Published in final edited form as:

Stat Methods Med Res. 2019 March ; 28(3): 822–834. doi:10.1177/0962280217737157.

Power and Sample Size for Multivariate Logistic Modeling of Unmatched Case-Control Studies

Mitchell H. Gail* and **Sebastien Haneuse**

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD (Mitchell H. Gail); Harvard T.H. Chan School of Public Health, Boston, MA (Sebastien Haneuse)

Abstract

Sample size calculations are needed to design and assess the feasibility of case-control studies. Although such calculations are readily available for simple case-control designs and univariate analyses, there is limited theory and software for multivariate unconditional logistic analysis of case-control data. Here we outline the theory needed to detect scalar exposure effects or scalar interactions while controlling for other covariates in logistic regression. Both analytical and simulation methods are presented, together with links to the corresponding software.

Sample size calculations are needed to design and assess the feasibility of case-control studies. Such calculations are readily available for simple case-control designs and univariate analyses (see e.g. Chapter 7 of (1)). More complex analyses, such as multivariate logistic regression that includes confounders as well as the exposure of interest and logistic models to test for interactions cannot be dealt with by these simple methods. For cohort data, Whittemore (2) and Demidenko (3) presented power calculations for the logistic model, but these cannot be used for case-control data. Theory in McKeown-Eyssen and Thomas (4) and in Lubin and Gail (5) can be used for general logistic models with case-control data. Nonetheless, a general treatment of the logistic model for unmatched case-control data is not available, nor is software. Here we outline that theory and refer to newly developed software.

Simulations provide an alternative approach for sample size and power calculations for unmatched case-control data (6, 7). Not only can simulation allow for multivariate logistic modeling, it can also handle other potential complications such as measurement error.

In this paper, we present examples to illustrate these methods and point the reader to the needed software. We also discuss some of the practical problems of obtaining information on the joint distribution of exposure and confounders as well as the associated log relative odds that are needed for these multivariate calculations.

*Correspondence to Dr. Mitchell H. Gail, National Cancer Institute, 9609 Medical Center Drive, Room 7E138, Rockville, MD 20850-9780 (gailm@mail.nih.gov).

METHODS

Motivating data: population-based case-control study of breast cancer in Poland

Women with incident breast cancer (cases) aged 20–74 years and diagnosed between 2000 and 2003 in Warsaw or Łódź were compared to controls who were frequency matched to cases on city and 5-year age group and sampled using the Polish Electronic System, which has demographic data on all residents of Poland (8). We use data from the 1968 cases and 2162 controls with complete information on age, education level (0=none, 1=completed elementary school, 2=some high school, 3=completed high school, 4=training after high school, 5=some college, 6=graduated college, 7=post-graduate/professional), age at first birth (years), age at menarche (years), history of a biopsy for benign breast disease (0=0, 1=at least one), family history of breast cancer (0=0, 1=at least one female relative), and body mass index (BMI in kg/m²).

The means and standard deviations of risk factor values are shown Table 1 separately for cases and controls. The ages are similar because of frequency matching. Increased risk is associated with higher education level, delayed first birth, early age at menarche, previous benign breast biopsy and family history. Increasing BMI is protective in these data, in both women under age 50 and women aged 50 or more, whereas other studies have found that increased BMI is protective in pre- but not post-menopausal women.

The estimated variances, covariances and correlations for the risk factors are presented separately for cases and controls in Table 2. Higher education level is positively correlated with age at first birth and benign breast biopsy, and inversely correlated with age, age at menarche, and BMI. We will use these data to motivate and provide parameter estimates for sample size calculations.

Model formulation and analytical sample size calculations

Consider a logistic risk model of the form

$$\text{logit}\{P(D = 1|\mathbf{X}_1, X_2)\} \equiv \log\{P(D = 1|\mathbf{X}_1, X_2)\}/\{P(D = 0|\mathbf{X}_1, X_2)\} = \beta_0 + \beta_1^T \mathbf{X}_1 + \beta_2 X_2 \quad (1)$$

where $D = 1$ if disease is present and 0 otherwise, $\mathbf{X}_1$ is a vector of covariates such as confounders, X_2 is a scalar covariate of primary interest, and the intercept β_0 pertains to the case-control population. For example, X_2 might be a scalar continuous, binary or ordinal exposure measurement, and one might want to test that there is no exposure effect, namely $\beta_2 = 0$. This formulation can also be used to test for an interaction between exposure (call it X_3) and a particular covariate (call it X_1) by setting $X_2 = X_1 X_3$ and $\mathbf{X}_1^T = (X_1, X_3)$. Then the interaction parameter is β_2 in the model $\text{logit}\{P(D = 1|\mathbf{X}_1, X_3)\} = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_2 X_1 X_3$, which has the same form as equation (1). Thus, this paper concerns sample size calculations to test $\beta_2 = 0$, which covers a variety of useful applications.

