

Outcome-Dependent Sampling

An Efficient Sampling and Inference Procedure for Studies With a Continuous Outcome

Haibo Zhou,* Jianwei Chen,† Tiina H. Rissanen,‡ Susan A. Korrick,§ Howard Hu,§
Jukka T. Salonen,‡ and Matthew P. Longnecker¶


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 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 outcome-dependent sampling. In this report, we examine the statistical efficiency of a plain random sample analyzed with standard methods, compared with that of data collected with outcome-dependent sampling and analyzed by either of 2 appropriate methods. In addition, 3 real datasets were analyzed using an outcome-dependent sampling approach. The results demonstrate the improved statistical efficiency obtained by using an outcome-dependent sampling, and its applicability in a wide range of settings. This design, coupled with an appropriate analysis, offers a cost-efficient approach to studying the determinants of a continuous outcome.

(*Epidemiology* 2007;18: 461–468)

Submitted 4 April 2006; accepted 27 March 2007.

From the *Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina; †Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York; ‡Research Institute of Public Health, University of Kuopio, Kuopio, Finland, and Department of Public Health and Clinical Nutrition, University of Finland, Finland; §Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; ¶Epidemiology Branch, National Institute of Environmental Health Sciences, National Institute of Health, Department of Health and Human Services, RTP, North Carolina.

Supported in part by NIH grant R01 CA79949 (H.Z., J.C.) and by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (M.L.).

 Supplemental material for this article is available with the online version of the journal at www.epidem.com; click on "Article Plus."

Correspondence: Dr. Haibo Zhou, Department of Biostatistics, CB 7420, The University of North Carolina, Chapel Hill, NC 27599-7420. E-mail: zhou@bios.unc.edu.

Copyright © 2007 by Lippincott Williams & Wilkins
ISSN: 1044-3983/07/1804-0461
DOI: 10.1097/EDE.0b013e31806462d3

The case-control study is a simple and familiar example of an outcome-dependent sampling design. The case-control 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,^{2–7} is to concentrate resources on observations carrying the greatest amount of information. Related ideas of response-based sampling have also been developed in economics and survey sampling.^{8–10}

A 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 same under retrospective sampling as it is under a cohort study, with the effect of sampling confined to the intercept.^{2,4} Thus, analyzing a data set from a case-control study as if it were from a cohort study will affect only the intercept. This property does not hold for a continuous response variable, or for regression models that are not logistic.

While designs for studying dichotomous outcomes have continued to develop, analogous work for studies of continuous outcomes has lagged. As the scope of epidemiologic inquiry grows, so does the need for efficient approaches to studying the determinants of the level of a continuous outcome, especially when the measurement of exposure is expensive.

As an example, we wanted to study in utero exposure to background-levels of the neurodevelopmental toxicant polychlorinated biphenyls (PCBs), and its relation to performance on the Bayley Scale of Infant Development. Maternal serum collected during pregnancy was available from a previously completed cohort study in which Bayley Scale had been measured. PCB concentration in the maternal serum provides 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, will distort the estimand and usually will result in a loss of information because it uses a lower-order scale for the response.¹³

To reap the benefits of a reduced sample size, one can employ outcome-dependent sampling, which is a case-

control-like sampling adapted for continuous outcomes. Zhou et al¹⁴ proposed a sampling scheme with 2 components. First, an overall random sample of the population base is taken. Second, one or more supplementary random samples are taken, in which the probability of selection depends on the level of the outcome variable (as with case-control sampling). For example, supplementary random samples might be taken only from the tails of the outcome distribution. In most settings, oversampling subjects in the tails achieves greater efficiency than a completely random sample of the same sample size, because the “tail” observations can provide greater influence on the parameter under study. Unlike the case-control study, the analysis must account for the biased sampling scheme with a continuous response to avoid biased estimates by the regression parameters (ie, to consistently estimate the same estimand as would have been obtained by studying the entire population). One commonly used method that can be adapted for this purpose is to conduct a weighted analysis (weighted estimating equation approach), with weights inversely proportional to the probability of being sampled (IPW).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, account for the sampling scheme in only an approximate way. A more efficient analysis, based on a likelihood function that precisely reflects the biased sampling design¹⁴ has been recently developed. The likelihood function used in our earlier paper¹⁴ reflects all observed data and characteristics of the outcome-dependent sampling design. No additional distributional assumptions about the exposure variable are needed, nor is enumeration of the base population required (as is the case with the inverse weighting method).

