $\hat{\beta}_{IPW}$ method needs about 83 percent. Further, for a given power, as the true value of β_1 is farther away from 0, relatively fewer subjects are needed to achieve the same power with $\hat{\beta}_Z$ as compared with $\hat{\beta}_P$. i.e., efficiency increased as β_1 is farther away from 0.

Figure 2 shows the impact of two factors in a given ODS setting. The impact of varying a, the *cut-point* that determined the strata of Y where the supplement random samples were drawn, is shown in the left graph. $a = 0$ means there was only plain random sampling and in this instance all three methods had the same power. As a increases, however, $\hat{\beta}_Z$ (solid line) has better power than the other two. With $\hat{\beta}_Z$, the increase in power appears to be monotonic in a. The graph on the right of Figure 2 shows the impact of ρ, the fraction of the overall random sample in the total ODS sample ($\rho = n_0/(n_0+n_1+n_3)$). At $\rho = 1$, there is no supplement sample and the ODS samples are plain random samples. However, as ρ decreases to below 0.7, $\hat{\beta}_Z$ (solid curve) is again the most powerful of the three estimators.

Table 3 investigates different ODS allocations when the exposure variable is skewed (X ~ lognormal). Compared with the results in the lognormal panel of Table 1, we see that allocating more of the ODS sample to the upper tail of Y improves the efficiency. The standard error of the slope estimator decreases as more of the ODS sample is shifted from the lower tail to the upper tail, i.e. 0.0222 → 0.0201 → 0.0192 as the ODS allocation changes from (200, 150, 50) → (200, 100, 100) → (200, 50, 150).

Real Data Example 1

In the motivating example noted in the Introduction, the setting was like one where if the outcome has been dichotomous a nested case-control study would have been implemented. However, the outcome of interest, the score on the Bayley Scale of Infant Development (BSID), was a continuous variable, and, as noted above, treating it as a dichotomous variable would have resulted in a loss of statistical power. Measurements of the exposure of interest here, polychlorinated biphenyls, are expensive, thus minimizing sample size while

maintaining power was especially important. Thus, the authors drew drew a random sample of cohort members and two additional random samples, one from each tail of the outcome distribution[19]. Using the inverse probability weighted estimator, the estimated $\hat{\beta}_{IPW}$ was 0.47 BSID units/μg/L PCB with estimated SE as 0.32 ($p = 0.14$)[14]. Using the Zhou et al. estimator, the estimated $\hat{\beta}_Z$ was 0.44 with $SE = 0.22$ ($p = 0.02$). Using the SRS data alone, the $\hat{\beta}_N$ was 0.29 ($SE = 0.29$, $p = 0.32$). Daniels et al. also examined and confirmed the shape of the dose-response relation in both the ODS and the SRS samples. The example demonstrates the improved efficiency obtained by the Zhou et al. estimator.

Real Data Example 2

Rissanen et al.[20] examined the relation of serum lycopene concentration to the thickness of carotid arteries among 1028 men in Finland. Using their data, we selected samples of total n=400 in two ways. First, we selected a random sample with n=400. Second, we selected a random sample of n=200, and, among those with carotid artery thickness above the 90th percentile we selected a random sample of n=100, and among those with carotid artery thickness below the 10th percentile we selected a random sample of n=100. We then analyzed the data using ordinary least squares with n=1028, and with the 400 all selected at random. In addition, we analyzed the 200-100-100 sample using inverse variance weights and then with the Zhou et al. estimator. In all models the lycopene results were adjusted for the same covariates (age, year, and sonographer). Results are given in Table 4. With n=1028 (the full data), the estimated coefficient for lycopene was −0.14 with an estimated SE at 0.04 and $p = 0.0011$. With n=400 (all random), the estimated effect $\hat{\beta}_P = -0.10$, with $SE = 0.06$ and $p = 0.096$. With the 200-100-100 sample, we had $\hat{\beta}_{IPW} = -0.19$ with $SE = 0.08$, $p = 0.017$ and $\hat{\beta}_Z = -0.24$, $SE = 0.07$ and $p = 0.0009$. This example suggests that use of an outcome-dependent sampling scheme and the Zhou et al. estimator obtained nearly as much power as analysis of the full dataset. Furthermore, the greater efficiency of the Zhou et al. estimator compared with IPW approach is clear. The larger β's obtained using outcome dependent sampling and estimators reflect the shape of the dose-response curve, which had larger negative slopes near the tails of the outcome. While we focused on analyzing the lycopene association as linear (trend test type approach), clearly a curvilinear approach would better describe the relation and could be easily accommodated with either the IPW or the Zhou et al. estimators.