Let $\mathbf{X} = (\mathbf{X}_1^T, X_2)^T$ be the vector of all covariates with distribution function F in the source population from which the case-control sample is drawn. If the prevalence of disease $\pi = P(D = 1)$ is known in the general population, the model corresponding to equation (1) in the source population with intercept β_0^* satisfies

$$\pi = \int P^*(D = 1|x) dF(x) \equiv \int \exp(\beta_0^* + \beta^T x) \{1 + \exp(\beta_0^* + \beta^T x)\}^{-1} dF(x), \quad (2)$$

which can be solved for $\beta_0^* = \beta_0 - \log(\text{case sampling fraction}) + \log(\text{control sampling fraction})$.

From Prentice and Pyke(1979), we know that the maximum likelihood estimate $\hat{\beta}_2$ has variance $\text{Var}(\hat{\beta}_2) = (I_{\beta_2\beta_2} - I_{\beta_2\beta} I_{\beta\beta}^{-1} I_{\beta\beta_2})^{-1}$, where I is minus the expectation of the Hessian of the log-likelihood for a cohort study and $\beta = (\beta_0, \beta_1^T)^T$ includes all parameters except β_2 . The “expectation” is with respect to the retrospective sampling, for which we need the distribution of $\mathbf{X}$ in cases,

$$F_1(x) = \pi^{-1} \int_{-\infty}^x P^*(D = 1|u) dF(u), \quad (3)$$

and the distribution of $\mathbf{X}$ in controls,

$$F_0(x) = (1 - \pi)^{-1} \int_{-\infty}^x P^*(D = 0|u) dF(u). \quad (4)$$

Details on how to compute the retrospective expectation of the information matrix are in the Appendix. The information also depends on the fraction λ of cases in the n cases plus controls. In particular, there are λn cases and $(1 - \lambda)n$ controls, and we calculate the total required sample size n .

The sample size depends on the specific statistical test used. Suppose we use a one-sided Wald test and reject the null hypothesis $\beta_2 = 0$ if $\hat{\beta}_2 > Z_{1-\alpha/2} \{n \hat{\text{Var}}_0(n^{1/2} \hat{\beta}_2)\}^{1/2} \equiv Z_{1-\alpha/2} n^{1/2} \hat{\sigma}_0$. Here we compute the information assuming the null hypothesis is true to calculate the null variance $\sigma_0^2 = \text{Var}_0(n^{1/2} \hat{\beta}_2)$. The quantity $Z_{1-\alpha/2}$ is the $(1 - \alpha/2)^{\text{th}}$ quantile of the standard normal distribution, chosen to guarantee size less than α for a two-sided test. For example, $Z_{1-\alpha/2} = 1.96$ controls α at 0.05. If the desired power is γ , standard methods (e.g. Chapter 7 in (1)) yield the sample size formula

$$n = [Z_\gamma \sigma + Z_{1-\alpha/2} \sigma_0]^2 / \beta_2^2. \quad (5)$$

In this equation, $\sigma^2 \equiv \text{Var}(n^{1/2} \hat{\beta}_2)$ is calculated by taking the expectation of the information matrix under the alternative hypothesis for β_2 . If instead we test the null hypothesis using $\hat{\beta}_2 > Z_{1-\alpha/2} \hat{\sigma}$, the required sample size is given by

$$n = \{Z_\gamma + Z_{1-\alpha/2}\}^2 \sigma^2 / \beta_2^2. \quad (6)$$

Instead of Wald tests, one can use tests based on the score statistic

$$U_{\beta_2: \beta_2=0} = \sum_{i=1}^{n_1} X_{2i} \{1 - P(D_i = 1 | X_{1i}; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 = 0)\} - \sum_{i=n_1+1}^n X_{2i} P(D_i = 1 | X_{1i}; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 = 0), \quad (7)$$

where the first sum is over cases and the second sum is over controls. Equation (7) requires maximum likelihood estimates $\hat{\beta}_0, \hat{\beta}_1$ obtained under the null hypothesis $\beta_2 = 0$. The variance $n\sigma_0^2 \equiv \text{Var}(U_{\beta_2: \beta_2=0}) = (I_{\beta_2 \beta_2} - I_{\beta_2, \beta} I_{\beta, \beta}^{-1} I_{\beta, \beta_2})$ is estimated by plugging in these estimates of $\beta = (\beta_0, \beta_1^T)^T$ and $\beta_2 = 0$. Under the alternative hypothesis $\beta_2 \neq 0$, the variance is $n\sigma^2 \equiv \text{Var}(U_{\beta_2: \beta_2 \neq 0}) = (I_{\beta_2 \beta_2} - I_{\beta_2, \beta} I_{\beta, \beta}^{-1} I_{\beta, \beta_2})$, where $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ are unconstrained maximum likelihood estimates at which the information is evaluated. The non-null expectation of $U_{\beta_2: \beta_2=0}$ is

$$E(U_{\beta_2: \beta_2=0}) = n[\lambda \int x_2 \{1 - P(D = 1 | \mathbf{x}_1, x_2; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 = 0)\} dF_1(\mathbf{x}_1, x_2) - (1 - \lambda) \int x P(D = 1 | \mathbf{x}_1, x_2; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 = 0) dF_0(\mathbf{x}_1, x_2)] \equiv n\theta. \quad (8)$$

