International Journal of Epidemiology, 2015, 1660–1672
doi: 10.1093/ije/dyv137
Advance Access Publication Date: 30 July 2015
Original article

Methodology

Standardized binomial models for risk or prevalence ratios and differences

David B Richardson,[1]* Alan C Kinlaw,[1] Richard F MacLehose[2] and Stephen R Cole[1]

[1]Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA and [2]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, USA

*Corresponding author. Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA 27599. E-mail: david.richardson@unc.edu

Accepted 12 June 2015

Abstract

Background: Epidemiologists often analyse binary outcomes in cohort and cross-sectional studies using multivariable logistic regression models, yielding estimates of adjusted odds ratios. It is widely known that the odds ratio closely approximates the risk or prevalence ratio when the outcome is rare, and it does not do so when the outcome is common. Consequently, investigators may decide to directly estimate the risk or prevalence ratio using a log binomial regression model.

Methods: We describe the use of a marginal structural binomial regression model to estimate standardized risk or prevalence ratios and differences. We illustrate the proposed approach using data from a cohort study of coronary heart disease status in Evans County, Georgia, USA.

Results: The approach reduces problems with model convergence typical of log binomial regression by shifting all explanatory variables except the exposures of primary interest from the linear predictor of the outcome regression model to a model for the standardization weights. The approach also facilitates evaluation of departures from additivity in the joint effects of two exposures.

Conclusions: Epidemiologists should consider reporting standardized risk or prevalence ratios and differences in cohort and cross-sectional studies. These are readily-obtained using the SAS, Stata and R statistical software packages. The proposed approach estimates the exposure effect in the total population.

Key words: Risk, prevalence, standardizations, regression models

Key Messages

- Standardized risk or prevalence ratios and differences are readily obtained using the SAS, Stata and R statistical software packages.
- These quantities may be obtained by fitting a marginal structural binomial regression model.
- This approach reduces problems with non-convergence in covariate-adjusted log binomial models.
- The proposed approach estimates an exposure's effect in the total study population, though other target populations are also accommodated.

Background

Logistic regression models are commonly used by epidemiologists to analyse binary outcome data from cohort and cross-sectional studies.[1] For rare events, the odds ratio from such a logistic regression model approximates the risk ratio well and is commonly used to do so. However, when logistic regression is used to model common events, the estimated odds ratio is not close to the risk ratio, and will be further from the null. Because the risk and odds ratios are based on the same quantities (i.e. the probabilities of the outcome among the exposed and unexposed), it might seem reasonable to conclude that one effect measure can be readily derived from the other. A simple algebraic conversion was described by, among others, Zhang and Yu (1998) for this purpose.[2]

However, when an adjusted odds ratio is estimated by a multivariable logistic regression model, the simple algebraic conversion proposed by Zhang and Yu will not yield the adjusted risk ratio one might calculate, for example, by classical Mantel-Haenszel methods.[2,3] The reason for this difference is that a summary adjusted odds ratio for an exposure contrast is obtained under the assumption of a constant odds ratio across strata of the covariates that are confounders. Unfortunately, if the odds ratio is constant over the strata defined by confounders, then the risk ratio will vary over these strata.[4]

Given these difficulties, many authors have advocated for direct modelling of the risk or prevalence ratio by fitting a multivariable log binomial regression model.[5–8] Of course if the underlying population risk model actually conforms to the logistic model, direct modelling of the risk or prevalence ratio will involve incorporating interaction terms in the model or ignoring potentially important effect measure modification by strata of covariates. Moreover, unlike logistic regression, multivariable log binomial regression models often suffer convergence problems.[8]

In the current paper we focus on the setting in which an investigator wishes to compare the risk or prevalence of disease between exposure groups, obtaining a summary measure that is standardized to the confounder distribution in the overall population. We describe the use of a standardized log binomial model to estimate risk and prevalence ratios in this setting. By employing standardization to control confounding rather than a model for the disease conditional on exposure and potential confounders, investigators will tend to experience fewer problems with model fit and convergence than those often encountered with multivariable log binomial regression. Moreover, when the effect measure (e.g. the risk ratio) varies across strata of covariates, this standardization approach yields a useful, standardized summary effect measure. We describe and illustrate this approach to estimate standardized risk and prevalence ratios. We further illustrate how this approach can facilitate evaluation of risk and prevalence differences using linear binomial regression, and assessment of whether the joint effects of explanatory variables conform to a linear model.

Methods

Consider a study in which D denotes disease status and is a binary outcome variable and E denotes exposure and is the explanatory variable of primary interest. Let $\boldsymbol{Z}$ denote a vector of explanatory variables that are potential confounders of the association between E and D; $\boldsymbol{Z}$ may include binary, categorical and continuous variables. The investigator wants to compare the risk or prevalence of D between groups defined by E, obtaining a summary measure that is standardized to the confounder distribution in the overall population.

Standardized binomial regression models

Direct standardization is used by epidemiologists as one method to control for confounding in analyses of exposure-disease associations. The observed data are weighted so that the exposure groups under comparison have the same distribution of potential confounding factors in the weighted data. Because standardization requires stratification by confounders, covariates that were originally measured on a continuous scale must be categorized for the purposes of standardization, and problems with instability of estimation may occur when the data are stratified by many confounders.[9]

Robins *et al.* proposed a flexible extension of classical standardization methods.[10] Under this approach, a weight is assigned to each person that is equal to the inverse probability of receiving their own exposure conditional on their confounder information. To estimate these weights in a parametric setting, the investigator fits a regression model for E as a function of the covariates, $\boldsymbol{Z}$.

If the exposure variable of primary interest is binary, then the weights needed for standardization may be estimated from a logistic regression model for E with $\boldsymbol{Z}$ as explanatory variables. The logistic model yields predicted probabilities of exposure as a function of covariates $\boldsymbol{Z}$. The weight assigned to each individual equals the inverse of the predicted probability that the person had the exposure level of $E = e$ that was in fact observed. In practice, weights often are stabilized by multiplying each weight by the overall unadjusted probability of that person having exposure level $E = e$.[10,11]

If E includes more than two levels, then a multinomial regression model for E may be fitted.[12] The model yields predicted probabilities for each level of E as a function of covariates $\boldsymbol{Z}$; again, the weight assigned to each individual equals the inverse of the predicted probability that the person had the exposure level of E that was in fact observed. In settings where the exposure categories are ordered, one may wish to employ an ordered logistic regression model to estimate the weights.[13] If the exposure variable of interest is continuous then, for practicality and simplicity, we focus on the setting in which the exposure variable is categorized into quantiles (e.g. deciles); however, weights could be formed using the density estimates returned from a linear regression model.[12] The weights needed for standardization may be estimated from a multinomial or ordered logistic regression model for this categorized version of E with $\boldsymbol{Z}$ as explanatory variables.[8] The model yields predicted probabilities for each quantile of E as a function of covariates $\boldsymbol{Z}$; and, the weight assigned to each person equals the inverse of the predicted probability that the person had the exposure quantile of E that was in fact observed. When the continuous exposure E is categorized, we suggest using the same exposure categories in the binomial regression model for the outcome. Use of a different, more finely categorized exposure variable in the binomial regression model may reintroduce within-category confounding, reflecting residual bias not accounted for during the weight construction.

