Published in final edited form as:

Stat Biosci. 2013 November 1; 5(2): . doi:10.1007/s12561-013-9080-2.

Using the Whole Cohort in the Analysis of Case-Control Data:

Application to the Women's Health Initiative

Norman E. Breslow,

Department of Biostatistics, University of Washington, Seattle, WA, USA, Tel.: +1-206-543-1044,
Fax: +1-206-616-2724

Gustavo Amorim,

Department of Statistics, University of Auckland, Auckland, NZ

Mary B. Pettinger, and

WHI Clinical Coordinating Center, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Jacques Rossouw

Division of Cardiovascular Sciences, National Heart, Lung and Blood Institute, Bethesda, MD,
USA

Norman E. Breslow: norm@uw.edu; Gustavo Amorim: gustavodecastro@hotmail.com; Mary B. Pettinger:
mpetting@whi.org; Jacques Rossouw: rossouw@nih.gov

Abstract

Standard analyses of data from case-control studies that are nested in a large cohort ignore information available for cohort members not sampled for the sub-study. This paper reviews several methods designed to increase estimation efficiency by using more of the data, treating the case-control sample as a two or three phase stratified sample. When applied to a study of coronary heart disease among women in the hormone trials of the Women's Health Initiative, modest but increasing gains in precision of regression coefficients were observed depending on the amount of cohort information used in the analysis. The gains were particularly evident for pseudo- or maximum likelihood estimates whose validity depends on the assumed model being correct. Larger standard errors were obtained for coefficients estimated by inverse probability weighted methods that are more robust to model misspecification. Such misspecification may have been responsible for an important difference in one key regression coefficient estimated using the weighted compared with the more efficient methods.

Keywords

Logistic regression; maximum likelihood; pseudolikelihood; calibration of sampling weights; model misspecification and survey sampling

1 Introduction

One of the most significant contributions of statistics to epidemiology was the development of concepts and methods for the design of, and analysis of data from, case-control studies.[3] By limiting data collection to rare disease cases and controls drawn randomly from the population at risk, such studies are highly cost effective. Since they provide information only about the covariate distributions in separate samples of cases and controls, however, one might wonder (and many did) what they tell us about the disease risks of primary interest. A seminal paper of Prentice and Pyke [26], building on early work by Cornfield [12], demonstrated that the maximum likelihood estimator of odds ratios in a population model for disease risk was obtained by applying logistic regression to the case-control data,

treating the known disease status as a random outcome. This is now the standard method of analysis of case-control data.[6]

Stratified case-control designs are especially valuable for sampling subjects from an existing cohort, using data available for all subjects to sample those most informative about the risk factors under investigation. Stored tissue samples for the sampled subjects may then be assayed for biomarkers of increased disease risk, for example. The Atherosclerosis Risk in Communities (ARIC) [37] and the Women's Health Initiative (WHI) [2] studies make frequent use of such designs. Since the investigators employed standard statistical methods for most of their published analyses, however, they routinely ignored extensive baseline information collected for the vast majority of subjects who were not sampled as cases or controls. Methods to recover some of this information, based on theory and methodology for two and three phase stratified sampling designs, have been shown in a few idealized situations to offer substantially more precise parameter estimates in comparison with standard methods.[5,16] Less is known about their performance in more realistic situations typified by ARIC and WHI.

This paper reviews selected methods of analysis of stratified case-control data that incorporate information available for all cohort members. The methods are illustrated and evaluated by application to a WHI case-control study of biomarkers for the risk of coronary heart disease (CHD).

2 Women's Health Initiative Study of Biomarkers for Coronary Heart Disease

Between September, 1993 and December, 1998, 27,347 postmenopausal women aged 50–79 years were randomized on one of two trials of hormone therapy (HT), with prevention of CHD the primary goal. The trial involving 16,608 women with an intact uterus who were randomized to a combination of estrogen plus progestin (E+P) vs placebo was stopped early because of an increase in CHD, stroke and venous thromboembolism (VTE) among those receiving treatment.[28]. The parallel trial involving 10,739 women with prior hysterectomy who were randomized to estrogen alone (E) vs placebo was also stopped early due to an increase in stroke and lack of CHD benefit.[1] Subsequently, three case-control studies were conducted to investigate the mechanisms through which HT might increase the risk of each of the cardiovascular events: CHD, stroke and VTE.

The case-control study of CHD involved 359 cases.[29] The 817 controls, matched by date of randomization, trial (E vs E+P), age, and prevalence of the particular cardiovascular disease (CVD) at baseline, were a pool drawn from all three case-control studies. Blood samples stored at baseline were assayed for inflammatory, lipid, thrombotic and genetic markers for cases and controls. The data were analyzed by ordinary logistic regression, with case-control status as the outcome and each of the biomarkers in turn as the primary risk factor. All analyses were adjusted for baseline age, trial, body mass index, waist-hip ratio, smoking, alcohol, physical activity, diabetes, history of high cholesterol, prevalent CVD, left ventricular hypertrophy, systolic blood pressure (SBP), use of anti-hypertensive medications, aspirin and statins.

Among the more interesting findings was a positive interaction between HT and low density lipoprotein cholesterol (LDL), such that the increased CHD risk for high LDL levels was even greater with treatment, and a negative interaction between HT and high density lipoprotein cholesterol (HDL), such that the decreased CHD risk for high HDL levels was even lower with treatment. The genetic polymorphism GpIIIa leu33pro (C/C+C/T vs T/T) was also associated with increased risk. Preliminary analyses showed that a model with

main and interactive effects for the log(LDL/HDL) ratio fit nearly as well as the model with main and interactive effects for both log(LDL) and log(HDL) (deviance difference 0.55, 2 degrees of freedom). Consequently, as a single comprehensive model for our methodological studies, we selected as variables of primary interest the log(LDL/HDL) ratio (centered at log 3), HT, their interaction and the GpIIIa leu33pro polymorphism. For adjustment we used only seven of the fifteen variables mentioned above. The other eight were retained for possible use as design or auxiliary variables.