Real Data Example 3

Korrick et al.[21] conducted a case-control study of hypertension and bone lead level.x For logistic reasons the sampling probabilities for cases and controls could not be determined. Measurements of blood pressure were available. The example, which is presented in the online supplementary material, shows that with their data, the proposed ODS approach could be used to estimate the coefficient for bone lead in a model of blood pressure.

DISCUSSION

Our results show that for a fixed total number of observations used to examine the linear relation between a continuous exposure and a continuous outcome, the ODS design is more efficient than a plain random sample design. An intuitive and simple expression of the benefit of the ODS approach using the Zhou *et al.* estimator was found in the simulations showing that it yields the same estimand as that from the underlying cohort study and that for a given desired statistical power, the number of observations needed could be reduced by about 40 percent compared with a plain random sample. This work shows that the benefits of outcome dependent sampling apply to continuous outcomes, not just to dichotomous ones through case-control designs. For binary response variable, our approach is equivalent to the case-control analysis[22].

To implement the weighted method, one needs to know or estimate the weights which requires at least empirical data about the distribution of Y. This may not be a problem for nested studies; however, it can be difficult to calculate the weight for studies that are not nested. The difficultly arises because good quality data on the distribution of Y, and an enumeration of potential subjects, may not be available for the base population. The IPW suffers from the fact that in the typical ODS setting the natural choice of number of categories of Y is not large enough to yield the variability in weights that would make it efficient[18]. Some recently developed methods may help to identify even more optimal weights[18,23]. The Zhou et al. method, on the other hand, does not used the selection probability and hence does not need to enumerate the base population.

Our results showed that with a up to about 1 and ρ near 0.5 and a total sample size of 400, ODS analysis conducted using the estimator of Zhou *et al.* increased power about twice as much as the weighted estimating equation approach (Figure 2). With larger values of a or smaller values of ρ, the advantage of the Zhou et al. estimator was greater. This reflects a greater influence on the regression parameters for data from the tail areas than the middle areas. In general, the efficiency gain of the IPW method over the simple random sample analysis is notable, thus we would suggest in practice that one should at least do the weighted analysis if one cannot implement the Zhou et al. method.

If an ODS procedure is to be used, questions will arise regarding the optimal choice of a and ρ. For example, consider an examination of the serum level of contaminant X among pregnant women in relation to a continuous outcome in offspring. Subject matter considerations might support large values of a, e.g., greater than 1, so that it corresponds to a clinically abnormal value. Values of a greater than 1 might seem appealing because of the resulting increase in power, especially with the Zhou et al. estimator (Figure 2). But the reward from choosing a relatively large value of a depends on an assumption about $f\beta(Y|X)$ across the range of Y. If, e.g., Y is intelligence quotient (IQ), and a is set at 2.5, and the only supplement sample is of those with IQ's less than 100 - 2.5 * SE_{IQ}, or 62.5, then the mentally retarded population in the supplement sample will include a larger proportion of genetically-determined retardation, on whom X might have essentially no effect. This "dilution" of the supplement sample with the "doomed" (or "immune") means estimates of $f\beta(Y|X)$ will be attenuated in proportion to the dilution. Smaller values of a, e.g., 1–1.5 would guard against such dilution while allowing one to get much more power with the Zhou et al. estimator than would be possible with the weighted estimated equation. Similarly, choice of $\rho > 0$ has several advantages over $\rho = 0$ (14), since including an overall random sample provides the flexibility of a cohort study and allows for model checking. In general, choosing a ρ from the range of [.2, 0.5] allows one to get much improved power with the Zhou et al. estimator than would be possible with the weighted estimated equation estimator, while still allocating enough observations to the SRS sample.

Similarly, the relative size of n_1 and n_3 might be affected by several factors that will vary across studies. For example, if the exposure variable is known to be skewed with a long tail to the right and β is known to be either zero or positive, then increasing the size of n_3 relative to n_1 would be sensible. A large n_3 relative to n_1 would also make sense when there is little interest in the determinants of a low value of the outcome variable.

