Published in final edited form as:

Biometrics. 2016 September ; 72(3): 965–975. doi:10.1111/biom.12489.

An Efficient Design Strategy for Logistic Regression Using Outcome- and Covariate-Dependent Pooling of Biospecimens Prior to Assay

Robert H. Lyles^{1,*}, Emily M. Mitchell², Clarice R. Weinberg³, David M. Umbach³, and Enrique F. Schisterman²

¹Department of Biostatistics and Bioinformatics, The Rollins School of Public Health of Emory University, 1518 Clifton Rd. N.E., Mailstop 1518-002-3AA, Atlanta, GA 30322

²Epidemiology Branch, Division of Intramural Population Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892

³Biostatistics and Computational Biology Branch, MD A303, National Institute of Environmental Health Sciences, National Institutes of Health, P.O. Box 12233, Research Triangle Park, NC 27709

Summary

Potential reductions in laboratory assay costs afforded by pooling equal aliquots of biospecimens have long been recognized in disease surveillance and epidemiological research and, more recently, have motivated design and analytic developments in regression settings. For example, Weinberg and Umbach (1999, *Biometrics* **55**, 718–726) provided methods for fitting set-based logistic regression models to case-control data when a continuous exposure variable (e.g., a biomarker) is assayed on pooled specimens. We focus on improving estimation efficiency by utilizing available subject-specific information at the pool allocation stage. We find that a strategy that we call “(y,c)-pooling,” which forms pooling sets of individuals within strata defined jointly by the outcome and other covariates, provides more precise estimation of the risk parameters associated with those covariates than does pooling within strata defined only by the outcome. We review the approach to set-based analysis through offsets developed by Weinberg and Umbach in a recent correction to their original paper. We propose a method for variance estimation under this design and use simulations and a real-data example to illustrate the precision benefits of (y,c)-pooling relative to y-pooling. We also note and illustrate that set-based models permit estimation of covariate interactions with exposure.

Keywords

Efficiency; Epidemiology; Bootstrap; Pooling; Study design

---

*phone: 404-727-1310; fax: 404-727-1370; rlyles@sph.emory.edu.

Supplementary Materials

Web Appendices and Supplementary Tables A–C referenced in Sections 3.1 and 3.2 are available with this paper at the *Biometrics* website on Wiley Online Library.

1. Introduction

Health-related research studies often require expensive laboratory assays of biospecimens to measure continuous biomarkers such as cytokines or endocrine disruptors. The notion of physically pooling specimens before assay has received considerable attention as a design strategy because it offers cost reduction, improved efficiency, and sample conservation, and can reduce the fraction falling below the assay limit of detection by pulling levels toward the mean. Concomitant statistical research has enabled the analyst to validly extract information from pooled measurements (e.g., Dorfman, 1943; Emmanuel et al., 1988; Kline et al., 1989; Lan, Hsieh and Yen, 1993; Brookmeyer, 1999; Schisterman and Vexler, 2008; Schisterman et al., 2010; Tebbs et al., 2014).

A potentially important epidemiologic application of specimen pooling is the use of biomarkers measured on pooled specimens as predictors in multivariable regression models of disease risk. A fundamental assumption is that the assay measures the arithmetic mean of the biomarker concentrations for the specimens included in the pool. The statistical goal is to construct a model applicable to measurements on pooled specimens that estimates the same risk parameters as would a corresponding model applied to measurements on individual specimens. Weinberg and Umbach (1999; hereafter, ‘WU’) were the first to address this problem. They established that, for studies involving a disease that follows a logistic model for risk among individuals, a suitable pooling strategy induces a corresponding set-based logistic model with the same risk parameters. They showed that the set-based model loses little statistical power compared to an individual-based analysis. All predictors, whether measured on pooled specimens or on individuals, enter set-based models as sums across individuals in the pooling set.

An important aspect of a set-based approach is the partitioning of subjects for construction of disjoint pooling sets. In general, one defines pooling strata and forms pooling sets by randomly assigning individuals in each pooling stratum into pooling sets. WU’s set-based models use offsets (predictors treated as having a known coefficient of 1.0) that reflect the allocation of individuals to pooling sets of different sizes, focusing on pooling sets formed within strata defined by disease status. We call this outcome-based pooling strategy “y-pooling.”