Estimation of a summary standardized effect measure

As in classical standardization, a summary effect measure is obtained after weighting the observed data such that the association between E and D in the weighted data is unconfounded by covariates $\boldsymbol{Z}$. This will yield the standardized risk or prevalence at each level of E. This approach is essentially a marginal structural binomial model for the effect of a point treatment. Relative and absolute effect measures (e.g. risk ratios and risk differences) may be obtained for contrasts defined by E, by means of a simple tabular analysis of the occurrence of D within levels of E in these weighted data or by means of a regression model that incorporates probability weights.

Because the standardized binomial regression model is a weighted m-estimator,[14] robust (Huber-White) or bootstrap confidence intervals (CIs) are recommended.[10] Figure 1 provides illustrative code for SAS, Stata and R to estimate stabilized weights for a binary exposure variable and then obtain an estimate (and associated 95% robust confidence interval) for the standardized relative risk or standardized risk difference. Suppose that a data set named *one* includes a unique identifier for each person (*id*), a binary indicator of disease status (D), a binary exposure variable (E) and covariates *Z1* and *Z2*. The code in Figure 1 estimates the predicted probability of exposure ($E = e$) given covariates *Z1* and *Z2* by fitting a logistic regression model, creating a new data set (*two*) that includes all of the observed data plus the predicted probability of exposure given covariates. Next, the marginal probability of exposure is estimated by fitting a logistic regression model with no explanatory variables, which serves as the numerator for the stabilized weight (*sw*), calculated in a new data set (*three*). Robust confidence intervals can be obtained simply by fitting a log binomial regression model for D with E as the only explanatory variable to the weighted data to estimate the risk ratio (or a linear binomial regression model to estimate the risk difference), using the *repeated* statement in SAS PROC GENMOD, the robust option in Stata or the *id* argument in the *geeglm* function in R. Percentile bootstrap confidence intervals can be calculated using the code provided in Appendix 1 (available as Supplementary data at *IJE* online).

Using the weighting approach described above, the total study population (exposed and unexposed) serves as the standard; this means that, in the weighted data, the exposed and unexposed each have the distribution of covariates, $\boldsymbol{Z}$, observed in the total group. The resultant standardized effect measure is interpretable as a contrast of the risk or prevalence of disease in the total group under complete exposure and complete non-exposure. If there is heterogeneity of the effect measure over strata of covariates, the resultant standardized summary measure (the difference or ratio of standardized risks or standardized prevalences) retains its intended meaning: it is an average of the effect of the exposure in a population with distribution of

SAS CODE

```
/*Calculate denominators used in inverse probability weights  */
        proc logistic data = one descending;
                model E = Z1 Z2;
                output out = two predicted=ps;

/*Create stabilized weights, using a null model with E as the dependent variable. */
        proc logistic data = two descending;
                model E =  ;
                output out=three predicted=marg_pr;
        data three;
                set three;
                sw = E*marg_pr/ps + (1-E)*(1-marg_pr)/(1-ps);

/*Fit a log binomial model to the weighted data for the E-D association, with robust variance */
        proc genmod data = three descending;
                class id;
                model D = E / link=log dist=bin ;
                weight sw;
                repeated subject=id / type=ind;
                estimate 'rr' E 1 / exp; run;


/*Fit a linear binomial model to weighted data for the E-D association with robust variance */
        proc genmod data = three descending;
                class id;
                model D = E / link=identity dist=bin ;
                weight sw;
                repeated subject=id / type=ind; run;
```

STATA CODE

```
*Calculate denominators used in inverse probability weights
        logit E Z1 Z2
        predict ps

*Create stabilized weights, using a null model with E as the dependent variable
        logit E
        predict marg  pr
        g sw=E*marg_pr/ps+(1-E)*(1-marg_pr)/(1-ps)

*Fit a log binomial model to the weighted data for the E-D association, with robust variance
        glm D E [pw=sw],family(binomial) link(log) robust

*Fit a linear binomial model to the weighted data for the E-D association, with robust variance
        glm D E [pw=sw],family(binomial) link(identity) robust
```

R CODE

```
# Calculate denominators used in inverse probability weights
        E.out=glm(E~Z1+Z2,family=binomial(link="logit"), data=one, na.action=na.exclude)
        ps=predict(E.out, type="response")

# Create stabilized weights, using a null model with E as the dependent variable
        sptw=one$E*mean(one$E)/ps+(1-one$E)*(1-mean(one$E))/(1-ps)

# Fit a log binomial model to the weighted data for the E-D association, with robust variance
# If 'geepack' package not installed, enter in console: install.packages('geepack') library(geepack)
        summary(geeglm(D~E, family=binomial(link="log"), weight=sptw, id=id, data=one))


# Fit a linear binomial model to the weighted data for the E-D association, with robust variance
        library(geepack)
        summary(geeglm(D~E, family=binomial(link="identity"), weight=sptw, id=id, data=one))
```

Figure 1. Sample code to estimate standardized relative risks and standardized risk differences.

covariates, **Z**, that was seen in this study. Alternatively, one could choose as the standard the exposed or unexposed group rather than the total study population.[9] Also, one could make comparisons other than that of everyone exposed vs no one exposed.[15,16]

Joint Effects of Two Exposures

Suppose that there are two exposure variables of interest, *E* and *F*. Assessment of whether there is a departure from additivity of effects may be of interest, for example to evaluate whether an intervention on *E* would have a larger absolute effect jointly with exposure *F* than without exposure *F*. A number of authors have considered assessment of additive interaction in analysis of binary outcome data using models for the log-odds of disease.[17–21] Such approaches require the assumption that the odds ratio approximates the risk or prevalence ratio, and that the relative excess risk due to interaction does not vary substantially across strata of covariates.[22,23] Standardized risk or prevalence can readily be compared in terms of risk or prevalence differences; we propose a novel approach for assessing additivity by fitting a weighted linear binomial regression model to assess whether the joint effects of explanatory variables conform to a linear model.