Equations (5) and (6) can be used for score tests by replacing $\hat{\beta}_2$ by $U_{\beta_2: \beta_2=0}/n$, β_2^2 by θ^2 , and the definitions of σ_0^2 and σ^2 for scores. Thus, there are potentially four tests of the null hypothesis and four sample size calculations.

Required parameters or data and software

To compute sample sizes, we need to have estimates of the log-odds ratios associated with various covariates in model (1) and we need to know the joint distribution of the covariates in the source population, F . In order to compute the distribution of covariates in cases and

controls from equations (3) and (4), we also need to know the incidence disease in the general population, which allows us to compute the conditional probability of disease in the source population, $P^*(D=1|x)$, by solving for β_0^* in equation (2).

If a previous case control study has been conducted, such as the Polish study of breast cancer (8), one can use previous estimates of log odds ratios to specify the effects of potential confounders and perhaps to get an estimate of the size of the exposure effect of interest. If the study is population-based, as was the Polish study, one also has an estimate of disease incidence. If the disease outcome is rare, the joint distributions of risk factors in unmatched controls approximates the distribution in the general population. If controls were matched to cases, for example on age, one needs to reweight the controls to the age distribution in the general population to estimate the distribution. F Once one has an empirical estimate $\hat{F}$ of the covariate distribution, with point masses at the unique values of the vector of covariates in controls, one can use the previous formulas and the formula for the information matrix in the Appendix, to compute sample size by replacing integrals involving F by weighted sums with weights $d\hat{F}$.

Another option is to assume a parametric model for F , such as a multivariate normal model, and to estimate its parameters from the controls in a case-control study or perhaps from a survey of covariates in the general population. Then the previous formulas can be used with parametric estimates of F replacing F .

Software that computes sample sizes either based on parametric models for F or based on empirical covariate distributions from unmatched controls or from a simple random sample from the general population is available in the R package, “samplesizelogisticcasecontrol” at CRAN.

Simulations to estimate power

With the same type of information required for analytical sample size calculations, one can use simulation to estimate the corresponding power for a given sample size, and hence to infer the required sample size. The simulation approach also offers the flexibility of allowing for other features, such as measurement error or missing data, that would be difficult to handle analytically in the multivariate setting. Our numerical example below concerns testing for an exposure effect, β_2 in the presence of a confounder X_1 . We also give an example that accommodates both confounding and non-differential recall error. For the first example, we assume we have an empirical distribution $\hat{F}(x_1, x_2)$ and parameter values (β_0^* , β_1 , β_2), but these methods easily extend to multiple covariates. The simulation proceeds as follows for each repetition $irep = 1, 2, \dots, IREP$.

- a. Create a large “source population” of N individuals by obtaining a random sample from $\hat{F}(x_1, x_2)$ of covariates (X_{1i}, X_{2i}) for $i = 1, 2, \dots, N$.
- b. From equation (1) with parameters (β_0^* , β_1 , β_2), for each $i = 1, 2, \dots, N$ compute $P(D_i = 1|X_{1i}, X_{2i})$ and generate a Bernoulli outcome $Y_i = 1$ with probability $P(D_i = 1|X_{1i}, X_{2i})$ and $Y_i = 0$ with probability $1 - P(D_i = 1|X_{1i}, X_{2i})$.

- c. Identify the N_1 diseased members of the population ($Y_i = 1$) and the $N_0 = N - N_1$ non-diseased members.
- d. Randomly sample n_1 cases from the N_1 diseased members in the population without replacement and randomly sample n_0 controls from the N_0 non-diseased members without replacement and “record” their confounder/exposure values, (X_{1i}, X_{2i}) .
- e. Use logistic regression on the cases and controls to test the null hypothesis $\beta_2 = 0$ using one of the four tests mentioned above, and record whether the null hypothesis is rejected.

Based on *IREP* repetitions of this process, the power is estimated as the number of times the null hypothesis was rejected divided by *IREP*. By repeating this for various choices of, $n = \lambda n + (1 - \lambda)n = n_1 + n_0$ one can generate power curves and use them to estimate the n needed to attain the desired power.

Software code for conducting such simulations in R is given in the Appendix. Furthermore, the algorithm for case-control studies outlined above is implemented in the “osDesign” package(6), available in CRAN, which also provides an analogous algorithm for two-phase designs(9).