We made the same exclusions for missing values as did the authors of the original paper. [29] The necessity that values for all fifteen adjustment variables used by them be known reduced the main cohort size from 27,347 to 23,301, including 276 cases. (The main cohort would have contained 24,232 subjects, including 289 cases, had we excluded only those with unknown values for the seven adjustment variables used by us.) Twenty-eight more cases and a proportionate number of controls were dropped due to missing values for the lipid levels or genotype, leaving 248 cases and 617 controls in the case-control analysis sample. Results of the standard logistic regression analysis are shown in Table 2.

The next section reviews basic methods of combining information from the case-control sample with information available for all cohort members, viewing the data as a two phase stratified sample.[19] The main cohort is the Phase I sample, assumed to have been drawn by simple random sampling from a “superpopulation”, *i.e.*, from a logistic regression model for CHD risk. The Phase II sample is the case-control sample, drawn by outcome dependent stratified sampling from the Phase I sample. A particular feature of the WHI study, however, invites consideration of a more complex design. In order to assess adherence to treatment, the investigators selected a 8.6% cohort random sample, stratified on trial and ethnicity, for assay of baseline and subsequent blood samples for selected biomarkers.[2] Baseline LDL and HDL levels were thus available for an additional 2,158 controls not sampled for the case-control study. The resulting data may be regarded as having arisen from a three phase design, with the main cohort as Phase I, all subjects with known LDL and HDL levels as Phase II and the case-control sample, for which genotype is also known, as Phase III.

3 Basic Methods for Two Phase Stratified Case-Control Samples

3.1 Notation and study design

The methods we first discuss assume (or pretend) that the Phase I information consists only of stratum frequencies. Let Y denote the binary outcome, X a p -vector of covariates and $S \in 1, \dots, J$ an indicator of population stratum depending on X and possibly also on auxiliary variables that can be used to predict missing covariate values. Interest is in estimation of β in a parametric regression model $f(y/x; \beta)$, here logistic regression. Denote the usual likelihood score by $U(\beta) = U(y/x; \beta) = \log f(y/x; \beta) / \beta$.

At Phase I we draw N observations from the population, observe Y and S for all of them, and record the frequencies N_{ij} of those having $Y = i$ and $S = j$. At Phase II n_{ij} subjects are drawn at random from the N_{ij} within each outcome $\times$ stratum category, and covariate values x_{ijk} , $k = 1, \dots, p$, are recorded. Each of the N Phase I subjects has a probability $\pi_{ij} = n_{ij}/N_{ij}$ of being sampled for Phase II depending on outcome i and stratum j .

As stratification variables for the WHI study we chose three factors on which cases and controls were matched: trial, prevalent CVD and age (Table 1). Although used for adjustment in the original analysis [29], the fact that they were matching variables precluded their evaluation as risk factors.[6, p. 31] Control sampling fractions ranged from 1.1% to 7.0% over the 12 strata. Sampled controls were older and more likely to have had prevalent CVD at baseline than cohort controls overall. One goal of two phase analyses is to correct

for such biased sampling. Cases were sampled at 81–100%, a consequence of the missing values of LDL, HDL or genotype for 28 of them. Two phase methods also account for such missing values so long as they are “missing at random” (MAR) in the sense of Little and Rubin.[17]

3.2 Analysis options

Denote by R_ℓ an indicator of whether or not the ℓ^{th} Phase I subject is sampled and by $\pi_\ell = \pi_{y \in \ell}$ the corresponding known sampling probability. The standard survey method for β estimation, termed here weighted likelihood or WL, is to solve the inverse probability weighted likelihood score equations

$$\sum_{\ell=1}^N \frac{R_\ell}{\pi_\ell} U_\ell(\beta) = 0. \quad (1)$$

Although less efficient, with large Phase I and II samples WL will faithfully reproduce the results that would have been obtained by fitting the model to a Phase I sample without missing values, even if the regression model is misspecified.[33,18]

Pseudolikelihood (PL), which relies more heavily on the model, takes for a score the logarithmic derivative, with respect to β , of the conditional density $f(y/x, R = 1; \beta, \pi)$. [4] The PL estimating equations are

$$\sum_{\ell=1}^N R_\ell \{U(Y_\ell | X_\ell; \beta) - E[U(Y_\ell | X_\ell, R_\ell = 1; \beta, \pi)]\} = 0. \quad (2)$$

The non-parametric maximum likelihood (ML) method bases inference on the profile likelihood obtained from

$$\prod_{j=1}^J \left\{ \prod_{i=0}^1 \left[\prod_{k=1}^{n_{ij}} \Pr(x_{ijk} | Y=i, S=j) \right] \Pr(Y=i | S=j)^{N_{ij}} \right\} = \prod_{j=1}^J \left\{ \prod_{i=0}^1 \left[\prod_{k=1}^{n_{ij}} f(i | x_{ijk}; \beta) g_j(x_{ijk}) \right] \left[\int f(i | x; \beta) dG_j(x) \right]^{N_{ij} - n_{ij}} \right\} \quad (3)$$

by maximizing, for fixed β , over stratum specific covariate distributions $G_j(\cdot)$ that put mass $g_j(\cdot)$ only on the observed covariate values. Computationally feasible solutions to the constrained optimization problem, involving only $p + J$ equations, were independently derived by Scott and Wild [34] and by Breslow and Holubkov [7]. (The latter considered a case-control sample also at Phase I and hence worked exclusively with logistic regression.) When the binary response model contains a parameter for each stratum, the empirical distributions of covariates within each stratum satisfy the relevant constraints and the PL and ML estimates coincide.[33] This result generalizes the finding of Prentice and Pyke for a single stratum.[26]

3.3 Results obtained with the basic methods

Table 2 shows the results of the standard (STD) case-control analysis in comparison with those for each of the basic two phase methods. These were obtained using the tps procedure in the R-package *osDesign* that implements the Breslow-Holubkov approach to ML estimation.[5,15] Walter Schill (personal communication) has noted that this approach, involving solution of concentrated Lagrange equations, sometimes yields estimates that do not satisfy the constraints on equality of observed and expected Phase II sample sizes.[7, Eq. 9] The problem seems to be associated with very large Phase I samples, however, and the

constraints were satisfied in all examples considered here. The more general program `bin2stg` in Chris Wild's *missreg* package is available as an alternative.[32]