One set of weights may be derived from a logistic regression model for *E* with **Z** as explanatory variables. A second set of weights can be derived from a logistic regression model for *F* with *E* and **Z** as explanatory variables. The vector of covariates **Z** in the regression model for *E* need not be identical to the vector of covariates **Z** in the regression model for *F*. It is important to note that when estimating the exposure prediction model for the second exposure *F*, the first exposure *E* must be included to properly recover the joint distribution of these two exposures, and the model for *F* may include product terms between *E* and covariates in vector **Z**. Moreover, if *F* is affected by *E*, then the ordering of models should follow this temporal relationship. A final weight for the individual is the product of the two weights (Appendix 2, available as Supplementary data at *IJE* online). Alternatively, one could create a categorical variable for all combinations of *F* and *E*. A multinomial model fit with the dummy variable as an outcome and **Z** as predictors could be used to estimate weights in a flexible manner.

Using these weighted data the analyst can estimate the risk or prevalence of disease (denoted by *R*) among those jointly unexposed, $R(\overline{E}\overline{F})$, exposed only to *E*, $R(E\overline{F})$, exposed only to *F*, $R(\overline{E}F)$, or jointly exposed to *E* and *F*, $R(EF)$. Departures from additive effects of *E* and *F* may be described in terms of the absolute excess risk due to interaction $= R(EF) - R(E\overline{F}) - R(\overline{E}F) + R(\overline{E}\overline{F})$.

Simulation example

Data were simulated for 1000 cohort studies, with 10 000 people in each cohort. Each simulated cohort had three explanatory variables: Z_1, Z_2 and Z_3. We assigned Z_1 as a random polytomous variable sampled from a multinomial distribution that took the values 1, 2, 3 with probabilities 0.5, 0.25 and 0.25. We assigned Z_2 as a random binary variable that took a value of 1 with probability $\exp(-1 - 1*Z_1)/(1+\exp(-1 - 1*Z_1))$, else 0; in each simulated cohort, the probability that Z_2 took a value of 1 was approximately 0.08. We assigned Z_3 as a random binary variable that took a value of 1 with probability $\exp(-0.1 - 1*Z_1 - 1*Z_2)/(1 + \exp(-0.1 - 1*Z_1 - 1*Z_2))$, else 0. We assigned the outcome *D* as a random binary variable that followed a log binomial distribution and took a value of 1 with probability $\exp(-0.1 - 1*Z_1 - 1*Z_2 - 1*Z_3)$. Z_3 is the exposure of interest in this scenario; in each simulated cohort the probability that Z_3 took a value of 1 was approximately 0.15. A correctly specified logistic model was fit to each simulated cohort to predict Z_3 as a function of Z_1 and Z_2; the inverse of the predicted probabilities served as the basis for weights in a marginal structural model. We stabilized these weights by multiplying them by the marginal probability of each person's observed exposure level. We estimated the effect of Z_3 on the outcome using a marginal structural log binomial model, with robust variance. From 1000 replications of the study, we computed the mean log risk ratio ('estimated log RR'), empirical standard error of the estimated log risk ratio and average of the estimated standard errors of the log risk ratio. In addition, we estimated the effect of Z_3 on the outcome using a marginal structural linear binomial model. Again, we computed the mean risk difference ('estimated RD'), empirical standard error of the estimated risk difference and average of the estimated standard errors of the risk difference. Finally, we attempted to fit multivariable log binomial regression models to each simulated cohort. From 1000 replications of the study, we tabulated the number of multivariable binomial models that converged.

Next, simulations were conducted in which Z_2 and Z_3 were two binary exposure variables of interest. A correctly specified logistic model was fit to each simulated cohort to predict Z_2 as a function of Z_1; the inverse of the predicted probabilities served as the basis for one set of weights, stabilized by marginal probabilities of each level of Z_2. A second correctly specified logistic model was fit to each simulated cohort to predict Z_3 as a function of Z_1 and Z_2, stabilized by marginal probabilities of each level of Z_3; the inverse of the predicted probabilities served as the basis for the second set of weights. The product of the two stabilized

weights served as the final weight in a marginal structural model. We computed the mean departure from additivity by fitting marginal structural linear binomial regression models with Z_2, Z_3 and Z_2 x Z_3 as explanatory variables, the last term quantifying deviation from additivity of effects; we also report the empirical standard error of the product term and the average of the estimated standard errors for this term.

Empirical example

Data were obtained from the Evans County study for a cohort of 609 White males who were followed for 7 years, with coronary heart disease (*CHD*) status as the outcome of interest. These data are publicly available [http://web1.sph.emory.edu/dkleinb/logreg3.htm#data] and used in a popular textbook on logistic regression.[18] The primary exposure variable of interest is *CAT*, a dichotomous variable indicating high or normal catecholamine level. Hypothesized confounders of the effect of *CAT* on *CHD* include *AGE* (a continuous variable for age in years), *CHL* (a continuous variable for cholesterol in mg/dl) and *SMK* (a dichotomous variable indicating whether the person ever smoked or never smoked). To calculate the denominators used to construct the weights for standardization, we estimated conditional probabilities using logistic regression with the exposure variable, *CAT*, as the dependent variable; explanatory variables were *AGE*, AGE^2, AGE^3, *CHL*, CHL^2, CHL^3, *SMK* and product terms of the form *AGE* x *SMK*, AGE^2 x *SMK* and AGE^3 x *SMK*. Quadratic and cubic functions of *AGE* and *CHL* and product terms between variables were included because these improved balance in covariate distributions across levels of the exposure variable. To stabilize these weights, we set the numerator equal to the marginal probability of each person's observed *CAT* level (i.e. 0.2 for those with $CAT = 1$ and 0.8 for those with $CAT = 0$). The mean weight was 0.98, the minimum weight was 0.26 and the 5th, 25th, 50th, 75th and 95th percentiles of weights were 0.39, 0.81, 0.86, 1.04 and 1.61, respectively. The maximum estimated weight was 6.85. To estimate the effect of *CAT* on *CHD*, we used a marginal structural log binomial model for the risk ratio, and a marginal structural linear binomial model for the risk difference, with robust variance estimators to obtain 95% confidence intervals.