RESULTS

In the following we present two examples based on the Polish breast cancer data. In both examples we take the primary goal to be estimation of power to detect a log odds ratio of $\beta_2 = 0.3$ (or odds ratio $\exp(\beta_2) = 1.35$) for the association of the presence of benign breast biopsy with breast cancer. In Example #1 we consider power calculations that do and do not account for including a confounder, specifically education, in the regression analysis. In Example #2, we consider power calculations that do and do not account for analytic approaches to correct for non-differential recall error with regard to the presence of previous benign breast biopsy, as well as for education as a confounder.

Example #1: Adjustment for confounding

In the observed data, the prevalence of benign breast biopsy in the Polish controls is 0.06475. Assuming equal numbers of cases and controls ($\lambda = 0.5$), $Z_{1-\alpha} = 1.96$ (for two-sided 0.05 level test), and $Z_\gamma = 1.282$ (for power $\gamma = 0.9$), the analytic sample size calculations presented above yield sample sizes n (cases plus controls) of 6,790 for the Wald test with equation (5), 6,863 for the Wald test with equation (6) and 6,775 and 6,702 for the corresponding score tests. These results can be obtained from the R program `samplesizelogisticcasecontrol` using the commands in the Appendix.

To compute the sample size adjusted for education level, a strong risk factor ($\beta_1 = 0.126$ from Table 1) with correlation 0.135 with benign breast biopsy, we read in the Polish case-control data, called `data`, with columns containing respectively the index of the entry, case-control status (1 or 0), benign breast biopsy status (0 or 1), and education level (0, 1, ..., 7). We created a separate data frame called “`controldata`” containing only rows of data

corresponding to controls. Using the R code in the Appendix, we found the respective confounder-adjusted sample sizes to be 6,742 and 6,810 for Wald tests and 6,705 and 6,602 for score tests. Thus, adjustment for this positively correlated confounder yielded very slightly decreased sample size requirements compared to no adjustment.

Figure 1 presents simulation-based estimates of power of the Wald test corresponding to equation (6) against $\beta_2 = 0.3$ as a function of sample size n , for the same model parameters used in the previous calculations. Simulation parameters were $N=1,000,000$ and $IREP=10,000$. From Figure 1, we find that the sample size required to achieve 90% power is approximately 6,900 without adjustment (solid curve) and 6,700 with adjustment for confounding (dashed curve). This small impact on sample size from confounder adjustment may be due to education's relatively weak association with the exposure of interest.

Example #2: Adjustment for misclassification of exposure

A common challenge in interpreting case-control studies is the potential for errors in reporting exposures due to the retrospective design. Particularly troublesome are error distributions that differ by disease status, resulting in "recall bias"(10). If recall bias is suspected, investigators may use any of a number of methods to adjust for measurement error. For example, Lyles and Lin(11) proposed using the observed data and investigator-supplied values for sensitivity and specificity to estimate the negative and positive predictive values for the true exposure (benign breast biopsy status here). These predictive values are then used as weights in fitting the model of interest to obtain unbiased estimates of odds ratios, with standard error estimates based on the jackknife. If one naively ignores measurement error at the design stage, the required sample size may be seriously underestimated. Here we use simulations to compare sample sizes estimated for a "naïve" analysis with no adjustments for misclassification of benign breast biopsy status, but allowing for adjustment on a confounder, education, with "realistic" sample sizes that account for both these factors. The measurement error is non-differential (the same distribution in cases and controls) so that there is no recall bias and tests of the null hypothesis of no association of disease status with benign breast biopsy are valid.

Building on the set-up for Example #1, we present simulation estimates of power of the Wald test corresponding to equation (6) to detect $\beta_2 = 0.3$ (Figure 2) based on: (i) a "naïve" calculation that assumes benign breast biopsy will be measured without error, but allow for a confounder, education, and (ii) a "realistic" calculation that allows for a confounder (education) and allows for an adjustment for misclassification error following Lyles and Lin (11). We assumed that 90% of those who truly have benign breast disease are correctly classified (i.e. 90% sensitivity), regardless of disease status, and that 90% of those who truly did not have benign breast disease are correctly classified (i.e. 90% specificity), regardless of disease status. Hence, the recall misclassification is "non-differential" with respect to outcome status.

From the solid line in Figure 2, we see that, consistent with the "realistic" results of Example #1, the naïve calculation suggests that just under 7,000 cases plus control would be needed to ensure power of at least 90%. From the dashed line, however, we find that a design in which 7,000 cases plus controls are sampled would only have 74% power for an

analysis (11) that adjusts for measurement error. Extrapolating the estimated power curve, we found that the required sample size for measurement error adjustment was more than 9,500.