All three two phase methods corrected for the serious bias in the control sample with respect to age and prevalent CVD. Consequently, the regression coefficients for these variables were considerably higher, and the standard errors lower, than for the standard method, reflecting the information contained in the Phase I stratum frequencies N_{ij} . (Table 1) WL estimated a noticeably smaller interaction between HT and the log(LDL/HDL) ratio. WL also produced standard errors that were generally *larger* than those for the other methods, including even the standard method, reflecting the well known and often serious loss of efficiency of WL in comparison with ordinary logistic regression for case-control samples. [33] Otherwise, the results for the four analyses were remarkably similar. The two phase methods only reduced standard errors for variables used in the Phase I stratification.

4 Accounting for more of the Phase I information

4.1 Calibration and estimation of the sampling weights

A serious limitation of the basic methods just discussed is the fact that the Phase I data must be summarized in stratum frequencies. The requirement that all strata contain at least one Phase II observation, to avoid infinite weights, limits the number of variables that may be accommodated and the number of intervals into which continuous variables may be discretized. A survey sampling technique to improve the estimation of finite population (Phase I) means is *calibration* of the weights.[14] Here we refer to sampling weights, also known as *design* weights, that are the inverses of known sampling fractions such as shown in Table 1. Calibration adjusts the design weights so that weighted estimates of auxiliary variables using only the Phase II sample equal their known Phase I totals. Another approach is to *estimate* the weights using a model for the probability of being sampled, $\pi(y, v; a) = \Pr(R = 1/y, v; a)$, that includes as predictors covariates and auxiliary variables known for all, which we denote by v . This Phase II inclusion model must include the stratification factor among the predictors so that the estimated weights account for the biased sampling.[27] Calibration and estimation of the weights are closely related, yielding identical results when both are applied using *post-stratification* to incorporate the additional information.[20,19] They work best when auxiliary variables are available that are highly correlated with the target variable whose mean is being estimated.

Adjusted weights are also used in weighted estimating equations such as those for the WL estimator shown in equation (1). An asymptotic expansion for the $\hat{\beta}_{WL}$ that solves this equation is

$$\sqrt{N} \left(\hat{\beta}_{WL} - \beta_0 \right) = \sum_{\ell=1}^N \frac{R_{\ell}}{\pi_{\ell}} \mathcal{J}^{-1} U_{\ell}(\beta_0) + o_p(1), \quad (4)$$

where $\mathcal{J} = EUU^T$ is the information. From a large sample perspective, therefore, one may view the problem as one of estimating the unknown finite population totals of the influence function components for the Phase I subjects.[10] In previous work we have used this insight to argue that the best choice for auxiliary variables are those that approximate the components of the expected influence function $E[\mathcal{J}^{-1} U \mathcal{J} V]$ given variables V known for everyone at Phase I.[8] We suggested a “plug-in” approach whereby one: (1) developed a prediction model for the partially known covariates, in our case the log(LDL/HDL) ratio and the genotype, as a function of V ; (2) used this model to impute the values of those variables for all Phase I subjects; (3) fit the interest model to the entire Phase I sample using the imputed values; (4) extracted the “delta-beta” influence function contributions; and (5) used

these as auxiliary variables in adjustment of the weights. Perhaps because no good surrogates were available for the partially missing variables, this method was unsuccessful in improving the basic results already obtained (Table 2). Consequently we adopted a simpler approach in which (continuous) age, smoking, diabetes, CVD, SBP, statins and trial, and their products with the case-control indicator, were used to adjust the weights.

Results obtained with Lumley's *survey* package[19] using calibration and estimation of the weights were very similar, and only the latter are shown in columns labeled WL in Table 3. These are noticeably changed from those obtained without adjustment of the weights (Table 2, WL). The coefficient for age increased, reflecting the fact that the matching of cases and controls on age was accounted for more completely by including continuous age, in addition to stratification by decade of age, in the Phase II inclusion model. Standard errors (SE) for the other adjustment variables were lower, reflecting the use of Phase I information about them not considered by the basic approach. The columns labeled SE1 and SE2 show the separate contributions to the overall SE from sampling at Phases I and II, respectively.[8,19] (The *survey* package uses the robust, "sandwich" formula for the former.) Nearly half the total variability can be attributed to Phase II sampling, *i.e.*, to not having complete covariate information for all Phase I subjects.

4.2 Estimated weights and pseudolikelihood

Scott and Wild [35] recently suggested estimating the sampling probabilities π that occur in the second term of the PL equations (2) using a parametric model $\pi(y,x;\alpha) = \Pr(R = 1|y,x;\alpha)$, as has just been illustrated for WL. Denote the PL equations with estimated π by $S_0(\beta, \alpha)$ and let

$$S_1(\alpha) = \sum_{\ell=1}^N \frac{R_{\ell} - \pi_{\ell}}{\pi_{\ell}(1 - \pi_{\ell})} \frac{\partial \pi_{\ell}}{\partial \alpha} \quad (5)$$

denote the likelihood score equations for α based on the inclusion (π) model. Solution of the equations $S(\beta, \alpha) = (S_0^T(\beta, \alpha), S_1^T(\alpha))^T = 0$ jointly for α and β leads to an enhanced PL estimator $\hat{\beta}_{PL}$ whose properties may be derived from standard theory for unbiased estimating equations. As would be expected from the work of Pierce [23], and as shown for WL by Robins *et al.* [27], estimation of the known sampling weights always reduces (or at least never increases) the asymptotic variance. A limitation of this approach with PL, however, is that the covariates x in the inclusion model must be a subset of those in the regression model $f(y|x;\beta)$ whose values are known for all Phase I subjects. The inclusion model may incorporate auxiliary covariates when used with WL. This feature was of importance when attempting to follow the five point imputation scheme mentioned at the end of §4.1 for construction of variables for calibration or estimation of the weights, because the prediction models for the partially missing covariates included auxiliary variables such as ethnicity and alcohol consumption that were not in the regression model. None of the results shown in the tables here, however, made use of this added generality.