Next, we fit a standardized binomial regression model with two binary exposure variables of interest: *SMK* and *CAT*. To calculate inverse probability weights, we fit a first logistic regression model for *SMK* in which *AGE*, AGE^2 and AGE^3 were explanatory variables. We fit a second logistic regression model for *CAT* with *SMK*, *AGE*, AGE^2, AGE^3, *CHL*, CHL^2, CHL^3 and product terms of the form

Table 1. Illustrative simulated data. Characteristics of covariates Z_1 and Z_2 by level of Z_3 in observed data and weighted data

	Exposed ($Z_3 = 1$)	Unexposed ($Z_3 = 0$)	Marginal over Z_3
Observed data			
$Z_1 = 2$	17.4%	26.4%	25.0%
$Z_1 = 3$	6.9%	28.3%	25.0%
$Z_2 = 1$	4.7%	8.1%	7.6%
D	9.7%	18.1%	16.8%
Weighted data			
$Z_1 = 2$	25.2%	25.0%	25.0%
$Z_1 = 3$	24.8%	25.0%	25.0%
$Z_2 = 1$	7.6%	7.6%	7.6%
D	7.1%	19.4%	17.5%

Table 2. Illustrative simulated data. Characteristics of covariate Z1 by level of Z2 and Z3 in observed data and weighted data

	$Z_3 = 1$ $Z_2 = 1$	$Z_3 = 1$ $Z_2 = 0$	$Z_3 = 0$ $Z_2 = 1$	$Z_3 = 0$ $Z_2 = 0$	Marginal Over Z_3, Z_2
Observed data					
$Z_1 = 2$	7.1%	17.7%	16.5%	27.3%	25.0%
$Z_1 = 3$	1.1%	7.2%	6.4%	30.2%	25.0%
D	4.2%	10.1%	10.3%	18.8%	16.8%
Weighted data					
$Z_1 = 2$	24.7%	25.0%	25.1%	25.0%	25.0%
$Z_1 = 3$	25.6%	25.0%	24.9%	25.0%	25.0%
D	2.7%	7.6%	7.7%	20.8%	17.6%

Table 3. Risk differences and risk ratios by level of Z_2 and Z_3. Results obtained over 1000 iterations of the simulation

	$Z_3 = 1$ $Z_2 = 1$	$Z_3 = 1$ $Z_2 = 0$	$Z_3 = 0$ $Z_2 = 1$	$Z_3 = 0$ $Z_2 = 0$
Risk difference (95% CI[a])	−0.18 (−0.22, −0.14)	−0.13 (−0.15, −0.12)	−0.13 (−0.15, −0.11)	0.0
Risk ratio (95% CI[a])	0.15 (0.05, 0.53)	0.37 (0.31, 0.44)	0.37 (0.28, 0.48)	1.0

[a]Robust confidence intervals were estimated to account for within−subject correlation induced by weighting.

$AGE \times SMK$, $AGE^2 \times SMK$ and $AGE^3 \times SMK$ as explanatory variables. We considered the model for *SMK* first since previous smoking status may be associated with *CAT*, the other exposure of interest, but is not affected by *CAT*. By ordering our two models based on appropriate temporal relationships between variables, we mitigated improper model specification in this example. We estimated the marginal probability of the dependent variable—*SMK* in the first model and *CAT* in the second—using a logistic regression model. We calculated stabilized weights for *CAT* and *SMK* separately, using the marginal probabilities of each person's observed level of each exposure as numerators for those weights. The mean, minimum and maximum values for the first set of stabilized weights were 1.00, 0.65 and 1.45, respectively. The mean, minimum and maximum values for the second set of weights were 0.98, 0.26 and 6.85, respectively. The product of the resultant weights ranged from 0.23 to 6.01 with a mean value of 0.99. We fit marginal structural log binomial and linear binomial regression models with *SMK, CAT* and their product term as explanatory variables.

Results

Simulation results

We first focused on estimation of the effect of Z_3, a binary exposure variable. For each of the 1000 simulations, we fitted a standardized log binomial regression model and estimated the risk ratio and robust standard error; the average of the estimated risk ratios was 0.37, the value specified under the simulation set-up (i.e. $\exp(-1) = 0.37$).

Table 4. Characteristics of covariates by level of catecholamine (*CAT*) in observed data and weighted data. Evans County study of 609 men

	Exposed ($CAT = 1$)	Unexposed ($CAT = 0$)	Total
Observed data	($n = 122$)	($n = 487$)	($n = 609$)
Age in years (mean)	61	52	54
Cholesterol in mg/dl (mean)	199	215	212
History of ever smoking	63%	64%	64%
Coronary heart disease	22.1%	9.0%	11.7%
Weighted data	($\sum_{\text{weights}}$[a] $= 111$)	($\sum_{\text{weights}} = 487$)	($\sum_{\text{weights}} = 598$)
Age in years (mean)	55	54	54
Cholesterol in mg/dl (mean)	206	212	211
History of ever smoking	58%	64%	63%
Coronary heart disease	27.3%	10.8%	13.8%

[a]Sum of weights. Weights were calculated from a logistic regression model for the dependent variable, *CAT*; weights were stabilized by a numerator equal to the marginal probability of ($CAT = 1$) in the study population. The mean weight was slightly less than 1 (0.98), therefore the total sample size for the weighted analysis is slightly smaller than for the unweighted analysis.

Table 5. Characteristics of covariates, risk of coronary heart disease (*CHD*) and *CHD* risk differences and risk ratios by high and low catecholamine level (*CAT*) and history of ever smoking (*SMK*) in observed data and weighted data. Evans County study of 609 men

	$CAT = 1$ $SMK = 1$	$CAT = 1$ $SMK = 0$	$CAT = 0$ $SMK = 1$	$CAT = 0$ $SMK = 0$	Total
Observed data	($n = 77$)	($n = 45$)	($n = 310$)	($n = 177$)	($n = 609$)
Age in years (mean)	61	61	51	54	54
Cholesterol in mg/dl (mean)	200	198	214	216	212
Coronary heart disease (CHD)	24.7%	17.8%	11.3%	5.1%	11.7%
Weighted data	($\sum_{\text{weights}}$[a] $= 65$)	($\sum_{\text{weights}} = 47$)	($\sum_{\text{weights}} = 312$)	($\sum_{\text{weights}} = 176$)	($\sum_{\text{weights}} = 600$)
Age in years (mean)	56	53	54	54	54
Cholesterol in mg/dl (mean)	204	213	211	213	211
Coronary heart disease (CHD)	31.6%	22.1%	14.7%	5.0%	14.3%

[a]Sum of final weights. Final weights were the product of weights calculated from two sequential logistic regression models fit for dependent variables *SMK* and *CAT*, respectively; resultant weights from each model were stabilized by the numerator equal to the marginal probability of the occurrence of the dependent variable ($SMK = 1$ in the first model and $CAT = 1$ in the second).The mean weight was slightly less than 1 (0.99); therefore the total sample size for the weighted analysis is slightly smaller than for the unweighted analysis.

Table 6. Coronary heart disease (*CHD*) risk differences and risk ratios by high and low catecholamine level (*CAT*) and history of ever smoking (*SMK*) in weighted data. Evans County study of 609 men

	$CAT=1$ $SMK=1$	$CAT=1$ $SMK=0$	$CAT=0$ $SMK=1$	$CAT=0$ $SMK=0$
CHD risk difference (95% CI[a])	0.27 (0.10, 0.44)	0.17 (−0.04, 0.38)	0.10 (0.04, 0.16)	0.0
CHD risk ratio (95% CI[a])	6.33 (2.76, 14.53)	4.44 (1.43, 13.75)	2.95 (1.42, 6.15)	1.0

[a]Robust confidence intervals were estimated to account for within-subject correlation induced by weighting.