DISCUSSION

We have described methods and software for sample size calculations to test for a scalar exposure effect or scalar interaction with unconditional logistic regression analysis of case-control data with an arbitrary ratio of cases to controls, $\lambda/(1 - \lambda)$. Analytic methods in the R program `samplesizelogisticcasecontrol` in CRAN can be used to adjust for an arbitrary number of confounders, whether testing for an exposure effect or interaction. Simulation code is also described in the Appendix. Alternative more general simulation methods were published previously(6).

Simulations have the advantage that special features such as measurement error or missing data, that are hard to address analytically in the multivariate setting, can be handled, provided one can model those special features in the simulation, as illustrated in Example #2 and in references (6, 7). Thus, the simulation method is more flexible and frees the investigator to try to define the particular characteristics of a study that need special treatment. The challenge is to understand the subtleties of a study well enough to simulate the data and to propose appropriate analytical procedures whose operating characteristics are to be studied in the simulations.

The main challenge to using these multivariate methods is specifying the log odds ratios for the various covariates and the joint distribution of covariates in the source population. Sometimes such information is available from a previous study, as illustrated by our use of the Polish data (8). Sometimes, initial data from the planned study can be used to refine sample sizes for the entire study(7). Sometimes one may need to posit various joint distributions and log odds ratios and examine the various sample size results in an analysis to see how sensitive the sample size calculations are to various assumptions. Here we show that analytical and simulation methods are available to refine sample size sample size calculations, whatever the source of the information on covariate effects and joint distributions.

Acknowledgments

Dr. Gail was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute. Dr. Haneuse was supported by the National Institutes of Health grant R-01 HL094786.

We thank Dr. William Wheeler at Information Management Services for developing the R package, `samplesizelogisticcasecontrol`, Dr. Kyu Ha Lee at the Forsyth Institute (Boston, MA) for help with simulations, Jonathan Bartlett for permission to adapt his code to produce the R function `pvw()` in the Appendix, and Dr. Montserrat Garcia-Closis at the Division of Cancer Epidemiology and Genetics, National Cancer Institute, for access to data from the case-control study of breast cancer in Poland.

APPENDIX

Calculation of the information matrix

The information matrix is the (retrospective) expectation of minus the matrix of cross-derivatives of the log-likelihood obtained by pretending that the case-control data were cohort data. For example the expectation of the term corresponding to the second derivative with respect to β_2 is

$$n\{\lambda \int x_2^2 P(D=1|\mathbf{x}_1, x_2) P(D=0|\mathbf{x}_1, x_2) dF_1(\mathbf{x}_1, x_2) + (1-\lambda) \int x_2^2 P(D=1|\mathbf{x}_1, x_2) P(D=0|\mathbf{x}_1, x_2) dF_0(\mathbf{x}_1, x_2)\}.$$

Other terms are computed similarly except that x_2^2 is replaced by quantities like 1, x_1 , x_2 , $x_1 x_2$, and x_1^2 , depending on the cross-derivative. Under the null hypothesis $\beta_2 = 0$, the other parameters β_0 and β_1 are the values to which the estimators converge when β_2 is set to zero. In the special case of only one exposure variable and no confounders (i.e. no $\mathbf{x}_2$), the expression above reduces to $n\lambda(1-\lambda)\{\lambda \int x_2^2 dF_1(\mathbf{x}_1, x_2) + (1-\lambda) \int x_2^2 dF_0(\mathbf{x}_1, x_2)\}.$

Commands to compute sample sizes in Example #1 with the R program

“samplesizelogisticcasecontrol”

The program “samplesizelogisticcasecontrol” is available in CRAN. The results for Example #1 without adjustment for a confounder were obtained with the R commands (separated by semicolons): `prev <- 0.01; logOR <- 0.3; p <- 0.06475; sampleSize_binary(prev, logOR, probXeq1=p).`

To adjust for covariates in Example #1, we created a separate data frame called “controldata” by using only rows of data corresponding to controls. Then we used the R code: `prev<- 0.01; logOR <- c(0.126, 0.3); sampleSize_data(prev, logOR,controldata).`

R code for simulations

Here we provide the code used for the simulations in Example #1 and Example #2. Since data from the Polish breast cancer study cannot be provided, the code will not run but can be easily modified for specific applications. Before providing the code for the simulation examples, we present three needed functions. The first function, `expit()`, is the inverse of the logit function, useful for computing probabilities from a logistic regression model. The second function, `eval.beta0()`, computes the absolute difference between the mean outcome probability for a given dataset and values for the regression coefficients and some user-specified overall prevalence for the outcome (denoted π_Y). This function is used to calculate the value of the intercept in a logistic regression model that corresponds to a particular dataset and overall prevalence. The final function, `pvw()`, implements the predictive value weighting algorithm of Lyles and Lin(11). The function we provide below was adaption from one provided with permission from the author, Jonathan Bartlett, at:

<http://thestatsgeek.com/2014/05/08/adjusting-for-covariate-misclassification-in-logistic-regression-predictive-value-weighting/>