Scott and Wild [35] also suggested that additional information about α might be obtained from the PL equations with estimated weights by solving the system

$$\begin{pmatrix} S_0(\beta, \alpha) \\ S_1(\alpha) - \hat{S}_1(\beta, \alpha) \end{pmatrix} = 0 \quad (6)$$

where

$$\tilde{S}_1(\beta, \alpha) = \sum_{\ell=1}^N R_{\ell} \frac{\partial \log f(y_{\ell}|x_{\ell}, R_{\ell}=1; \beta, \pi(y_{\ell}, x_{\ell}; \alpha))}{\partial \alpha}. \quad (7)$$

Indeed, when utilized with a subset of the covariates in the regression model that have finite support and a saturated inclusion model, solution of (6) yields the nonparametric maximum likelihood estimator discussed in §3. Otherwise it is not maximum likelihood, although incorporation of the S_1 term in equation (6) can never increase and may decrease the asymptotic variance. We continue to use the label “ML” for this method nonetheless.

One of us (GA) developed an R function that solves equations (6), with (ML) or without (PL) the S_1 term, and that calculates the estimated covariance matrix for the resulting $\hat{\beta}_{PL}$ or $\hat{\beta}_{ML}$, respectively, using results of Scott and Wild.[35] This function also implements WL with estimated weights. When applied with weights estimated using a saturated model for the combination of case-control indicator, trial, prevalent CVD and decade of age (Table 1), it effectively replicated the results shown in Table 2, with slight differences in the SEs and none in the coefficients. When applied with weights estimated using also the adjustment variables as described in §4.1, it likewise effectively replicated the results for WL found with the *survey* package (Table 3). It was then used to generate the last four columns of Table 3.

Just as shown in Table 2 for the basic estimators, the PL and ML estimators that incorporated the additional Phase I information yielded a noticeably larger coefficient for the $\log(\text{LDL}/\text{HDL}) \times \text{HT}$ interaction term than did WL. Standard errors for PL and ML were identical for most covariates and noticeably improved in comparison to those for WL. They were also lower, sometimes substantially so, in comparison with the corresponding SE in Table 2 that utilized only the stratification data at Phase I.

4.3 Methods for three phase samples

These ideas were further developed by Scott and Wild [35] for three phase sampling designs. Suppose that at Phase I the outcome y and certain components of the covariate vector, say $x^{(1)}$, are observed for all subjects. These data are used to construct strata for a stratified random sample of Phase II subjects ($R_1 = 1$), and additional covariate components $x^{(2)}$ are measured for those who are sampled. A further stratified sampling of Phase II subjects for Phase III ($R_2 = 1$) is conducted using strata based on $(y, x^{(1)}, x^{(2)})$, and the entire covariate vector x is observed for those with $R_1 = R_2 = 1$. Two inclusion models are developed

$$\begin{aligned} \pi_1(y, x^{(1)}; \alpha_1) &= \Pr(R_1=1|y, x^{(1)}; \alpha_1) \quad \text{and} \\ \pi_2(y, x^{(1)}, x^{(2)}; \alpha_2) &= \Pr(R_2=1|R_1=1, y, x^{(1)}, x^{(2)}; \alpha_2). \end{aligned}$$

As for two phase designs, efficiency may be improved by enriching the inclusion models at each phase with available covariate components besides those actually used to carry out the stratified sampling. For PL estimation, the likelihood scores $S_1(\alpha_1)$ and $S_2(\alpha_2)$, defined as in equation (5), are combined with $S_0(\beta, \alpha_1, \alpha_2) =$

$$\sum_{\ell=1}^N R_{1\ell} R_{2\ell} \frac{\partial \log f(y_{\ell}|x_{\ell}, R_{1\ell}=R_{2\ell}=1; \beta, \pi_1(y_{\ell}, x_{\ell}^{(1)}; \alpha_1), \pi_2(y_{\ell}, x_{\ell}^{(1)}, x_{\ell}^{(2)}; \alpha_2))}{\partial \beta}$$

into joint estimating equations for $(\beta, \alpha_1, \alpha_2)$. For “ML” estimation, these equations are augmented with terms $S_k(\beta, \alpha_1, \alpha_2) =$

$$\sum_{\ell=1}^N R_{1\ell} R_{2\ell} \frac{\partial \log f(y_\ell | x_\ell, R_{1\ell}=R_{2\ell}=1; \beta, \pi_1(y_\ell, x_\ell^{(1)}; \alpha_1), \pi_2(y_\ell, x_\ell^{(1)}, x_\ell^{(2)}; \alpha_2))}{\partial \alpha_k} \quad (8)$$

for $k = 1, 2$ to become

$$\begin{pmatrix} S_0(\beta, \alpha_1, \alpha_2) \\ S_1(\alpha_1) - \tilde{S}_1(\beta, \alpha_1, \alpha_2) \\ S_2(\alpha_2) - \tilde{S}_2(\beta, \alpha_1, \alpha_2) \end{pmatrix} = 0. \quad (9)$$

When saturated models are used with discrete $x^{(1)}$ and $x^{(2)}$ at both phases, solution of (9) does in fact yield the nonparametric maximum likelihood estimator.[16,35]

Amorim’s R function handles both two and three phase samples and produced the results shown in Table 4. Recall that the Phase II sample in the three phase design consisted of subjects with known LDL and HDL levels. Most of these were from the 8.6% cohort random sample that was stratified on trial and ethnicity. Since it was necessary to put the stratum factor in the Phase III inclusion model to account for the biased sampling, we were obliged in principle to include ethnicity and its interaction with trial also in the regression model as additional adjustment variables. This may be avoided, however, if we are willing to assume that they are conditionally independent of outcome given the covariates already in the model. This was a reasonable assumption on general grounds and in fact, when ethnicity and its interaction with trial were included in the model, their coefficients were all near zero. Whether or not they were included made little difference in the results for the other covariates. Consequently, to achieve greater comparability with earlier results, they were excluded from the regression model for which results are shown in Table 4.