The empirical standard error of the log risk ratio was 0.095, and the average of the estimated standard errors of the log risk ratio was 0.095. We also fitted a standardized linear binomial regression model and estimated the standardized risk difference and robust standard error; the average of the estimated risk differences was −0.123, the empirical standard error of the risk difference estimate was 0.008 and the average of the estimated standard errors of the risk difference was 0.008.

Table 1 reports the distributions of Z_1 and Z_2 between subgroups defined by Z_3 in the original (unweighted) data and in the weighted data. The distributions of Z_1 and Z_2 differed between subgroups defined by Z_3; in the weighted data, the distributions of Z_1 and Z_2 were similar between the subgroups defined by Z_3 and these distributions were equal to the marginal distribution of Z_1 and Z_2 in the total study group. In the weighted data, the (standardized) risk of the outcome was 0.071 among the exposed and 0.194 among the unexposed. The ratio of these standardized risks (0.071/0.194) equals 0.37, the value specified under the simulation set-up (i.e. $\exp(-1)=0.37$) and the difference in these standardized risks (0.071 − 0.194) equals −0.123. For this simulation scenario, a multivariable log binomial model with main effects for Z_1, Z_2 and Z_3 converged for each of the 1000 simulations, whereas a multivariable linear binomial model failed to converge in 968 of the 1000 simulations.

Next, we assessed whether the joint effects of two binary exposure variables, Z_2 and Z_3, are additive. Table 2 reports the standardized risk of the outcome for the four possible exposure levels: exposed to both Z_2 and Z_3, exposed only to Z_3, exposed only to Z_2 and unexposed to Z_2 and Z_3; and Table 3 reports standardized risk differences and risk ratios with robust 95% confidence intervals. Potential departure from additivity under the excess risk model can be quantified as the absolute excess risk due to interaction. For the illustrative data in Table 2, this value is $0.027 - 0.076 - 0.077 + 0.208 = 0.082$. Over the 1000 simulations, the average of the estimated absolute excess risk due to interaction was 0.084; this evidence of departure from additivity is consistent with the simulation setting in which the joint effects conform to a multiplicative (log binomial) model. The empirical standard error of the absolute excess risk due to interaction was 0.029, and the average of the estimated standard errors was 0.024.

Empirical results

In the Evans County data, the cumulative incidence of *CHD* over 7 years was 11.7%. The prevalence of the primary exposure, high or normal catecholamine level (*CAT*), was 20.0%. A crude comparison of the risk of *CHD* between categories of *CAT* yielded a risk ratio of 2.45 (95% CI: 1.58, 3.79). Table 4 reports the characteristics of *AGE*, *CHL* and *SMK* at each level of *CAT* in the observed data and in the weighted data. In the observed data, the average *AGE* among those with high *CAT* was 9 years greater than among those with low *CAT*, whereas the average *CHL* was lower among those with high *CAT* than among those with low *CAT* (Table 4). In the weighted data, mean values for *AGE* and *CHL* were similar between *CAT* groups. The variable *SMK* was slightly imbalanced between exposed ($CAT=1$) and unexposed ($CAT=0$) in the weighted data, although the difference was not large. In the weighted data, the (standardized) risk of the outcome was 0.273 among the exposed and 0.108 among the unexposed. The ratio of these standardized risks was 2.54 (robust 95% CI: 1.44, 4.46; bootstrapped 95% CI: 1.35, 4.36) and the difference in these standardized risks was 0.165 (robust 95% CI: 0.030, 0.300; bootstrapped 95% CI: 0.041, 0.309). Because the outcome was a common event, the risk ratio diverged from the odds ratio; expressing the exposure contrast in terms of the odds ratio yielded an effect measure that was further from the null than the risk ratio (standardized odds ratio = 3.11; robust 95% CI: 1.48, 6.52). We attempted to fit a log binomial regression model for the association between *CHD* and *CAT*, adjusting for *AGE*, AGE^2, AGE^3, *CHL*, CHL^2, CHL^3, *SMK* and product terms of the form *AGE* x *SMK*, AGE^2 x *SMK*, and AGE^3 x *SMK*; however, reliable estimates for the model parameters could not be obtained due to problems of poor model convergence.

In Table 5 we examine the joint effects of *CAT* and *SMK*. The upper half of Table 5 reports the characteristics of *AGE* and *CHL* as well as the risk of *CHD* at each level

of *CAT* and *SMK* in the observed data from Evans County. The lower half of Table 5 reports results from the weighted data at each level of *CAT* and *SMK*, including the characteristics of *AGE* and *CHL*, as well as standardized risk of *CHD* at each level of *CAT* and *SMK*. In the weighted data, the covariates *AGE* and *CHL* have similar distributions at each level of *CAT* and *SMK*; therefore, potential confounding by these variables was largely nullified. Table 6 reports standardized risk differences and risk ratios with robust 95% confidence intervals. Potential departure from additivity under the excess risk model can be quantified as the absolute excess risk due to interaction, $0.316-0.221-0.147+0.050=-0.003$. To obtain empirical standard error estimates and robust confidence interval for the standardized absolute excess risk due to interaction (95% CI: −0.27, 0.27), we fit a linear binomial model with *CAT*, *SMK* and their product term (the estimated coefficient for the product term corresponding to the absolute excess risk due to interaction). There was little observed departure from risk difference additivity.

Discussion

Epidemiologists often wish to compare the risk of disease, or prevalence of disease, between two or more groups of people. In observational studies, a routine concern is that the groups under comparison may differ with respect to other risk factors for disease. A common approach to deal with such concerns is to fit a multivariable regression model for the outcome. However, another approach that has a long tradition in epidemiology is to use standardization, so that in the weighted data the exposure groups under comparison are similar with respect to other disease risk factors.[19] The problem of confounding is dealt with by weighting the observed data so that the exposure groups under comparison are similar with respect to covariates of concern.