Prospective designs coupled with relatively expensive measures of exposure are being used with increasing frequency in epidemiologic research. Furthermore, the scope of epidemiologic research increasingly includes outcomes best measured on a continuous scale. Given these trends, methods that allow cost cutting while maintain statistical efficiency are likely to see greater use. Recently, similar ideas using the ODS design has been extended to the situation where in addition to the ODS sample, information other than exposure variable

are also available for the rest of the base population[24]. Methods have been developed that account for an ordinal outcome variables in a generalized linear model setting, as have been methods that incorporate auxiliary information about the exposure variable that is available for the entire base population[25]. Much work, however, remains to be done, e.g., how to use ODS with longitudinal data is still an open question. A survey of the statistical research on ODS design can be found in Zhou and You[26].

Acknowledgments

The authors thank Dr. Clare Weinberg and Beth Gladen for their careful reading of the paper and helpful suggestions. We also thank the reviewers for their helpful suggestions that lead to a much more complete version of the manuscript. This research was supported in part by a grant from NIH (CA 79949, for H. Zhou and J. Chen) and by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (for M. Longnecker).

Abbreviations

BSID	Bayley Scale of Infant Development
CI	Confidence interval
IQ	intelligence quotient
ODS	outcome dependent sampling
SE	standard error

REFERENCES

1. Cornfield J. A method of estimating comparative rates from clinical data. Applications to cancer of lung, breast, and cervix. Journal of National Cancer Institute. 1951; 11:1269–1275.
2. Anderson JA. Separate sample logistic discrimination. Biometrika. 1972; 59:19–35.
3. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. Biometrika. 1988; 75:11–20.
4. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika. 1979; 71:101–113.
5. Prentice RL. A case-cohort design for epidemiologic studies and disease prevention trials. Biometrika. 1986; 73:1–11.
6. White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. American Journal of Epidemiology. 1982; 115:119–128. [PubMed: 7055123]
7. Wacholder S, Weinberg CR. Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling. Biometrics. 1994; 50(2):350–357. [PubMed: 8068835]
8. Imbens GW, Lancaster T. Efficient estimation and stratified sampling. Journal of Econometrics. 1996; 74:289–318.
9. Cosslett SR. Maximum likelihood estimator for choice-based samples. Econometrika. 1981; 49:1289–1316.
10. Holt D, Smith TMF, Winter PD. Regression analysis of data from complex surveys. Journal of the Royal Statistical Society, A. 1980; 143:474–487.
11. Li R, Folsom AR, Sharrett AR, Couper D, Bray M, Tyroler HA. Interaction of the glutathione S-transferase genes and cigarette smoking on risk of lower extremity arterial disease: the Atherosclerosis Risk in Communities (ARIC) study. Atherosclerosis. 2001 Feb 15; 154(3):729–738. [PubMed: 11257276]
12. Iribarren C, Folsom AR, Jacobs DR Jr, Gross MD, Belcher JD, Eckfeldt JH. Association of serum vitamin levels, LDL susceptibility to oxidation, and autoantibodies against MDA-LDL with carotid atherosclerosis. A case-control study. The ARIC Study Investigators. Atherosclerosis Risk in Communities. Arterioscler Thromb Vasc Biol. 1997 Jun; 17(6):1171–1177. [PubMed: 9194770]

13. Suissa S. Binary methods for continuous outcome: a parametric alternative. Journal Clinical Epidemiology. 1991; 44:241–248.

14. Zhou H, Weaver MA, Qin J, Longnecker MP, Wang MC. A semiparametric empirical likelihood method for data from an outcome-dependent sampling design with a continuous outcome. Biometrics. 2002; 58:413–421. [PubMed: 12071415]

15. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association. 1952; 47:663–685.

16. Flanders WD, Greenland S. Analytical methods for two-stage case-control studies and other stratified designs. Statistics in Medicine. 1991; 10:739–747. [PubMed: 2068427]

17. Zhao LP, Lipsitz S. Designs and Analysis of Two-Stage Studies. Statistics in Medicine. 1992; 11:769–782. [PubMed: 1594816]

18. Godambe VP, Vijyan K. Optimal estimation for response-dependent retrospective sampling. Journal of the American Statistical Association. 1996; 91:1724–1734.

19. Daniels JL, Longnecker MP, Klebanoff MA, Gray KA, Brock JW, Zhou H, Chen Z, Needham LL. Prenatal exposure to low-level polychlorinated biphenyls in relation to mental and motor development at 8 months. Am J Epidemiol. 2003; 157:485–492. [PubMed: 12631537]

20. Rissanen T, Voutilainen S, Nyyssonen K, Salonen R, Salonen JT. Low plasma lycopene concentration is associated with increased intima-media thickness of the carotid artery wall. Arterioscler Thromb Vasc Biol. 2000 Dec; 20(12):2677–2681. [PubMed: 11116071]

21. Korrick SA, Hunter DJ, Rotnitzky A, Hu H, Speizer FE. Lead and hypertension in a sample of middle-aged women. J Public Health. 1999 Mar; 89(3):330–335.