An important limitation associated with y-pooling, however, is that one cannot accommodate interaction terms involving the predictor that was assessed on pooled specimens, fundamentally because a sum of products does not equal a product of sums. Consequently, WU recommended an alternative pooling strategy for analyzing such interactions: creating pooling sets within strata defined jointly by disease status and any categorical covariates $\mathbf{C}$ considered potential effect modifiers. We call this strategy “(y,c)-pooling.”

The primary purpose of this paper is to examine the potential efficiency benefits of (y,c)-pooling over y-pooling for set-based logistic analyses when either the variable(s) in $\mathbf{C}$ are confounders or the investigator wishes to retain flexibility to assess possible interactions between variables in $\mathbf{C}$ and the exposure. We hypothesize that increasing the homogeneity within pooling sets (and the heterogeneity between them) would increase the precision of

set-based estimates. If so, (y,c)-pooling, which uses more subject-specific information in pool allocation, should provide more precise estimation than y-pooling.

Variance estimation, however, is not straightforward. Because $\mathbf{C}$ is a random variable, the numbers of individuals and of pooling sets in each (y,c) stratum are random. Thus, model-based estimates of standard errors derived assuming fixed offsets are too small for any parameters whose estimates depend critically on the offsets. We provide two bootstrap-based approaches to estimate standard errors.

We use simulations to evaluate the proposed methods and assess efficiency benefits. Finally, we illustrate the approach using data from a substudy of the Collaborative Perinatal Project (Hardy, 2003; Whitcomb et al., 2007), where the goal was to associate cytokine levels and other subject characteristics with the risk of spontaneous abortion.

2. Methods

2.1 Risk model for individuals

Consider a binary outcome Y , a continuous exposure variable X , and a T -dimensional set of covariates $\mathbf{C}$ (e.g., possible confounders) that are of interest. Covariates may be continuous or categorical; if categorical, we code the variables in the models as 0–1 indicator functions of category membership. We use the following notation to specify the assumed logistic risk model for individual subjects:

$$\text{logit } \{\Pr(Y_{ij}=1|x_{ij}, \mathbf{c}_{ij})\} = \alpha + \beta x_{ij} + \sum_{t=1}^T \gamma_t c_{ijt}, \quad (1)$$

($i=1, \dots, k$; $j=1, \dots, g$; $t=1, \dots, T$). Here, i indexes the pooling set, j indexes the g_i individuals within pooling set i , and t indexes covariates. Inference concerns the individual-level parameters of model (1), with the challenge being that exposure (X) is not measured individually but in pools.

2.2 Construction of pooling sets

For y-pooling, the outcome determines two pooling strata; for (y,c)-pooling, the outcome and selected covariates jointly determine the pooling strata. Denote the number of individuals in each stratum as n_y for y-pooling or as n_{yc} for (y,c)-pooling. For either pooling strategy, one must also specify pooling partitions based on the sets of pool sizes to be used within each pooling stratum. The list of possible pooling-set sizes is often the same for every stratum. Set-based models require that the corresponding $Y=1$ and $Y=0$ partitions within each $\mathbf{C}$ stratum have at least one pool of each selected pooling set size. Denote the number of pooling sets of size g in each stratum as m_{yg} for y-pooling or as m_{yeg} for (y,c)-pooling. Notice that $g \times m_{yeg}$ is the number of subjects allocated to pooling sets of size g in stratum (y,c), which we denote as n_{yeg} . The investigator randomly partitions subjects in each stratum into pooling sets according to the specified pooling design, ensuring at least one pool of each

desired size in every (y,c) stratum. Of course, a pooling design may specify that some pools have $g_i=1$, i.e., a single specimen (see Section 4).

2.3 The WU approach to set-based modeling

WU used a convolution-based argument to derive their set-based logistic model for risk, assuming that risk in individuals obeys a logistic model and that the individuals in a pooling set are random realizations from their outcome category. Although WU present their formal models without multiple covariates, the arguments readily generalize, so we employ model (1) to illustrate their approach. The predictors in these set-based models are sums of exposures or of individually-ascertained covariate values across members of the pooling set. Both y-pooling and (y,c)-pooling create pool sets that are matched on the outcome. Let $Y_i=1$

or 0 for a “case” or “control” pooling set. Further, let $x_{i\bullet} = \sum_{j=1}^{g_i} x_{ij}$; the observed value of $x_{i\bullet}$ for pooling set i is the measured exposure concentration in the pooled specimen multiplied

by g_i . Finally, let $c_{i\bullet t} = \sum_{j=1}^{g_i} c_{ijt}$, that is, the sum of the measured values of the t -th covariate across the g_i individuals in pooling set i .