Following code for these functions, we provide the code for example #1 and example #2 in successive sections.

```
#####
#####
## Functions that are used in the simulations in Examples #1 and #2
#####
#####
##
expit <- function(x) exp(x)/(1 + exp(x))
##
eval.beta0 <- function(beta0, betaX, Xmat, piY) abs(piY -
mean(expit(as.numeric(Xmat %*% c(beta0, betaX)))))
##
pvw <- function(data, indices, sens, spec)
{
  localData <- data[indices,]
  n <- nrow(localData)
  zyc_mod <- glm(benignMiss ~ Y + educat + Y:educat, family="binomial",
data=localData)
  localData$piycstar <- predict(zyc_mod, type=c("response"))
  localData$se <- rep(sens, n)
  localData$se[localData$piycstar>=localData$se] <- localData
$piycstar[localData$piycstar>=localData$se]+0.001
  localData$sp <- rep(spec, n)
  localData$sp[localData$piycstar<=(1-localData$sp)] <- 1-localData
$piycstar[localData$piycstar<=(1-
localData$sp)]+0.001
  localData$term1 <- (localData$se-1)*localData$piycstar*(localData
$se*(localData$piycstar-1))^(1)
  localData$term2 <- (localData$sp-1)*(localData$piycstar-1)*(localData
$sp*localData$piycstar)^(1)
  localData$det <- 1/(localData$term1*localData$term2-1)
  localData$ppv <- localData$det*(localData$term2-1)
  localData$npv <- localData$det*(localData$term1-1)
  expandedData <- rbind(localData, localData)
  expandedData$benignNew <- 0
  expandedData$benignNew[(n+1):(2*n)] <- 1
  expandedData$wgt <- 0
  expandedData$wgt[expandedData$benignNew==1 & expandedData$benignMiss==1] <-
expandedData$ppv[expandedData$benignNew==1 & expandedData$benignMiss==1]
```

```

    expandedData$wgt[expandedData$benignNew==0 & expandedData$benignMiss==1] <
- 1-
expandedData$ppv[expandedData$benignNew==0 & expandedData$benignMiss==1]
    expandedData$wgt[expandedData$benignNew==1 & expandedData$benignMiss==0] <
- 1-
expandedData$npv[expandedData$benignNew==1 & expandedData$benignMiss==0]
    expandedData$wgt[expandedData$benignNew==0 & expandedData$benignMiss==0] <-
expandedData$npv[expandedData$benignNew==0 & expandedData$benignMiss==0]
    wgtMod <- glm(Y ~ benignNew, family="quasibinomial", weights=wgt,
data=expandedData)
return(coef(wgtMod))
}
#####
##
## Code for running the simulation in Example #1
#####
##
##
Npop <- 1e6
piY <- 0.01
## Create a simulation population of size Npop from the controls in the
Polish data
##
Xmat.F <- cbind(1, poland[poland$Y == 0,])
expand.popn <- sample(1:nrow(Xmat.F), Npop, replace=TRUE)
simData <- as.data.frame(Xmat.F[expand.popn,-c(1,9)])
## Determine the value of the intercept given:
## - the design matrix based on the 'true' benign breast disease status
## - an effect of 0.3 for benign breast disease
## - an overall prevalence of piY
##
Xmat.S.1 <- cbind(1, simData$benign)
Xmat.S.2 <- cbind(1, simData$benign, simData$educat)
##
betaX.S.1 <- c(0.3)
betaX.S.2 <- c(0.3, 0.126)
##
beta0.S.1 <- optimize(eval.beta0, interval=c(-5, 0), betaX.S.1, Xmat.S.1,
piY)$minimum
beta0.S.2 <- optimize(eval.beta0, interval=c(-5, 0), betaX.S.2, Xmat.S.2,
piY)$minimum
## True data generating probabilities
##
piY.S.1 <- expit(as.vector(Xmat.S.1 %*% c(beta0.S.1, betaX.S.1)))