22. Qin J, Zhang B. A goodness-of-fit test for logistic regression models based on case-control data. Biometrika. 1997; 84(3):609–618.

23. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association. 1994; 89:846–866.

24. Weaver M, Zhou H. An Estimated Likelihood Method for Continuous Outcome Regression Models with Outcome-Dependent Subsampling. J. Am. Stat. Assoc. 2005; 100:459–469.

25. Wang X, Zhou H. A Semiparametric Empirical Likelihood Method For Biased Sampling Schemes In Epidemiologic Studies With Auxiliary Covariates. Biometrics. 2006 (in press).

26. Zhou, H.; You, J. Semiparametric Methods for Data from an Outcome-Dependent Sampling Scheme. In: Hong, D.; Shyr, Y., editors. Quantitative Medical Data Analysis Using Mathematical Tools and Statistical Techniques. Singapore: World Scientific Publications; 2006. (in press)

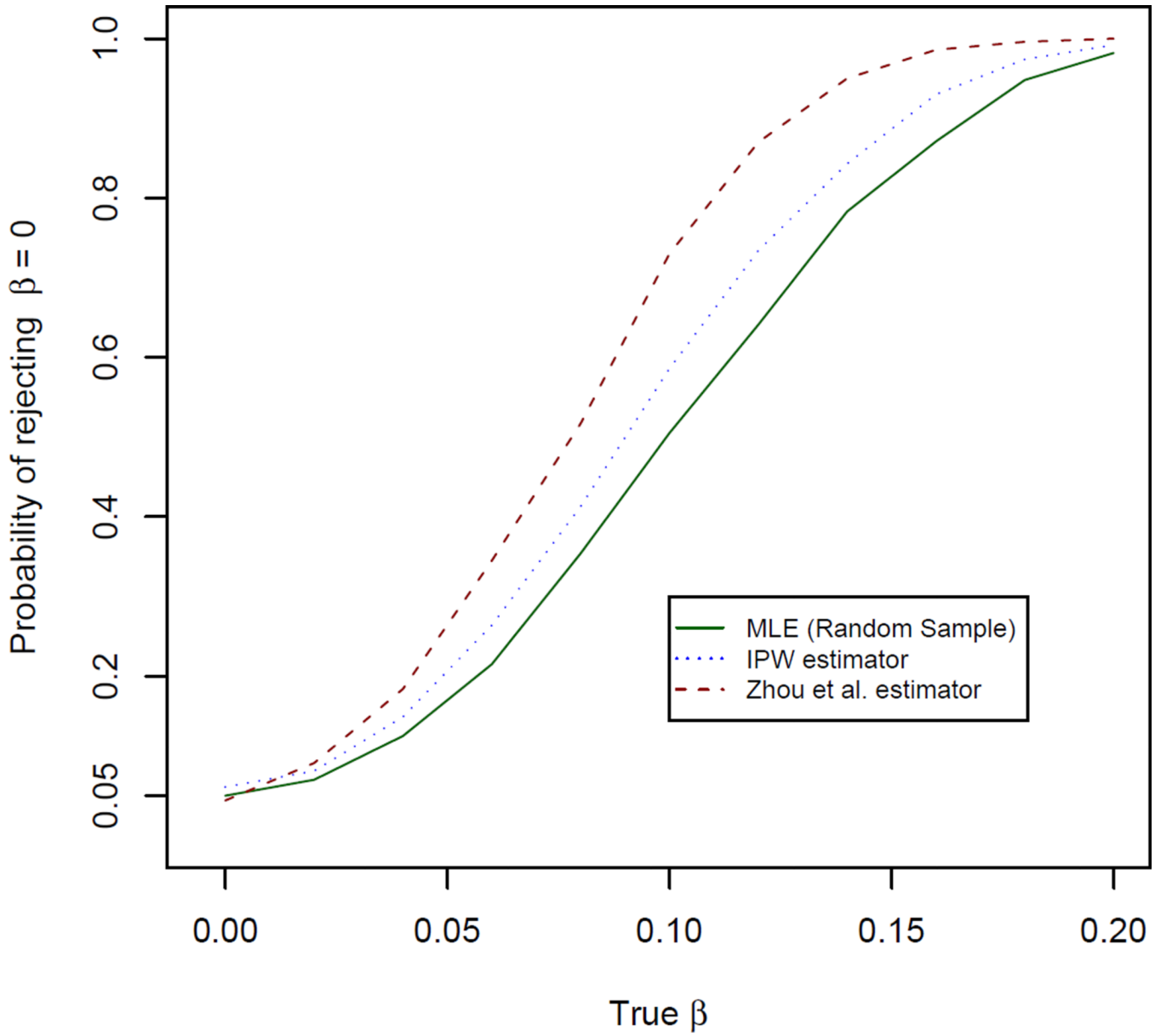

Figure 1.
Simulation results for the power of testing $H_0 : \beta_1 = 0$. v.s. $H_1 : \beta_1$ = true value under the model in the top panel of Table 1. The results are based on 1000 simulations with $n_0 = 200$ and $n_1 = n_3 = 100$.

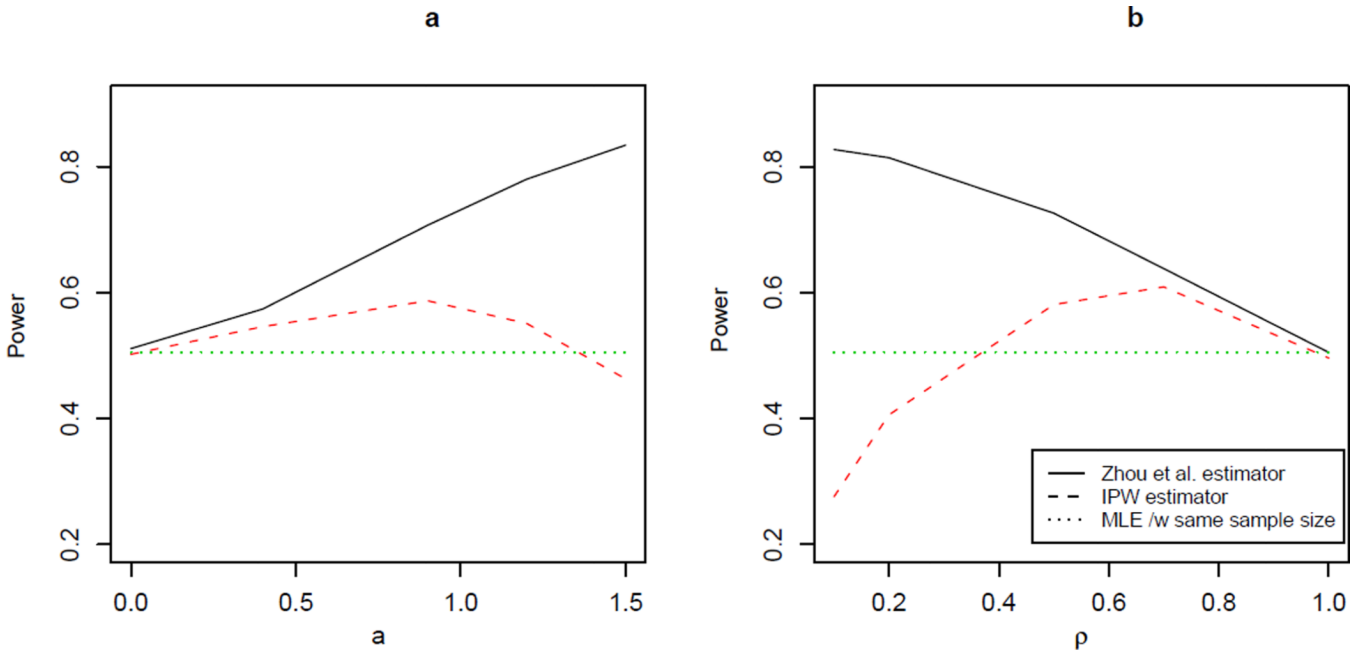

Figure 2.

The power for the testing H_0: $\beta_1 = 0$. v.s. $H_1 : \beta_1 = 0.1$. (a). The power as a, cut-point for defining the strata in term of σ_Y, varies with sample size $(n_0, n_1, n_2) = (200, 100, 100)$; (b). The power as ρ, the fraction of SRS sample, varies with $n = 400$, $n_1 = n_2$, and $a = 1$.