The standardization approach described in the current paper is essentially a simple marginal structural binomial regression model for a point exposure study.[10] Whereas marginal structural models are often discussed for handling of time-varying confounding in longitudinal data analyses, as we illustrate here, in some settings standardization by inverse probability of exposure weighting can offer a useful approach for handling potential confounders in analyses of cross-sectional data and cohort data in which prevalences or incidence proportions are compared between groups defined by baseline characteristics. The use of a standardized log binomial model for estimation of adjusted risk ratios is similar to the approach proposed by Greenland for the calculation of standardized risks when covariates are categorical.[24] However, marginal structural models with inverse probability of exposure weighting readily allow incorporation of continuous covariates into the exposure prediction model (i.e. without forming categories); and methods for robust variance estimation can be readily used to obtain (conservative) confidence intervals.[10]

By using weighting to deal with covariate adjustment, one can obtain adjusted estimates of risk for groups defined by exposure categories. When there is heterogeneity in risk ratios over strata of covariates, as happens for example if the underlying population risk model does not conform to the exponential risk form (or determinants of the outcome are unobserved), a standardized risk ratio derived under this approach retains a very useful interpretation: an average ratio of the expected risks if everyone in the cohort had been exposed vs unexposed. Although in this paper we focus on an approach in which the total study population (exposed and unexposed) serves as the standard, one could choose as the standard the exposed or unexposed group rather than the total study population,[9] or one could make comparisons other than that of everyone exposed vs no one exposed.[15,16]

Of course, investigators sometimes wish to describe heterogeneity in associations rather than obtain a summary risk ratio or difference. We illustrate how to estimate the marginal risks within levels defined by the joint distribution of two exposure variables. The approach allows evaluation of departure from additivity, or departure from multiplicativity, of standardized risks or prevalences. As we illustrate, departure from additivity can be defined by the quantity $R(EF)-R(E\overline{F})-R(\overline{E}F)+R(\overline{E}\overline{F})$. The proposed approach offers a simple approach to estimate the departure from additive effects of the marginal (and standardized) risk differences or ratios. The approach that we propose for estimation of departure from additive joint effects of exposures shares similarities with the approach for estimation of additive odds described by Vanderweele and Vansteelandt;[25] however, whereas they focused on estimation of odds ratios using unmatched cumulative design case-control data under the condition of a rare outcome, we focus on analysis of risk ratios and risk differences (or prevalence ratios and differences) in cohort and cross-sectional data involving common outcomes.

Importantly, positivity and correct specification of the exposure prediction models used to obtain the weights applied to the observed data are required to obtain consistent estimates of risk ratios or prevalence ratios from a weighted (standardized) log binomial model. An important limitation of the current paper is the small set of simulations conducted. We did not, for example, investigate performance of this approach in settings in which there was incorrect specification of the exposure models. This is one important area for future work. When the number of

groups under comparison is small, a simple assessment of balance in covariates between exposure groups in the weighted data suffices to address concerns regarding residual confounding by measured variables. The importance of correct specification of the exposure prediction model is underscored by considerations of the requirements necessary for consistent effect estimates when the exposure variable of primary interest is continuous. We focused on scenarios in which a continuous variable was binned into groups defined by quantiles of the exposure distribution. This approach allows exploration of modelling the exposure-disease association of primary interest using the explanatory variable on its original continuous scale; for the standardized binomial regression model using the continuous form of exposure, residual confounding may occur within the quantiles (e.g. deciles) used to develop the weights, but in many settings of practical importance such bias is likely to be small. However, we note that given a continuous explanatory variable, the investigator may again find that convergence of the log binomial model for the outcome given exposure is problematic; use of categories in a log binomial model will facilitate model convergence. Therefore, we have focused on describing the association by modelling a categorized version of the underlying variable.

Similar methods for obtaining effect estimates have been proposed for estimating standardized ratio or difference measures in Stata.[26,27] Those methods use built-in Stata commands to estimate risk ratios or differences. The current paper extends those earlier papers by making the method easily accessible in SAS, Stata and R, as well as elaborating on the connection to marginal structural models and the interpretation of model fit. The method used in the current paper will typically estimate effects under assumptions different from the methods used in the earlier papers. The modelling assumptions in the current paper are largely contained in the regression model of the exposure conditional on the other covariates whereas, in the earlier papers, the modelling assumptions were largely contained in the regression model of the outcome conditional on the covariate and exposure. Given differences in modelling assumptions, the two approaches could usefully serve as sensitivity analyses, whereas extensions of these approaches lead to doubly robust regression methods.[28,29]

The proposed approach reduces problems of model convergence typical of binomial regression by shifting all explanatory variables except the exposures of primary interest from the linear predictor of the outcome regression model to a model for the standardization weights. There are several approaches that may permit model convergence if a conditional effect measure is desired, for example in a legal or medical setting where an estimate of risk is desired for a claimant or patient with a specified covariate pattern.[30] Analysis using a log-linear Poisson regression model is an alternative to log binomial regression to obtain model convergence;[6] however, such models do not constrain the upper bound on predicted risks and the prospect of predicted probabilities outside their logical range is often unappealing. Constrained optimization avoids this problem, but will often lead to some data points having very high influence and may be sensitive to the parameterization of constraints imposed on the set of covariate predictor variables. Data augmentation can be used with log binomial models to obtain convergence,[31,32] but the resultant maximum likelihood estimates in the modified data may not be close to the maximum likelihood estimate in the original data (if, in fact, the maximum likelihood estimate exists in the original data). However, if the desired target of inference is the exposure effect in the total population, the investigator may find that a standardized comparison of disease risk between exposure groups is readily estimable, whereas problems of convergence occur if one attempts to derive the conditional estimate by fitting a multivariable log binomial regression model.

The proposed approach offers a simple solution to an important set of problems routinely encountered in analyses of epidemiological cohort and cross-sectional data when attempting to estimate adjusted risk or prevalence ratios and differences. For an important class of research questions, the approach presented in this paper may facilitate calculations to obtain adjusted risk or prevalence that are often difficult to obtain by fitting conditional models for the binary outcome given the exposures and covariates.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

S.C. was funded in part by NIH grants R01AI100654, R24AI067039 and P30AI50410.

Conflict of interest: None declared.

References

1. Agresti A. *Categorical Data Analysis. Probability and Mathematical Statistics*. New York, NY: John Wiley, 1990.
2. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* 1998;**280**:1690–91.
3. McNutt LA, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol* 2003;**157**:940–43.