The inclusion model for Phase II had as predictors: the three-way combination of outcome, trial and ethnicity; the four way combination involving the sampling strata and outcome as shown in Table 1 and used to obtain the results in Tables 2 and 3; and age, treatment, smoking, prevalent CVD, diabetes, SBP, statins plus the product of each with the case-control indicator. The inclusion model for Phase III added to these terms the three way combination of outcome, treatment and log(LDL/HDL). The idea was to try to bring into the analysis the information available from both phases on marginal associations between outcome and each of the covariates.

Incorporation of the additional information on LDL and HDL from the 8.6% sample had a noticeable effect on precision for three of the variables of prime interest, namely, log(LDL/HDL), HT and their interaction. In comparison with results from Table 3, SEs for these variables were reduced for all three methods (WL, PL, ML). There was stronger evidence from PL and ML for an interaction between HT and the log ratio; the results from WL were still inconclusive. The SEs for the adjustment variables decreased for WL whereas several increased slightly for PL and WL in comparison with the two phase results. Consequently, the loss of precision from using WL relative to PL/ML was reduced.

4.4 Summary of results

Bar plots in Figure 1 show relative efficiencies (in percent) of the estimation methods considered in this paper relative to the standard method of logistic regression analysis of data from the case-control sample. These were estimated by squaring the ratios of standard

errors for the two methods being compared (standard vs other). Each of the eight panels, corresponding to eight of the eleven variables in the regression equation, contains three pairs of bars. The left bar in each pair shows the relative efficiency of WL estimation, the right bar that of PL estimation. ML results are not shown since they were so close to those for PL with these data. The left most pair of bars, labelled “Strata”, is for the basic stratified sampling methods described in §3 where the only Phase I data used were the 24 stratum frequencies (Table 1). The middle pair is for the methods described in §4.1 and §4.2 that estimated the Phase II inclusion probabilities by adding to the stratum indicators already in the inclusion model the other regression variables (age in years, HT, etc.), and products of these with case-control status. The last pair is for the three phase methods described in §4.3, where the Phase III inclusion model contains the three way interaction involving case-control status, HT and the log(LDL/HDL) ratio, plus other terms.

Results for the eight covariates shown in Figure 1 fall naturally into four groups. The first contains one covariate, the genotype, which was available only for subjects in the case-control (Phase III) sample and was essentially independent of the other covariates. Since the Phase I and Phase II data provided no information about genotype, the standard errors were virtually unchanged from those obtained with the standard estimator. Those for WL were greater than for PL, as noted earlier, but otherwise varied little. The second group of three covariates comprises HT, the log(LDL/HDL) ratio and their interaction. Here there was a noticeable improvement in precision resulting from inclusion of the Phase II information on LDL and HDL in the three phase analysis. There was also a slight improvement for terms involving HT, which is known at Phase I, in the two phase analysis. The third group of covariates comprises age and history of CVD, which were used for stratification of the case-control sample. Inclusion of information about them via the stratum frequencies was essential to avoid bias. It also increased precision, with the increase for the basic stratified analysis comparable to the increase for the two and three phase analyses with estimated weights. The fourth and last group contains the remaining covariates, not used for stratification, for which information was available only at Phase I. This group comprises statins, diabetes and trial (results not shown) in addition to smoking status and SBP (results shown in Figure 1). The pattern was the same for all five covariates in this group. No increase in precision over the standard logistic regression analysis was obtained from the basic stratified analysis. Using the Phase I information on these covariates in the two and three phase analyses, however, did improve precision, in some cases substantially. Overall, the results were as expected. Increase in precision of a regression coefficient was observed only with analyses that included, through estimation of the weights, additional information about the marginal association of the corresponding covariate and outcome.

5 Discussion

Use of data available for subjects who were not sampled for the case-control study helped to correct bias in odds ratio estimates for variables used to match controls with cases, and increased the precision of all estimates. Accounting for control sampling fractions that, by design, depended on age and prevalent CVD (Table 1) yielded results that clearly demonstrated the important association of CHD risk with these two factors (Table 2). Whereas the standard analysis returned an age coefficient near zero, as expected due to the matching, in the adjusted analysis each decade of age was estimated to increase risk by a factor of 1.8. The broad intervals used for age stratification, however, provided incomplete adjustment for the dependence of the sampling fractions on age. When the inclusion models used actual age in addition to age interval, risk was estimated to double with each 10 year increase (Tables 3 and 4). A similarly dramatic change was observed in the coefficient for CVD.

One focus of the analysis was on a possible statistical interaction between the log(LDL/HDL) ratio and HT. The SE estimated by WL for the interaction term was 0.51–0.53 for analyses that failed to incorporate the extra data on LDL and HDL levels (Tables 2 and 3), but dropped to 0.44 when this information was incorporated in the three phase analysis (Table 4). Similarly, the SEs estimated by PL and ML dropped from 0.47–0.48 to 0.41. By contrast, the SE for the GpIIIa genotype was not improved in comparison with the standard analysis, reflecting the fact that only subjects in the case-control sample had relevant information about the genotype.

There was a notable difference between the WL and PL/ML analyses in the coefficient of the interaction term and none of the WL analyses suggested a statistically significant interaction effect. Differences for other covariates were not nearly so pronounced. The discrepancy probably resulted from model misspecification. When quadratic terms in log(LDL/HDL) were added to the models, the quadratic interaction term had a coefficient of about -1.5 with a SE of 1.0, though this varied a little depending on the analysis. Although not strong evidence for misspecification, the result certainly suggests that the rate of increase of the HT effect with log(LDL/HDL) slowed with increasing values of the LDL/HDL ratio. Scott and Wild [31] have shown that, when a linear logistic regression model is fitted to simulated case-control data but the true model is negative quadratic, the WL estimator gives a better, *i.e.*, lower, estimate of the slope for the small percent of the population that is most at risk (higher LDL/HDL) whereas the standard (ML) estimator provides a better (higher) estimate of the slope for the majority at lower risk. It seems likely, though remains to be proven, that a similar phenomenon occurs with the more complex interaction model considered here. This would explain the fact that WL gave a noticeably lower estimate for the interaction. The choice between WL and either PL or ML will depend on how much credence one is willing to give the assumed model and how important it is to approximate the result that “would have been obtained” with complete data in case of misspecification. For ordinary case-control studies without stratification, the standard and generally accepted unweighted logistic regression approach yields the PL/ML estimator.