The present report builds on our previous, technically oriented piece¹⁴ in several ways. First, we provide a more intuitive explanation for why estimator based on outcome-dependent sampling is statistically more 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 outcome-dependent sampling design and estimator.

METHODS

A Semiparametric Inference Procedure for Outcome-Dependent Sampling With a Continuous Outcome

In this section, we give a brief overview of the outcome-dependent sampling design,¹⁴ and statistical inference under such a design. Let Y denote the continuous outcome variable and X the exposure. Assume that each Y falls into 1 of 3 mutually exclusive intervals: a lower tail strata, a middle section strata, and an upper tail strata. The general structure

of the proposed design consists of 2 components: an overall random sample, and a supplement random sample from each of the 3 strata of Y . Let C_k , $k = 1, 2, 3$, denote the strata in Y . The data structure in the above design is as follows: one observes the supplement random samples conditional on Y being in strata C_k , ie, $\{Y_{ki}, X_{ki} | Y \in C_k\}$, where $i = 1, 2, \dots, n_k$; one also observes an overall simple random sample whose individuals are denoted by $\{Y_{0i}, X_{0i}\}$, where $i = 1, 2, \dots, n_0$. The total sample size in the outcome-dependent sampling 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; for example, when $n_1 = n_2 = n_3 = 0$, then the outcome-dependent sampling design reduces to the simple random sample or cohort design; alternatively, when Y is binary and $n_0 = 0$, the outcome-dependent sampling 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 outcome-dependent sampling 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 simple random 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

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

$L(\beta, G)$ now has 3 components: the specified regression model $f_{\beta}(Y | X)$ in the first bracket, the unspecified $g(X)$ in the second bracket, and the outcome-dependent 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 outcome-dependent sampling design; ignoring it in the analysis would result in biased estimates of β . 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 al¹⁴ developed a maximum likelihood based approach that maximizes $L(\beta, G)$ by modeling G nonparametrically. This approach used the profile likelihood idea where it fixes β in equation (3) and solves 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. 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 has been provided.¹⁴ 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 Thompson¹⁵ can also be adapted in this situation by crudely treating all observed data, including the simple random sample, as if it were sampled from 3 strata, each with a given selection probability. Like the Zhou et al estimator, the inverse probability weighted approach also yields a consistent estimate for β . This method is commonly used with data from a 2-stage study.^{16,17} If all N individuals were fully observed in the 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. The inverse probability weighted estimator $\hat{\beta}_{IPW}$ is the solution to the following weighted score equation

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

where p_k can be estimated by $\frac{n_k}{N_k}$ if there is a complete information. Note that this method needs more information than the likelihood approach employed by Zhou et al method since it requires the sampling probabilities to be known. However, since the inverse-probability weighted approach is based on crudely accounting for the sampling scheme and on an estimating-equations approach, it may not be as efficient as a likelihood-based estimator. When the number of categories of Y is small, the estimating-equation method is not efficient.¹⁸ The realistic settings for the outcome-dependent sampling design we considered had k between 2 to 4. Generally speaking, if the same amount of information is used, among the various statistical approaches to computing any estimator, the maximum likelihood approach is always the most efficient.