4. Greenland S, Holland P. Estimating standardized risk differences from odds ratios. *Biometrics* 1991;**47**:319–22.
5. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol* 1986;**123**:174–84.
6. Zou G. A modified poisson regression approach to prospective studies with binary data. *Am J Epidemiol* 2004;**159**:702–06.
7. Spiegelman D, Hertzmark E. Easy SAS calculations for risk or prevalence ratios and differences. *Am J Epidemiol* 2005;**162**:199–200.
8. Chu H, Cole SR. Estimation of risk ratios in cohort studies with common outcomes: a Bayesian approach. *Epidemiology* 2010;**21**:855–62.
9. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology* 2003;**14**:680–86.
10. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;**11**:550–60.
11. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008;**168**:656–64.
12. Naimi AI, Moodie EE, Auger N, Kaufman JS. Constructing inverse probability weights for continuous exposures: a comparison of methods. *Epidemiology* 2014;**25**:292–99.
13. Howe CJ, Cole SR, Ostrow DG, Mehta SH, Kirk GD. A prospective study of alcohol consumption and HIV acquisition among injection drug users. *AIDS* 2011;**25**:221–28.
14. Stefanski LA, Boos DD. The Calculus of M-Estimation. *Am Stat* 2002;**56**:29–38.
15. Morgenstern H, Bursic ES. A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population. *J Community Health* 1982;**7**:292–309.
16. Westreich D. From exposures to population interventions: an example in pregnancy and response to HIV therapy. *American Journal of Epidemiology* 2014;**179:**797–806.
17. Hosmer DW, Lemeshow S. Confidence interval estimation of interaction. *Epidemiology* 1992;**3**:452–56.
18. Richardson DB, Kaufman JS. Estimation of the relative excess risk due to interaction and associated confidence bounds. *Am J Epidemiol* 2009;**169**:756–60.
19. Chu H, Nie L, Cole SR. Estimating the relative excess risk due to interaction: a bayesian approach. *Epidemiology* 2011;**22**:242–48.
20. Nie L, Chu H, Li F, Cole SR. Relative excess risk due to interaction: resampling-based confidence intervals. *Epidemiology* 2010;**21**:552–56.
21. Assmann SF, Hosmer DW, Lemeshow S, Mundt KA. Confidence intervals for measures of interaction. *Epidemiology* 1996;**7**:286–90.
22. Skrondal A. Interaction as departure from additivity in case-control studies: a cautionary note. *Am J Epidemiol* 2003;**158**:251–58.
23. Greenland S. Additive risk versus additive relative risk models. *Epidemiology* 1993;**4**:32–36.
24. Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol* 2004;**160**:301–05.
25. VanderWeele TJ, Vansteelandt S. A weighting approach to causal effects and additive interaction in case-control studies: marginal structural linear odds models. *Am J Epidemiol* 2011;**174**:1197–203.
26. Localio AR, Margolis DJ, Berlin JA. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *J Clin Epidemiol* 2007;**60**:874–82.
27. Muller CJ, MacLehose RF. Estimating predicted probabilities from logistic regression: different methods correspond to different target populations. *Int J Epidemiol* 2014;**43**:962–70.
28. Funk MJ, Westreich D, Wiesen C, Sturmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *Am J Epidemiol* 2011;**173**:761–67.
29. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J Am Stat Assoc* 1999;**94**:1096–120.
30. Lumley T, Kronmal R, Ma S. *Relative Risk Regression in Medical Research: Models, Contrasts, Estimators, and Algorithms*. UW Biostatistics Working Paper 293. Seattle, WA: University of Washington, 2006.
31. Deddens JA, Petersen MR. Approaches for estimating prevalence ratios. *Occup Environ Med* 2008;**65**:481, 501–06.
32. Deddens JA, Petersen MR. Re: "Estimating the relative risk in cohort studies and clinical trials of common outcomes". *Am J Epidemiol* 2004;**159**:213–14; author reply 214–15.

APPENDIX 1 – Implementation of percentile bootstrap confidence intervals

The SAS, Stata, and R code shown below can be used to obtain percentile bootstrap confidence intervals for the standardized estimates calculated in Figure 1. The code below uses 2 000 samples, but this can be changed by the user.

```
SAS CODE
* Set up bootstrap resampling;
        data boot;
                do sample=1 to 2000;
                        do i=1 to nobs;
                        pt=round(ranuni(12)*nobs);
                        set one nobs=nobs point=pt ;
                        output;
                        end;
                end;
        stop; run;

        proc logistic data = boot descending; by sample;
                model E = Z1 Z2;
                output out = boot predicted=ps;

        proc logistic data = boot descending; by sample;
                model E = ;
                output out=boot predicted=marg_pr;

        data boot;
                set boot;
                sw = E*marg_pr/ps + (1-E)*(1-marg_pr)/(1-ps); run;

* Obtain bootstrap confidence intervals for risk ratio ;

        ods output Estimates=rr_est ;
        proc genmod data = boot descending; by sample;
                model D = E / link=log dist=bin ;
                weight sw; estimate 'rr' E 1 / exp; run;
                ods rtf close;

        data rr;
        set rr_est;
                if Label ne 'rr'; epred=LBetaEstimate; run;

        proc univariate data=rr;
                var epred; output out=rr_cis pctlpts=2.5 97.5 pctlpre=rr_cis; run;

        proc print data=rr_cis noobs label; run;

* Obtain bootstrap confidence intervals for risk difference ;

        ods output Estimates=rd_est ;
        proc genmod data = boot descending; by sample;
                model D = E / link=identity dist=bin ;
                weight sw; estimate 'rd' E 1;  run;
                ods rtf close;

        data rd;
        set rd_est;
                epred=LBetaEstimate; run;

        proc univariate data=rd;
                var epred; output out=rd_cis pctlpts=2.5 97.5 pctlpre=rd_cis; run;

        proc print data=rd_cis noobs label; run;


STATA CODE
*For the RR
        program margrr, rclass

*Calculate denominators used in inverse probability weights
        logit E Z1 Z2
        predict den

*Create stabilized weights, using a null model with E as the dependent variable
        logit E
        predict num
        g sw=E*num/den+(1-E)*(1-num)/(1-den)

*Fit a log binomial model to the weighted data for the E-D association, with robust
variance
        glm D E [pw=sw],family(binomial) link(log) robust
        matrix b=e(b)
        local b=el(b,1,1)
        return scalar beta =`b'
        drop num den sw
        end


        bootstrap b=r(beta): margrr
        estat bootstrap, eform

*For the RD
        program margrd, rclass

*Calculate denominators used in inverse probability weights
        logit E Z1 Z2
        predict den

*Create stabilized weights, using a null model with E as the dependent variable
        logit E
        predict num
        g sw=E*num/den+(1-E)*(1-num)/(1-den)

*Fit a linear binomial model to the weighted data for the E-D association, with robust
variance
        glm D E [pw=sw],family(binomial) link(id) robust
        matrix b=e(b)
        local b=el(b,1,1)
        return scalar beta =`b'
        drop num den sw
        end

        bootstrap b=r(beta): margrd
        estat bootstrap,


R CODE

    # If 'boot' package not installed, enter in console: install.packages('boot')
        library(boot)

    # Specify starting value for random number generation for re-sampling
        set.seed(12)

    # For the risk ratio, create wrapper function for bootstrap procedure...
    # in which the propensity score is re-estimated in each re-sampling
        sptw.wrap=function(dat,indices)
        {
        dat=dat[indices,]
        E.out=glm(E~Z1+Z2,family=binomial(link="logit"),data=dat,na.action=na.exc
lude)
        ps=predict(E.out,type="response")
        new.sptw=dat$E*mean(dat$E)/ps+(1-dat$E)*(1-mean(dat$E))/(1-ps)
        coef(glm(D~E,family=binomial(link="log"),weight=new.sptw,data=dat))[2]
        }

    # Invoke wrapper function to perform bootstrap using the dataset of interest for
2000 samples
        boot.out=boot(one,sptw.wrap,2000)
        boot.out

    # Display percentile bootstrap point and 95% confidence interval estimates
        median(boot.out$t)
        boot.ci(boot.out,type="perc",conf=0.95)

    # plot density of bootstrap resamples
        plot(density(boot.out$t))


    # For the risk difference, create wrapper function for bootstrap procedure...
    # in which the propensity score is re-estimated in each re-sampling
        sptw.wrap=function(dat,indices)
        {
        dat=dat[indices,]

        E.out=glm(E~Z1+Z2,family=binomial(link="logit"),data=dat,na.action=na.exc
lude)
        ps=predict(E.out,type="response")
        new.sptw=dat$E*mean(dat$E)/ps+(1-dat$E)*(1-mean(dat$E))/(1-ps)

        coef(glm(D~E,family=binomial(link="identity"),weight=new.sptw,data=dat))[2
]
        }

    # Invoke wrapper function to perform bootstrap using the dataset of interest for
2000 samples
        boot.out=boot(one,sptw.wrap,2000)
        boot.out

    # Display percentile bootstrap point and 95% confidence interval estimates
        median(boot.out$t)
        boot.ci(boot.out,type="perc",conf=0.95)

    # plot density of bootstrap resamples
        plot(density(boot.out$t))
```