A referee suggested that an alternative explanation for the difference in results between WL and PL might be the presence of “outliers”. The subject with the greatest influence on the HT*log(HDL/LDL) interaction coefficient in the standard analysis was an untreated CHD case with the fifth lowest value for the HDL/LDL ratio. When this subject was omitted, the interaction coefficients shown in Table 2 dropped from 0.99 ± 0.49 to 0.86 ± 0.49 for PL and from 0.56 ± 0.50 to 0.44 ± 0.50 for WL. While not providing the explanation for the WL-PL discrepancy, this nonetheless serves as an important reminder that individual subjects can have a noticeable impact on results.

The precision achieved by PL and ML, as reflected in the SEs in Tables 2–4, was essentially the same for the basic stratified and more complex two and three phase analyses with estimated weights. On the one hand this is reassuring, since the ML estimate requires more calculation and is susceptible to convergence problems. On the other hand, ML was substantially more efficient than PL in limited simulations.[35] Further numerical work will be required to assist in evaluating the tradeoff between possible bias due to model misspecification and increased efficiency stemming from heavier reliance on the model.

The description of the study design in §2 called attention to the exclusions, from both the main cohort and the case-control (Phase II) sample, mandated by missing covariate values. Indeed, to construct a data set as close as possible to that used for the original report,[29] 931 subjects were excluded from the main cohort due to missing values for eight secondary covariates not used here. Some of this information could be recovered, at least in principle, by adding one more phase of sampling to the two or three phases already considered.

Important gains in precision would only be possible, however, if the secondary covariates were highly correlated with those included in the model. Since the addition of 931 subjects to the main cohort (Phase I sample) would only increase its size by 4%, the extra complexity does not appear warranted.

Exclusions from the case-control sample were due to missing values for the biomarkers, particularly LDL and genotype. Over 28% of the missing genotype values were from a single clinical center, out of a total of 40, that had difficulty in providing quality DNA. Overall, the fraction of subjects with missing genotype declined with year of randomization. Neither of these features call into question the MAR assumption; there was no obvious dependence of the indicator of missing genotype with measured covariates. On the other hand, the MAR assumption for LDL is not tenable. This is because LDL is a derived variable, calculated from values for total cholesterol, HDL and triglyceride. The calculation is not considered valid for a small number of subjects, including in particular a few with uncontrolled diabetes, who have elevated triglyceride, and these subjects therefore have missing LDL. Consequently, the results reported both here and in the original article [29] are best considered as applicable only to the sub-population of subjects, fortunately the vast majority, who do not have elevated triglyceride.

One key feature of the data not utilized in our analyses is the statistical independence between HT and covariates, a consequence of treatment randomization. Various methods of analysis exploit this additional information.[38,11,36] When the disease is rare and one is interested only in interactions between covariates and HT, a “case-only” analysis of ordinary case-control data yields a more efficient estimate than provided by the interaction term in standard logistic regression.[38] Prentice himself adopted this method to investigate interactions between HT and dietary factors.[25] If terms other than the interactions are of interest, methods developed and implemented by Shin and colleagues in the *luca* package are available.[36] Their application with the WHI CHD case-control data improved precision even more than the two phase methods considered here. The coefficient $\pm$ SE for the interaction term was 0.75 ± 0.38 , for the main effect of $\log(\text{LDL}/\text{HDL})$ 0.94 ± 0.31 and for the main effect of treatment 0.19 ± 0.15 , which results may be compared with those for the standard estimator shown in Table 2. Dai and colleagues [13] proposed an extension that incorporates the two phase data *and* exploits the independence assumption, using another WHI case-control study for illustration. Their method was not considered here, however, since at present it assumes that the only Phase I data available consist of the outcome and the randomized treatment.

Although the original paper stated that the 359 cases of CHD occurred within four years of randomization,[29] in fact all cases identified by February, 2001 were utilized both in the original analyses and in those reported here, including 43 cases that occurred between 4.0 and 5.7 years from randomization. The binary outcome had a different interpretation depending on year of randomization, since those randomized earlier had a longer period during which to develop CHD. Those randomized during 1998 did not have opportunity even for a full four year follow-up. In principle, therefore, it may have been more appropriate to consider the outcome as time-to-event and to analyse the data with Cox regression rather than logistic regression. In this context, the two phase design may be viewed as a stratified version of Prentice’s case-cohort design.[24] One advantage of weighted, sample survey methods is that they may be used with whatever analysis model one chooses. The design object constructed by Lumley’s *survey* package, and used to produce the results for estimated weights shown in columns labelled WL of Table 3, may also be used to fit the analogous Cox model.[9] Results for the Cox model, as it turned out, were nearly indistinguishable from those shown in Table 3. The coefficient $\pm$ SE for age, for example, was 0.68 ± 0.13 and for the $\log(\text{LDL}/\text{HDL})\pm\text{HT}$ interaction 0.52 ± 0.51 . From a

practical standpoint, it made little difference that the follow-up times were ignored. More efficient methods for the Cox model, analogous to ML for logistic regression, have been proposed.[30,22] Both are computationally complex and do not appear to have been implemented in a form suitable for general use with continuous covariates and stratified two phase samples. One feasible alternative approach is multiple imputation, though the increased efficiency is gained at the price of assuming a correct imputation model.[21]

In conclusion, recently developed statistical methods can increase the precision of results of case-control and case-cohort studies, either by using more of the available data or by exploiting statistical independence when one of the risk factors is a randomized treatment. Software is available for the R package to implement these methods. We hope that they will experience greater use in future years.