A SIMULATION STUDY

We designed a simulation to study the efficiency of different methods under a variety of conditions that one might face in real applications. The basic simulation setting is modeled after a study by Daniels et al¹⁹ of prenatal exposure to low levels of PCB in relation to mental and motor development. An outcome-dependent sampling 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 Z s are independent covariate variables, and e is a standard normal random error. We generate X from several distributions that include 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 3 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 outcome-dependent 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 supplemental random sample of size n_1 from the lower tail of Y defined by $\{Y < \bar{Y} - a * \sigma_Y\}$ and a supplemental 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 supplemental random samples ($p = n_0/(n_0 + n_1 + n_3)$). 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 4 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 ($\hat{\beta}_P$); and (iv) 2 logistic regression estimators ($\hat{\beta}_{L_k}$) 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 various inconsistent estimates ($\hat{\beta}_{L_0}$ and $\hat{\beta}_{L_1}$) 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 $\hat{\beta}_{L_1}$, compared with $P < 0.05$ for all other methods using continuous response. We do not present results for these 2 estimators in future comparisons. The other 3 methods all yielded consistent estimates of β_1 . The actual

coverage of their nominal 95% confidence intervals (CIs) coverage are all close to 95%, 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 $\hat{\beta}_P$ 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 portion 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 (0.05)). When compared with 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 outcome-dependent sampling design for this situation is certainly warranted.

Figure 1 shows the power for testing $H_0: \beta_1 = 0$ versus $H_1: \beta_1 = \text{true value}$, for $n = 400$ and type I error fixed at 5%.

TABLE 1. Simulation Results for Different Exposure Effects With $a = 1$ and $\rho = 0.5$

Term of X in the Model	(n_0, n_1, n_3)	β_1	Methods	Mean	SE	\widehat{SE}	95% CI Coverage
Linear X $X \sim N(0, 1)$	(200, 100, 100)	0.1	β_N	0.167	0.066	0.051	0.671
			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 $X \sim N(0, 1)$	(200, 100, 100)	0.1	β_N	0.156	0.042	0.034	0.587
			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
			β_N	0.137	0.026	0.021	0.554
			β_P	0.100	0.025	0.025	0.953
Linear $X \sim LN(0, 1)$	(200, 100, 100)	0.1	β_{IPW}	0.102	0.023	0.028	0.977
			β_Z	0.101	0.020	0.020	0.947
			β_N	0.166	0.299	0.234	0.863
			β_P	0.110	0.237	0.234	0.948
Linear $X \sim \text{Bernoulli}(0.05)$	(200, 100, 100)	0.1	β_{IPW}	0.103	0.224	0.220	0.930
			β_Z	0.102	0.183	0.184	0.961

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.

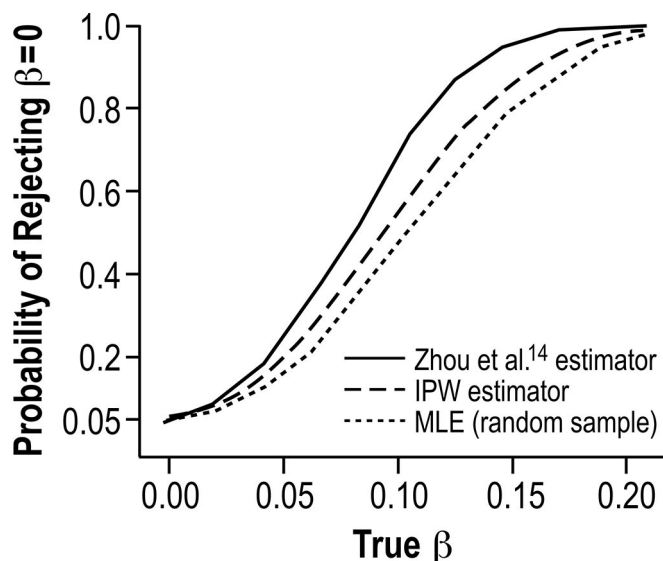


FIGURE 1. Simulation results for the power of testing $H_0: \beta_1 = 0$ versus $H_1: \beta_1 = \text{true value}$ under the model in the top panel of Table 1. The results are based on 1000 simulations with $n_0 = 200$ and $n_1 = n_3 = 100$.