```

```

piY.S.2 <- expit(as.vector(Xmat.S.2 %*% c(beta0.S.2, betaX.S.2)))
##
results <- array(NA, c(length(nCC), 2, 2, R))
for(r in 1:R)
{
  ##
  simData$Y.S.1 <- rbinom(Npop, 1, piY.S.1)
  simData$Y.S.2 <- rbinom(Npop, 1, piY.S.2)
  ##
  for(j in 1:length(nCC))
  {
    ##
    simData$Y <- simData$Y.S.1
    keep.Y0 <- sample(c(1:Npop)[simData$Y == 0], nCC[j]/2)
    keep.Y1 <- sample(c(1:Npop)[simData$Y == 1], nCC[j]/2)
    simDataCC <- simData[c(keep.Y0, keep.Y1),]
    fitCC <- summary(glm(Y ~ benign, data=simDataCC, family=binomial))$coef
    results[j,1,r] <- c(fitCC[2,1], as.numeric(fitCC[2,4] < 0.05))
    ##
    simData$Y <- simData$Y.S.2
    keep.Y0 <- sample(c(1:Npop)[simData$Y == 0], nCC[j]/2)
    keep.Y1 <- sample(c(1:Npop)[simData$Y == 1], nCC[j]/2)
    simDataCC <- simData[c(keep.Y0, keep.Y1),]
    fitCC <- summary(glm(Y ~ benign + educat, data=simDataCC,
family=binomial))$coef
    results[j,2,r] <- c(fitCC[2,1], as.numeric(fitCC[2,4] < 0.05))
  }
  ##
  if((r %% 10) == 0)
  {
    print(c(R, r))
    save("results", file=paste("Output/Sim-21/21_SimOutput-", in1, ".dat",
sep=""))
  }
}

#####
#####
## Code for running the simulation in Example #2
#####
#####
##
Npop <- 1e6
piY <- 0.01
## Create a simulation population of size Npop from the controls in the

```

```

Polish data
##
Xmat.F <- cbind(1, poland[poland$Y == 0,])
expand.popn <- sample(1:nrow(Xmat.F), Npop, replace=TRUE)
simData <- as.data.frame(Xmat.F[expand.popn,-c(1,9)])
## Simulate the misclassified benign breast disease variable
## - currently 90% sensitive and 90% specific
##
simData$benignMiss <- NA
simData$benignMiss[simData$benign == 0] <- rbinom(sum(simData$benign == 0),
1, 0.1)
simData$benignMiss[simData$benign == 1] <- rbinom(sum(simData$benign == 1),
1, 0.9)
## Determine the value of the intercept given:
## - the design matrix based on the 'true' benign breast disease status
## - an effect of 0.3 for benign breast disease
## - an overall prevalence of piY
##
Xmat.S <- cbind(1, simData$benign)
betaX.S <- 0.3
beta0.S <- optimize(eval.beta0, interval=c(-5, 0), betaX.S, Xmat.S, piY)
$minimum
## True data generating probabilities
##
piY.S <- expit(as.vector(Xmat.S %*% c(beta0.S, betaX.S)))
## Run the simulation
R <- 10000
nCC <- seq(from=4000, to=7000, by=1000)
results <- array(NA, c(length(nCC), 3, 3, R))
for(r in 1:R)
{
  ##
  print(c(R, r))
  ##
  simData$Y.S <- rbinom(Npop, 1, piY.S)
  ##
  for(j in 1:length(nCC))
  {
    ##
    simData$Y <- simData$Y.S
    keep.Y0 <- sample(c(1:Npop)[simData$Y == 0], nCC[j]/2)
    keep.Y1 <- sample(c(1:Npop)[simData$Y == 1], nCC[j]/2)
    simDataCC <- simData[c(keep.Y0, keep.Y1),]
    ##

```

```

fitCC      <- summary(glm(Y ~ benign, data=simDataCC, family=binomial))$coef
results[j,1,,r] <- c(fitCC[2,1:2], as.numeric(fitCC[2,4] < 0.05))
##
fitCC      <- summary(glm(Y ~ benignMiss, data=simDataCC, family=binomial))
$coef
results[j,2,,r] <- c(fitCC[2,1:2], as.numeric(fitCC[2,4] < 0.05))
##
temp.Est <- pvw(simDataCC, c(1:nCC[j]), 0.9, 0.9)
temp.SE  <- matrix(NA, nrow=nCC[j], ncol=length(temp.Est))
for(k in 1:nCC[j])
{
  if((k %% 10) == 0) print(c(R, r, nCC[j], k))
  temp.SE[k,] <- pvw(simDataCC, c(1:nCC[j])[-k], 0.9, 0.9)
}
temp.SE <- apply(temp.SE, 2, FUN=sd) * sqrt(nCC[j]-1)/sqrt(nCC[j]) *
sqrt(nCC[j]-1)
results[j,3,,r] <- c(temp.Est[2], temp.SE[2], as.numeric(abs(temp.Est/
temp.SE)[2] > qnorm(0.975)))
}
}