WU focused on the retrospective (case-control) sampling design where, even for individual-level logistic models, the intercept cannot be estimated without external knowledge of the relative sampling fractions (Prentice and Pyke, 1975). WU showed that, under y-pooling, model (1) implies the following set-based logistic regression model:

$$\text{logit } \{\Pr(Y_i=1|x_{i\bullet}, c_{i\bullet})\} = g_i \alpha^* + \beta x_{i\bullet} + \sum_{t=1}^T \gamma_t c_{i\bullet t} + \ln \left\{ \frac{\Pr(\text{case set}|g)}{\Pr(\text{control set}|g)} \right\} \quad (2)$$

where in practice one estimates the offset as $\ln(m_{1g_i}/m_{0g_i})$ where m_{1g_i} (m_{0g_i}) denotes the number of case (control) pooling sets of size g_i . Note that m_{1g_i} and m_{0g_i} must both be positive to avoid infinite offsets. In (2), β and γ are identical to the corresponding parameters in (1) and are estimated consistently using set-based analysis; however, α^* in (2) differs from α in (1).

As discussed by WU, interactions between exposures measured on pools and other covariates are not estimable through y-pooling. To allow set-based estimation of an interaction between exposure (X) measured on pools and a categorical covariate (C) measured on individuals, WU proposed (y,c)-pooling. Within pooling sets created for (y,c)-pooling, the categorical covariate is homogeneous and, because it enters the model as a category indicator variable (i.e., being either 0 or 1), the product of the common indicator value in a pooling set and the pooled-exposure sum equals the sum of the products of the corresponding individual values.

WU (1999), however, erred in presenting their set-based model for (y,c)-pooling. The offset they specified was incomplete so that estimates of the covariate main effect (γ) would be biased (though the exposure effect β and covariate interaction effects with exposure would

be estimated consistently). This flaw, addressed obliquely by Saha-Chaudhuri and Weinberg (2013, pg. 129) and specifically by Lyles and Mitchell (2013), was corrected by Weinberg and Umbach (2014).

2.4 WU approach for (y,c)-pooling

Mimicking their original y-pooling convolution argument and explicitly conditioning on accrual (denoted by A) of individuals, Weinberg and Umbach (2014) derived an appropriate offset for (y,c)-pooling for a single binary covariate (C) that interacts with the exposure variable (X) that is subject to pooling. The offset readily generalizes to accommodate (y,c) pooling with multiple categorical covariates **C** (which may or may not interact with exposure) as follows:

$$g_i \ln \left\{ \frac{\Pr(Y_{ij}=0|\mathbf{C}, A)}{\Pr(Y_{ij}=1|\mathbf{C}, A)} \right\} + \ln \left\{ \frac{\Pr(Y_i=1|g_i, \mathbf{C}, A)}{\Pr(Y_i=0|g_i, \mathbf{C}, A)} \right\} \quad (3)$$

where subscripts on **C** and A are suppressed, being the same for all individuals in pooling set i. For fitting, one replaces the theoretical representation of the offset in (3) for pooling set i with the sample-based version:

$$r_{c_{g_i}} = g_i \ln \left(\frac{n_{0c}}{n_{1c}} \right) + \ln \left(\frac{m_{1c_{g_i}}}{m_{0c_{g_i}}} \right) \quad (4)$$

where n_{1c} (n_{0c}) denotes the number of individuals in the pooling stratum for $Y=1$ ($Y=0$) with covariate values **c**, and $m_{1c_{g_i}}$ ($m_{0c_{g_i}}$) denotes the number of $Y=1$ ($Y=0$) pooling sets of size g_i in the same stratum. Thus, under (y,c)-pooling in practice, the WU set-based model corresponding to individual-level model (1) is:

$$\text{logit} \{ \Pr(Y_i=1|x_{i\bullet}, c_{i\bullet}) \} = g_i \alpha^* + \beta x_{i\bullet} + \sum_{t=1}^T \gamma_t c_{i\bullet t} + r_{c_{g_i}} \quad (5)$$

As in (2), the parameters β and γ in (5) are identical to the corresponding parameters in (1) and are estimated consistently using a set-based analysis. In this case, the asterisk on α in (5) is needed only under outcome-dependent (e.g., case-control) sampling (see Section 2.5). Interactions between categorical covariates and the pool-based exposure could be included in (5) as was WU's original intent; the same offsets apply.