Acknowledgments

To Ross, with thanks for your support and leadership

References

1. Anderson GL, Limacher M, Assaf AR, Bassford T, Beresford SAA, Black H, Bonds D, Brunner R, Brzyski R, Caan B, Chlebowski R, Curb D, Gass M, Hays J, Heiss G, Hendrix S, Howard BV, Hsia J, Hubbell A, Jackson R, Johnson KC, Judd H, Kotchen JM, Kuller L, LaCroix AZ, Lane D, Langer RD, Lasser N, Lewis CE, Manson J, Margolis K, Ockene J, O'Sullivan MJ, Phillips L, Prentice RL, Ritenbaugh C, Robbins J, Rossouw JE, Sarto G, Stefanick ML, Van Horn L, Wactawski-Wende J, Wallace R, Wassertheil-Smoller S. Effects of conjugated, equine estrogen in postmenopausal women with hysterectomy - The Women's Health Initiative randomized controlled trial. *JAMA- Journal of the American Medical Association*. 2004; 291:1701–1712.
2. Anderson GL, Manson J, Wallace R, Lund B, Hall D, Davis S, Shumaker S, Wang CY, Stein E, Prentice RL. Implementation of the Women's Health Initiative study design. *Annals of Epidemiology*. 2003; 13:S5–S17. [PubMed: 14575938]
3. Breslow NE. Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*. 1996; 91:14–28. [PubMed: 12155399]
4. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. *Biometrika*. 1988; 75:11–20.
5. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcomes applied to Wilms tumor prognosis. *Applied Statistics*. 1999; 48:457–468.
6. Breslow, NE.; Day, NE. *Statistical Methods in Cancer Research I: The Analysis of Case-Control Studies*. International Agency for Research on Cancer; Lyon: 1980.
7. Breslow NE, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society (Series B)*. 1997; 59:447–461.
8. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences*. 2009; 1:32–49. [PubMed: 20174455]
9. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology*. 2009; 169:1398–1405. [PubMed: 19357328]
10. Breslow NE, Wellner JA. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*. 2007; 34:86–102.
11. Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*. 2005; 92:399–418.
12. Cornfield J. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*. 1951; 11:1269–1275. [PubMed: 14861651]

13. Dai JY, LeBlanc M, Kooperberg C. Semiparametric estimation exploiting covariate independence in two-phase randomized trials. *Biometrics*. 2009; 65:178–187. [PubMed: 18479485]
14. Deville JC, Särndal CE, Sautory O. Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*. 1993; 88:1013–1020.
15. Haneuse S, Saegusa T, Lumley T. osDesign: An R package for the analysis, evaluation, and design of two-phase and case-control studies. *Journal of Statistical Software*. 2011; 43:1–29. [PubMed: 22003319]
16. Lee AJ, Scott AJ, Wild CJ. Efficient estimation in multi-phase case-control studies. *Biometrika*. 2010; 97:361–374.
17. Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. Wiley; New York: 2002.
18. Lumley, T. Robustness of Semiparametric Efficiency in Nearly-Correct Models for Two-Phase Samples; UW Biostatistics Working Paper. 2009. p. 351 www.bepress.com/uwbiostat/paper351
19. Lumley, T. *Complex Surveys: A Guide to Analysis Using R*. John Wiley & Sons, Inc; Hoboken, N.J: 2012.
20. Lumley T, Shaw PA, Dai JY. Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*. 2011; 79:200–220. [PubMed: 23833390]
21. Marti H, Chavance M. Multiple imputation analysis of case-cohort studies. *Statistics in Medicine*. 2011; 30:1595–1607. [PubMed: 21351290]
22. Nan B. Efficient estimation for case-cohort studies. *Canadian Journal of Statistics*. 2004; 32:403–409.
23. Pierce DA. The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Annals of Statistics*. 1982; 10:475–478.
24. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
25. Prentice RL, Caan B, Chlebowski RT, Patterson R, Kuller LH, Ockene JK, Margolis KL, Limacher MC, Manson JE, Parker LM, Paskett E, Phillips L, Robbins J, Rossouw JE, Sarto GE, Shikany JM, Stefanick ML, Thomson CA, Van Horn L, Vitamins MZ, Wactawski-Wende J, Wallace RB, Wassertheil-Smoller S, Whitlock E, Yano K, Adams-Campbell L, Anderson GL, Assaf AR, Beresford SAA, Black HR, Brunner RL, Brzyski RG, Ford L, Gass M, Hays J, Heber D, Heiss G, Hendrix SL, Hsia J, Hubbell FA, Jackson RD, Johnson KC, Kotchen JM, LaCroix AZ, Lane DS, Langer RD, Lasser NL, Henderson MM. Low-fat dietary pattern and risk of invasive breast cancer - the Women's Health Initiative randomized controlled dietary modification trial. *JAMA-Journal of the American Medical Association*. 2006; 295:629–642.
26. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.
27. 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.
28. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, Jackson RD, Beresford SAA, Howard BV, Johnson KC, Kotchen M, Ockene J. Risks and benefits of estrogen plus progestin in healthy postmenopausal women - Principal results from the Women's Health Initiative randomized controlled trial. *JAMA-Journal of the American Medical Association*. 2002; 288:321–333.
29. Rossouw JE, Cushman M, Greenland P, Lloyd-Jones DM, Bray P, Kooperberg C, Pettinger M, Robinson J, Hendrix S, Hsia J. Inflammatory, lipid, thrombotic, and genetic markers of coronary heart disease risk in the Women's Health Initiative trials of hormone therapy. *Archives of Internal Medicine*. 2008; 168:2245–2253. [PubMed: 19001202]
30. Scheike TH, Martinussen T. Maximum likelihood estimation for Cox's regression model under case-cohort sampling. *Scandinavian Journal of Statistics*. 2004; 31:283–293.
31. Scott A, Wild C. On the robustness of weighted methods for fitting models to case-control data. *Journal of the Royal Statistical Society Series B-Statistical Methodology*. 2002; 64:207–219.
32. Scott, AJ.; Wild, C. Calculating efficient semiparametric estimators for a broad class of missing-data problems. In: Liski, EP.; Isotalo, J.; Niemelä, J.; Puntanen, S.; Styan, GPH., editors.