APPENDIX 2 – Implementation of standardized (weighted) estimates for two dichotomous exposure variables of interest

SAS CODE

```
/*Calculate denominators for weights. Logistic regression model for SMK. */
        proc logistic data = EVANS descending;
        model smk = age age*age age*age*age;
        output out = outpssmk predicted=ps; run;

/*Fit a second logistic regression model with CAT as the dependent variable. */
        proc logistic data = EVANS descending;
        model cat = smk age age*age age*age*age chl chl*chl chl*chl*chl age*smk
age*age*smk age*age*age*smk;
        output out = outpscat predicted=pc; run;

/*Create one dataset with conditional probabilities for SMK (ps) and CAT (pc) for
each obs. */
        proc sort data = outpssmk; by id;  proc sort data = outpscat; by id; run;

        data margstruc;
        merge outpssmk outpscat; by id; run;

/*Create stabilized weights for SMK, first using a null model with SMK as dependent
variable.*/
        proc logistic data = margstruc descending;
        model smk = ; output out=iptw2 predicted =marg_pr_smk; run;

        data iptw2; set iptw2;
        swsmk = smk*marg_pr_smk/ps + (1-smk)*(1-marg_pr_smk)/(1-ps);run;

/*Create stabilized weights for CAT first using a null model with CAT as the
dependent variable*/
        proc logistic data = iptw2 descending;
        model cat =; output out=iptw2 predicted =marg_pr_cat; run;

        data iptw2; set iptw2;
        swcat = cat*marg_pr_cat/pc + (1-cat)*(1-marg_pr_cat)/(1-pc); run;

/*Compute final regression weights. */
        data iptw2; set iptw2;
        swfinal = swsmk*swcat; run;

/*Fit log binomial model for standardized risk ratios with robust variance. */
        proc genmod data = iptw2 descending;
        class id;
        model chd = cat smk cat*smk / link=log dist=bin covb;
        weight swfinal;
        repeated subject=id / type=ind; run;

/*Fit linear binomial model for standardized risk differences with robust variance. */
        proc genmod data = iptw2 descending;
        class id;
        model chd = cat smk cat*smk / link=identity dist=bin covb;
        weight swfinal;
        repeated subject=id / type=ind;
        run;
```

STATA CODE

```
*Calculate denominators for weights. Model for SMK
        logit smk  age age*age age*age*age
        predict ps

*Second model for CAT
        logit cat  smk age age*age age*age*age chl chl*chl chl*chl*chl age*smk
age*age*smk age*age*age*smk
        predict pc


*Create stabilized weights, using a null model with SMK as the dependent variable
        logit smk
        predict marg_prsmk
        g swsmk=smk*marg_prsmk/ps+(1-smk)*(1-marg_prsmk)/(1-ps)

*Create stabilized weights, using a null model with CAT as the dependent variable
        logit cat
        predict marg_prcat
        g swcat=cat*marg_prcat/pc+(1-cat)*(1-marg_prcat)/(1-pc)

 *Compute final regression weights
      g swfinal = swsmk*swcat

*Fit a log binomial model for standardized risk ratios with robust variance
        glm chd cat smk cat*smk [pw=swfinal],family(binomial) link(log) robust

*Fit a linear binomial model for standardized risk differences with robust variance
        glm chd cat smk cat*smk [pw=swfinal],family(binomial) link(id) robust
```

R CODE

```
# variable definitions:
    # age2 = age*age
    # age3 = age*age*age
    # chl2 = chl*chl
    # chl3 = chl*chl*chl
    # a1s = age*smk
    # a2s = age*age*smk
    # a3s = age*age*age*smk
    # c1s = cat*smk

# Calculate denominators for weights. Logistic regression model for SMK
    smk.out=glm(smk~age+age2+age3,family=binomial(link="logit"),data=evans,n
    a.action=na.exclude)
    ps=predict(smk.out,type="response")
    summary(smk.out)

# Fit a second logistic regression model with CAT as the dependent variable
    cat.out=glm(cat~smk+age+age2+age3+chl+chl2+chl3+a1s+a2s+a3s,family=bin
    omial(link="logit"),data=evans,na.action=na.exclude)
    pc=predict(cat.out,type="response")
    summary(cat.out)

# Create stabilized weights for SMK, using the "mean" operator to create
numerators
    swsmk=evans$smk*mean(evans$smk)/ps+(1-(evans$smk))*(1-
    mean(evans$smk))/(1-ps)
    summary(swsmk)

# Create stabilized weights for CAT, using the "mean" operator to create
numerators
    swcat=evans$cat*mean(evans$cat)/pc+(1-(evans$cat))*(1-
    mean(evans$cat))/(1-pc)
    summary(swcat)

# Compute final regression weights
    swfinal = swsmk*swcat
    summary(swfinal)

# Fit log binomial model for standardized risk ratios with robust variance
    library(geepack)
    summary(geeglm(chd~cat+smk+c1s,family=binomial(link="log"),
    weight=swfinal, id=id, data=evans))

# Fit linear binomial model for standardized risk differences with robust variance
    library(geepack)
    summary(geeglm(chd~cat+smk+c1s,family=binomial(link="identity"),
    weight=swfinal, id=id, data=evans))
```