The points corresponding to the true value of $\beta_1 = 0$ shows the empirical type I error rate for each test. All 3 methods yield close to 5% type I error. For all 3 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 3 values of β_1 according to type of estimator under the same settings used in the top panel of Table 1. Use of outcome-dependent sampling with an appropriate estimator requires a smaller sample size. In this setting, the $\hat{\beta}_Z$ method on average needs about 60% of the subjects

TABLE 2. Sample Size Needed for Testing $H_0: \beta_1 = 0$ for a Given Statistical Power for Model in Table 1

Power	True β_1	Sample Sizes for n		
		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

The results are based on 1000 simulations with $\alpha = 1.0$, $\rho = 0.5$ and $n_1 = n_2$.

who would be needed if the study were conducted with a simple random sampling scheme; the $\hat{\beta}_{IPW}$ method needs about 83%. 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 $\hat{\beta}_P$. That is, efficiency increased as β_1 is farther away from 0.

Figure 2 shows the impact of 2 factors in a given setting of outcome-dependent sampling. The impact of varying a , the cut-point that determined the strata of Y where the supplement random samples were drawn, is shown in Figure 2A. When $a = 0$, there was only plain random sampling, and in this instance all 3 methods had the same power. As a increases, however, $\hat{\beta}_Z$ (solid line) has better power than the other 2. With $\hat{\beta}_Z$, the increase in power appears to be monotonic in a . Figure 2B shows the impact of ρ , the fraction of the overall random sample in the total outcome-dependent sample ($\rho = n_0/(n_0 + n_1 + n_3)$). At $\rho = 1$, there is no supplemental sample and the outcome-dependent 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 3 estimators.

Table 3 presents different allocations of outcome-dependent sampling when the exposure variable is skewed ($X \sim \text{lognormal}$). Compared with the results in the lognormal panel of Table 1, we see that allocating more of the sample to the upper tail of Y improves the efficiency. The standard error of the slope estimator decreases as more of the sample is shifted from the lower tail to the upper tail, ie, $0.0222 \rightarrow 0.0201 \rightarrow 0.0192$ as the allocation changes $(200, 150, 50) \rightarrow (200, 100, 100) \rightarrow (200, 50, 150)$.

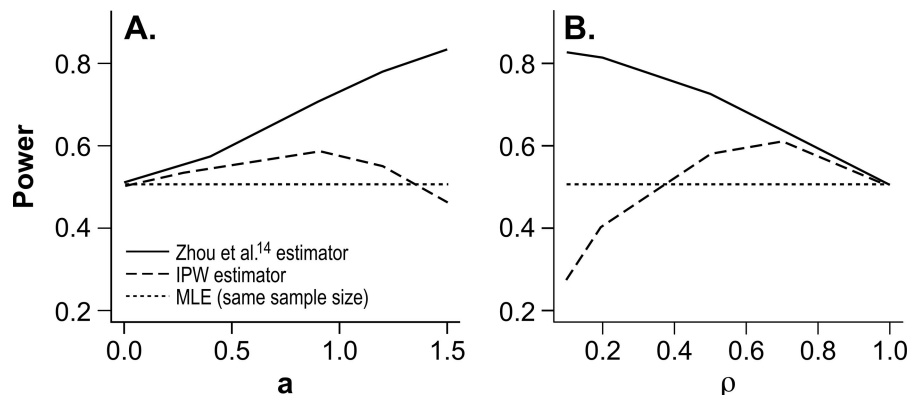
Real Data Example 1

In the motivating example noted in the introductory text, a nested case-control study would have been implemented if the outcome had been dichotomous. However, the outcome of interest, (the score on the Bayley Scale of Infant Development), was a continuous variable, and, treating it as a dichotomous variable would have resulted in a loss of statistical power. Measurements of polychlorinated biphenyls (the exposure of interest here) are expensive; thus, minimizing sample size while maintaining power was especially important. The authors drew a random sample of cohort members and 2 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/ $\mu\text{g/L}$ 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 simple random sample data alone, the $\hat{\beta}_N$ was 0.29 (SE = 0.29, $P = 0.32$). Daniels et al¹⁹ also examined and confirmed the shape of the dose-response relation in both the outcome-dependent and the simple random samples. The example demonstrates the improved efficiency obtained by the Zhou et al estimator.