- Festschrift for Tarmo Pukkila on his 60th Birthday. Dept. of Mathematics, Statistics and Philosophy, University of Tampere; 2006. p. 301-314.
33. Scott AJ, Wild CJ. Fitting logistic models under case-control or choice based sampling. *Journal of the Royal Statistical Society (Series B)*. 1986; 48:170–182.
 34. Scott AJ, Wild CJ. Fitting regression models to case-control data by maximum likelihood. *Biometrika*. 1997; 84:57–71.
 35. Scott AJ, Wild CJ. Fitting regression models with response-biased samples. *Canadian Journal of Statistics-La revue canadienne de statistique*. 2011; 39:519–536.
 36. Shin JH, McNeney B, Graham J. Case-control inference of interaction between genetic and non-genetic risk factors under assumptions on their distribution. *Statistical Applications in Genetics and Molecular Biology*. 2007; 6:1–41.
 37. The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) study: Design and objectives. *American Journal of Epidemiology*. 1989; 129:687–702. [PubMed: 2646917]
 38. Umbach DM, Weinberg CR. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine*. 1997; 16:1731–1743. [PubMed: 9265696]

Fig. 1.
Relative efficiency (%) of two and three phase sampling methods vs ordinary logistic regression

Table 1
Phase I stratum frequencies $N_{i,j}$ and Phase II sampling fractions $\pi_{i,j}$, in percent, by stratum (j)

Age	50-59 years				60-69 years				70-79 years			
	No		Yes		No		Yes		No		Yes	
CVD												
Trial	E	E+P	E	E+P	E	E+P	E	E+P	E	E+P	E	E+P
j	1	2	3	4	5	6	7	8	9	10	11	12
Frequencies and sampling fractions of cases												
$N_{1,j}$	11	27	2	3	35	47	22	20	38	43	13	15
$\pi_{1,j}$	82	81	100	100	91	96	95	90	89	86	85	93
Frequencies and sampling fractions of controls												
$N_{0,j}$	2561	4440	156	124	3595	5979	424	358	1814	2707	330	339
$\pi_{0,j}$	1.2	1.1	4.5	4.0	2.9	2.5	5.2	3.9	4.0	4.4	7.0	5.9

Table 2

Results of standard case-control and two phase analyses; phase I variables limited to indicators of strata used for sampling

Model term	STD		WL		PL		ML	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
log(LDL/HDL)	0.83	0.35	0.98	0.38	0.82	0.35	0.83	0.35
HT	0.16	0.17	0.07	0.18	0.16	0.17	0.15	0.17
HT*log(L/H)	0.97	0.49	0.56	0.50	0.99	0.49	0.99	0.49
GpIIa	0.43	0.18	0.44	0.19	0.44	0.18	0.43	0.18
Age/10	0.07	0.14	0.56	0.13	0.58	0.12	0.59	0.12
Current smoker	1.45	0.24	1.49	0.25	1.42	0.24	1.42	0.24
Diabetes	1.30	0.28	1.52	0.28	1.35	0.28	1.35	0.28
Prevalent CVD	0.64	0.21	1.19	0.19	1.18	0.17	1.18	0.17
SBP/100	2.30	0.49	2.41	0.56	2.29	0.49	2.35	0.49
Statins	0.67	0.25	0.60	0.27	0.64	0.25	0.65	0.25
Trial E+P	0.24	0.17	0.10	0.16	0.11	0.15	0.13	0.15

Abbreviations: STD=standard (logistic regression); WL=weighted likelihood

PL=pseudolikelihood; ML=maximum likelihood

Table 3

Results of two phase analyses: stratum indicators plus additional covariates at Phase I

Model term	WL			PL			ML		
	Coef.	SE	SE1	SE2	Coef.	SE	Coef.	SE	SE
log(LDL/HDL)	1.00	0.39	0.30	0.25	0.83	0.34	0.83	0.35	
HT	0.15	0.17	0.13	0.10	0.18	0.17	0.23	0.15	
HT*log(L/H)	0.53	0.52	0.36	0.37	0.96	0.47	0.96	0.48	
GpIIa	0.45	0.20	0.14	0.15	0.45	0.18	0.45	0.18	
Age/10	0.67	0.13	0.11	0.08	0.71	0.12	0.69	0.12	
Current smoker	1.49	0.23	0.17	0.16	1.41	0.18	1.40	0.18	
Diabetes	1.28	0.25	0.18	0.18	1.15	0.19	1.14	0.19	
Prevalent CVD	1.17	0.18	0.15	0.11	1.16	0.18	1.16	0.18	
SBP/100	1.99	0.51	0.41	0.29	1.88	0.42	1.93	0.42	
Statins	0.67	0.24	0.18	0.16	0.70	0.21	0.70	0.21	
Trial E+P	0.11	0.17	0.13	0.10	0.12	0.15	0.13	0.15	

Table 4
Results of three phase analyses: stratum indicators plus additional covariates at both phases

Model term	WL		PL		ML	
	Coef.	SE	Coef.	SE	Coef.	SE
log(LDL/HDL)	1.18	0.36	0.99	0.31	1.03	0.31
HT	0.09	0.16	0.21	0.15	0.16	0.15
HT*log(L/H)	0.63	0.47	1.08	0.42	1.00	0.43
GpIIa	0.45	0.20	0.46	0.18	0.45	0.18
Age/10	0.68	0.13	0.63	0.12	0.65	0.12
Current smoker	1.40	0.21	1.26	0.19	1.26	0.19
Diabetes	1.27	0.22	1.12	0.20	1.02	0.19
Prevalent CVD	1.11	0.19	1.13	0.18	1.11	0.18
SBP/100	1.89	0.51	1.71	0.43	1.85	0.43
Statins	0.61	0.23	0.67	0.21	0.68	0.21
Trial E+P	0.09	0.17	0.08	0.15	0.13	0.15