Table 1

Simulation results for different exposure effects with $\alpha = 1$ and $\rho = 0.5$

Term of X in the model	(n_0, n_1, n_3)	β_1	Methods	Mean	SE	$\widehat{SE}$	95% C.I. Coverage
Linear X	(200,100, 100)	0.1	β_N	0.167	0.066	0.051	0.671
$X \sim N(0, 1)$			β_P	0.099	0.051	0.051	0.950
			β_{IPW}	0.101	0.048	0.047	0.938
			β_Z	0.101	0.040	0.040	0.947
			β_{L_0}	0.237	0.113	0.107	0.756
			β_{L_1}	0.221	0.131	0.128	0.852
Quadratic X^2	(200,100,100)	0.1	β_N	0.156	0.042	0.034	0.587
$X \sim N(0,1)$			β_P	0.100	0.035	0.036	0.961
			β_{IPW}	0.101	0.034	0.041	0.975
			β_Z	0.100	0.029	0.029	0.950
Linear	(200,100,100)	0.1	β_N	0.137	0.026	0.021	0.554
$X \sim LN(0,1)$			β_P	0.100	0.025	0.025	0.953
			β_{IPW}	0.102	0.023	0.028	0.977
			β_Z	0.101	0.020	0.020	0.947
Linear	(200,100,100)	0.1	β_N	0.166	0.299	0.234	0.863
$X \sim Bernoulli(0.05)$			β_P	0.110	0.237	0.234	0.948
			β_{IPW}	0.103	0.224	0.220	0.930
			β_Z	0.102	0.183	0.184	0.961

NOTE: Results are based on the model $Y = \beta_0+\beta_1 X+\beta_2 Z_2+\beta_3 Z_3+\beta_4 Z_4+e$, where $e \sim N(0, 1)$,$\beta_0 = 1.5$, $\beta_2 = -0.5$, $\beta_3 = 0.02$ and $\beta_4 = 0.05$. β_N is the estimator based on ignoring the ODS sampling scheme. β_P is the maximum likelihood estimator based on a plain random sample of the same sample size. β_{IPW} is the inverse probability weighted estimator and β_Z is the Zhou *et al.* estimator. β_{L_1} is estimator from a logistic regression analysis where outcome variable to be one if $Y \geq \text{mean}(Y) + \sigma_Y$ and 0 otherwise. β_{L_2} is estimator from a logistic regression analysis where outcome variable to be one if $Y \geq \text{mean}(Y)$ and 0 otherwise.

Table 2

Sample size needed for testing $H_0 : \beta_1 = 0$ for a given statistical power. The results are based on 1000 simulations with $\alpha = 1.0$, $\rho = 0.5$ and $n_1 = n_2$.

		Sample Sizes for		
Power	**Trueβ_1**	**β_P**	**β_{IPW}**	**β_Z**
0.80	0.05	3000	2500	1900
	0.10	790	670	470
	0.15	360	310	220
0.85	0.05	3600	2900	2250
	0.10	960	780	530
	0.15	400	340	245
0.90	0.05	4200	3400	2500
	0.10	1070	870	630
	0.15	485	400	280
0.95	0.05	5100	4300	3080
	0.10	1320	1080	770
	0.15	625	510	350

NOTE: X follows a log-normal distribution, other details see footnote to Table 1.

Table 3

Simulation study for different ODS allocation with skewed exposure effect.

(n_0,n_1,n_3)	β_1	Methods	Mean	SE	$\widehat{\text{SE}}$	95% C.I.
(200; 50; 150)	0.1	β_N	0.116	0.021	0.019	0.850
		β_P	0.100	0.025	0.025	0.953
		β_{IPW}	0.102	0.023	0.028	0.986
		β_Z	0.100	0.019	0.019	0.947
(200,150,50)		β_N	0.147	0.031	0.025	0.503
		β_P	0.100	0.025	0.025	0.953
		β_{IPW}	0.102	0.025	0.031	0.977
		β_Z	0.101	0.022	0.022	0.943

Table 4

Results of fitting adjusted models of the carotid artery thickness in relation toserum lycopene concentration, according to sampling method and method of data analysis.

Sampling method	n	Analysis method	$\hat{\beta}$[1]	SE(β)	p
All available subjects	1028	OLS	−0.14	0.04	0.0011
Random sample	400	OLS	−0.10	0.06	0.096
Random sample plus outcome tail samples[2]	400	weighted OLS	−0.19	0.08	0.017
Random sample plus outcome tail samples[2]	400	Zhou et al. estimator	−0.24	0.07	0.0009

NOTE:

[1] Units are mm per mol/L, adjusted for age, year, and sonographer.

[2] 200 subjects selected at random, 100 with carotid artery thickness above the 90th percentile selected a random, and 100 with carotid artery thickness below the 10th percentile selected a random (see text).