```

References

1. Breslow, NE, Day, NE. Statistical Methods in Cancer research, Volume II: The Design and Analysis of Cohort Studies. Lyon: International Agency for Research on Cancer; 1987.
2. Whittemore AS. Sample size for logistic-regression with small response probability. Journal of the American Statistical Association. 1981; 76(373):27–32.
3. Demidenko E. Sample size determination for logistic regression revisited. Stat Med. 2007; 26(18): 3385–97. [PubMed: 17149799]
4. McKeown-Eyssen GE, Thomas DC. Sample-size determination in case control studies - the influence of the distribution of exposure. J Chronic Dis. 1985; 38(7):559–68. [PubMed: 4008598]
5. Lubin JH, Gail MH. On power and sample-size for studying features of the relative odds of disease. Am J Epidemiol. 1990; 131(3):552–66. [PubMed: 2301364]
6. 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(11):1–29.
7. Haneuse S, Schildcrout J, Gillen D. A two-stage strategy to accommodate general patterns of confounding in the design of observational studies. Biostatistics. 2012; 13(2):274–88. [PubMed: 22130627]
8. Garcia-Closas M, Brinton LA, Lissowska J, et al. Established breast cancer risk factors by clinically important tumour characteristics. Br J Cancer. 2006; 95(1):123–9. [PubMed: 16755295]
9. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. Journal of the Royal Statistical Society Series C-Applied Statistics. 1999; 48:457–68.
10. Coughlin SS. RECALL BIAS IN EPIDEMIOLOGIC STUDIES - REPLY. J Clin Epidemiol. 1990; 43(12)
11. Lyles RH, Lin J. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. Stat Med. 2010; 29(22):2297–309. [PubMed: 20552681]

Figure 1.

Simulation-based estimated power curves with confounding, as a function of sample size n , for detecting $\beta_2 = 0.3$, the log odds of breast cancer associated with presence of benign breast disease, in a hypothetical case-control study with equal numbers ($n/2$) of cases and controls. The solid curve (“naïve”) is for a model that includes benign breast disease only; the dashed curve (“realistic”) is for a model that also includes education level, a potential confounder that is associated with benign breast disease and outcome.

Figure 2.

Simulation-based estimated power curves with measurement error and confounding, as a function of sample size n , for detecting $\beta_2 = 0.3$, the log odds of breast cancer associated with presence of benign breast disease, in a hypothetical case-control study with equal numbers ($n/2$) of cases and controls. The solid curve (“naïve”) is for a simulation study that adjusts for the confounder, education, but naively ignores any eventual need to control for non-differential recall error regarding benign breast biopsy status. The dashed curve (“realistic”) is for a simulation study that accounts for the confounder and for non-differential recall error using the method of Lyles and Lin(6, 11).

Table 1

Means and standard deviations (SDs) of risk factor levels in cases and controls in the Polish breast cancer study, together with log odds ratio (OR) estimates, standard error (SE) estimates, and p-values from separate univariate logistic regressions.

Risk Factor	Cases n ₁ =1,968		Controls n ₀ =2,162		Estimated Log OR (SE)	Two-Sided P Value
	Mean	SD	Mean	SD		
Age(yrs) ^a	55.7	9.94	55.9	9.98	-0.020(0.031)	0.519
Education (0-7)	3.4	1.76	2.9	1.60	0.169(0.019)	<0.001
Age 1 st birth(yrs)	24.5	4.61	23.6	4.22	0.220(0.036)	<0.001
Age menarche (yrs) ^b	13.5	1.71	13.7	1.72	-0.061(0.018)	0.001
Benign biopsy (0,1)	0.10	0.30	0.06	0.25	0.485(0.115)	<0.001
Family history (0,1)	0.11	0.31	0.06	0.23	0.678(0.118)	<0.001
BMI (kg/m ²) ^b	27.4	5.38	28.0	5.33	-0.116(0.029)	<0.001

^alog odds ratio for a 10-year contrast

^blog odds ratio for a 5-year contrast

Table 2

Variances, covariances (above the diagonal) and correlations (below the diagonal) for the cases and the controls in the Polish breast cancer data.

Cases	Age	Education	Age first birth	Age menarche	Benign biopsy	Family history	BMI
Age	98.5	-2.53	-2.75	3.22	-0.15	0.12	16.1
Education	-0.15	3.09	2.61	-0.31	0.05	0.01	-1.50
Age first birth	-0.06	0.32	21.3	0.22	0.06	0.00	-1.85
Age menarche	0.19	-0.10	0.03	2.91	0.00	0.01	-0.17
Benign biopsy	-0.05	0.09	0.04	0.00	0.09	0.01	-0.14
Family history	0.04	0.02	0.00	0.02	0.06	0.10	0.02
BMI	0.30	-0.16	-0.08	-0.02	-0.84	0.01	29.0
Controls	Age	Education	Age first birth	Age menarche	Benign biopsy	Family history	BMI
Age	99.6	-3.01	-1.81	1.91	-0.03	0.11	11.5
Education	-0.19	2.56	2.18	-0.32	0.05	0.02	-1.46
Age first birth	-0.04	0.32	17.8	0.44	0.05	0.01	-3.12
Age menarche	0.11	-0.12	0.06	2.95	-0.02	0.00	-0.32
Benign biopsy	-0.01	0.14	0.05	-0.04	0.06	0.01	-0.04
Family history	0.05	0.06	0.01	0.07	0.08	0.05	0.05
BMI	0.22	-0.17	-0.14	-0.03	-0.03	0.04	28.4