Real Data Example 2

Rissanen et al²⁰ examined the relation of serum lycopene concentration to the thickness of carotid arteries among 1028 men in Finland. Using their data, we selected 400 samples in 2 ways. First, we selected a random sample of

FIGURE 2. The power for the testing $H_0: \beta_1 = 0$ versus $H_1: \beta_1 = 0.1$. (A) The power as a , cut-point for defining the strata in term of σY , varies with sample size $(n_0, n_1, n_2) = (200, 100, 100)$; (B) The power as ρ , the fraction of a simple random sample, varies with $n = 400$, $n_1 = n_2$, and $a = 1$.



400. Second, we selected a random sample of 200, and, plus a random sample of 100 from those with carotid artery thickness above the 90th percentile, and a random sample of 100 from those with carotid artery thickness below the 10th percentile. We then analyzed the data using ordinary regression model with all 1028 samples, and with the 400 selected at random. In addition, we analyzed the 200-100-100 sample using inverse probability-weighted method and then with the Zhou et al estimator. The lycopene results were adjusted for the same covariates (age, year, and sonographer) in all models. Results are given in Table 4. With the full sample of 1028, the estimated coefficient for lycopene was -0.14 with an estimated SE at 0.04 and $P = 0.0011$. With the random sample of 400, the estimated effect $\hat{\beta}_P$ was -0.10 , with SE = 0.06 and $P = 0.096$. With the 200-100-100 sample, $\hat{\beta}_{IPW}$ was -0.19 with SE = 0.08 , $P = 0.017$ and $\hat{\beta}_Z$ was -0.24 , SE = 0.07 and $P = 0.0009$. This example suggests that the outcome-dependent sampling scheme and the Zhou et al estimator provided nearly as much power as analysis of the full dataset. Furthermore, the Zhou et al estimator clearly had greater efficiency compared with the inverse-probability weighted approach. 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), a curvilinear approach would better describe the relation and could be easily

accommodated with either the inverse-probability weighted or the Zhou et al estimators.

Real Data Example 3

Korrick et al²¹ conducted a case-control study of hypertension and bone lead level. For logistic reasons the sampling probabilities for cases and controls could not be determined. Measurements of blood pressure were available. The example, presented as Appendix 1 in the online supplementary material, available with the online version of the article, shows that the proposed outcome-dependent sampling approach could be used to estimate the coefficient for bone lead in a model of blood pressure.

DISCUSSION

When there is a fixed, limited number of observations for examining the linear relation between a continuous exposure and a continuous outcome, outcome-dependent sampling is more efficient than a random sample. The simulations provided an intuitive and simple expression of this benefit: the estimator yields the same estimand as an analysis of the underlying complete population. Stated another way, for a given level of statistical power, the number of observations can be reduced by about 40% compared with a random sample. Thus, the benefits of outcome-dependent sampling apply to continuous outcomes, as well as to dichotomous ones through case-control designs. For a binary response variable, our approach, implemented with logistic regression, is equivalent to the case-control analysis.²²

To implement the weighted method, one needs to know or estimate the weights, this requires at least empirical data about the distribution of Y . Such information may not be a problem for nested studies; however, it can be difficult to calculate the weight for studies that are not nested. The difficulty 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 inverse-probability weighting suffers from the fact that in the typical setting of the outcome-dependent sampling, 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

TABLE 3. Simulation Study for Various Allocations of Outcome-Dependent Sampling With Skewed Exposure Effect

(n_0, n_1, n_2)	β_1	Methods	Mean	SE	\widehat{SE}	95% CI
(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)	0.1	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

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

TABLE 4. Results of Fitting Adjusted Models of the Carotid Artery Thickness in Relation to Serum Lycopene Concentration, According to Sampling Method and Method of Data Analysis

Sampling Method	<i>n</i>	Analysis Method	$\hat{\beta}^*$	SE(β)	<i>P</i>
All available subjects	1028	Linear regression	-0.14	0.04	0.0011
Random sample	400	Linear regression	-0.10	0.06	0.096
Random sample plus outcome tail samples [†]	400	Inverse-probability weighted estimator	-0.19	0.08	0.017
Random sample plus outcome tail samples ²	400	Zhou et al ¹⁴ estimator	-0.24	0.07	0.0009

*Units are mm/mol/L, adjusted for age, year, and sonographer.

[†]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).

hand, does not use 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, outcome-dependent sampling using the estimator of Zhou et al increased power about twice as much as the weighted estimating equation approach (Fig. 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 inverse-probability weighted 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 outcome-dependent sampling 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 (eg, >1) so that a corresponds to a clinically abnormal value. Values of $a > 1$ might seem appealing because of the resulting increase in power, especially with the Zhou et al estimator (Fig. 2). But this reward depends on an assumption about $f_{\beta}(Y, X)$ across the range of Y (see the online Appendix 2 for discussion of an example). 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 $[0.2, 0.5]$ provides much improved power with the Zhou et al estimator compared with the weighted-estimated-equation estimator, while still allocating enough observations to the simple random 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 maintaining statistical efficiency are likely to see greater use. Recently, similar ideas using outcome-dependent sampling have been extended to the situation in which information other than the exposure variable are also available for both the sample and the rest of the base population.²⁴ Methods have been developed that account for an ordinal outcome variable in a generalized linear model setting; other methods incorporate auxiliary information about the exposure variable that is available for the entire base population.²⁵ Much work, however, remains to be done; for example, how to use outcome-dependent sampling with longitudinal data is still an open question. A survey of the recent statistical research on outcome-dependent sampling design can be found in Zhou and You.²⁶

ACKNOWLEDGMENTS

We thank 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.

REFERENCES

1. Cornfield J. A method of estimating comparative rates from clinical data. Applications to cancer of lung, breast, and cervix. *J Natl Cancer Inst.* 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. *Am J Epidemiol.* 1982;115:119–128.
7. Wacholder S, Weinberg CR. Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling. *Biometrics.* 1994;50:350–357.
8. Imbens GW, Lancaster T. Efficient estimation and stratified sampling. *J Econom.* 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. *J R Stat Soc, A.* 1980;143:474–487.
11. Li R, Folsom AR, Sharrett AR, et al. 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;154:729–738.
12. Iribarren C, Folsom AR, Jacobs DR Jr, et al. 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;17:1171–1177.
 13. Suissa S. Binary methods for continuous outcome: a parametric alternative. *J Clin Epidemiol*. 1991;44:241–248.
 14. Zhou H, Weaver MA, Qin J, et al. A semiparametric empirical likelihood method for data from an outcome-dependent sampling design with a continuous outcome. *Biometrics*. 2002;58:413–421.
 15. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc*. 1952;47:663–685.
 16. Flanders WD, Greenland S. Analytical methods for two-stage case-control studies and other stratified designs. *Stat Med*. 1991;10:739–747.
 17. Zhao LP, Lipsitz S. Designs and analysis of two-stage studies. *Stat Med*. 1992;11:769–782.
 18. Godambe VP, Vijayan K. Optimal estimation for response-dependent retrospective sampling. *J Am Stat Assoc*. 1996;91:1724–1734.
 19. Daniels JL, Longnecker MP, Klebanoff MA, et al. Prenatal exposure to low-level polychlorinated biphenyls in relation to mental and motor development at 8 months. *Am J Epidemiol*. 2003;157:485–492.
 20. Rissanen T, Voutilainen S, Nyyssonen K, et al. Low plasma lycopene concentration is associated with increased intima-media thickness of the carotid artery wall. *Arterioscler Thromb Vasc Biol*. 2000;20:2677–2681.
 21. Korrick SA, Hunter DJ, Rotnitzky A, et al. Lead and hypertension in a sample of middle-aged women. *J Public Health*. 1999;99:330–335.
 22. Qin J, Zhang B. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*. 1997;84:609–618.
 23. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 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;62:1149–1160.
 26. Zhou H, You J. Semiparametric methods for data from an outcome-dependent sampling scheme. In: Hong D, Shyr Y. eds. *Quantitative Medical Data Analysis Using Mathematical Tools and Statistical Techniques*. Singapore: World Scientific Publications. 2006.