In an unpublished technical report, Lyles and Mitchell (2013) proposed inverse propensity weighting (IPW) to alleviate bias induced by the incorrect offset for (y,c)-pooling in WU (1999). Their approach led to alternative offsets that are algebraically equivalent (in large samples) to those in (4). We use the offsets in (4) in what follows, however, consistent with the original development by Weinberg and Umbach (1999).

2.5 Accommodating cohort or cross-sectional sampling in set-based models

WU (1999) focused on the case-control sampling design. Nevertheless, the argument in their 2014 erratum does not require outcome-dependent sampling and the offsets in (4) and model (5) for (y,c)-pooling apply more broadly. In particular, they point out that the bias implicit in the intercept α^* in model (5) depends on outcome-specific accrual probabilities but is zero for designs where accrual is independent of disease status. This includes prospective designs in which subjects are followed disease-free for the same length of time or present with disease before that time, and to cross-sectional designs that target adjusted prevalence odds ratios. Under such designs, (y,c)-pooling permits valid estimation of all parameters in model (1), including the intercept. Hence, precision gains for estimating any such parameters (and functions of them, such as prevalences within subgroups) play into the potential overall efficiency benefits of (y,c)-pooling.

For y-pooling under the aforementioned prospective or cross-sectional designs, the WU (1999) offsets in (2) yield invalid intercept estimates (but valid estimates of other parameters). Suppressing the conditioning on covariates in the WU erratum (2014) yields an appropriate correction. Thus, for valid estimates of intercepts as well as other model parameters under y-pooling for these additional designs, one modifies the offsets in (4) by dropping the c subscripts.

2.6 Variations on (y,c)-pooling

To this point, we have focused on model (1) and implicitly assumed that all covariates in $\mathbf{C}$ are categorical and indexed by dummy variables to define (y,c)-strata. In practice, however, one may wish to include some covariates in model (1) that are not used for (y,c)-stratification. Consider the following extension of model (1) under (y,c)-pooling (i.e., the pool allocation process ignores $\mathbf{Z}$):

$$\text{logit } \{\Pr(Y_{ij}=1|x_{ij}, c_{ij}, z_{ij})\} = \alpha + \beta x_{ij} + \sum_{t=1}^T \gamma_t c_{ijt} + \sum_{s=1}^S \psi_s z_{ijs}, \quad (6)$$

In the corresponding extension of model (2), the sums of the covariates in $\mathbf{Z}$ enter the model just as do the sums of the covariates in $\mathbf{C}$. Further, because the $\mathbf{Z}$'s are not utilized for pool allocation, no new offset-based correction is required in the set-based model corresponding to (6) to restore valid estimation of the coefficients $\boldsymbol{\psi}$ (or $\boldsymbol{\beta}$). The model can thus be fit using the offsets in (4), which materially affect the estimates of $(\alpha, \boldsymbol{\gamma})$ by restoring their consistency.

Next, consider the situation in which one or more of the covariates measured in advance and then used for defining pooling strata are continuous. Some discretization of such continuous covariates is required for (y,c)-pooling (WU, 1999; Lyles and Mitchell, 2013), perhaps most logically based on marginal sample percentiles (e.g., the median or quartiles, depending on available total and stratum-specific sample sizes). If interactions are not a concern, including those continuous covariates (and not the discretized form used for pooling) in the model should improve model aptness and efficiency. Thus, we consider the situation where

continuous covariates are discretized only to allocate subjects to pooling sets, but they enter the set-based models as continuous, again by summing across values in each pooling set. For example, suppose gender and age (in years) are the covariates involved. Age could be divided into two groups at the median so that there would be 8 (y,c) strata based on the outcome, gender, and dichotomized age. Random pooling would proceed within each unique (y,c) stratum, and the set-based counterpart of (6) would be fit using offsets r_{cgi} and with any continuous C's still entering as sums of the non-discretized levels across individual members of the pool. This analysis can be justified using the WU approach as illustrated in Appendix 1.

2.7 Standard error calculations

Our recommended analytical approach subsequent to (y,c)-pooling is to fit a set-based logistic model, using the offsets r_{cgi} in (4). Because the pooling stratum-specific sample sizes (n_{yc}) vary across repeated samples and the probabilities in (3) are estimated from the data, the offsets are subject to sampling variation. Estimated standard errors (SEs) obtained directly from the fit of set-based models (5) and (6) will thus be biased downward for estimates that depend fundamentally on the offsets, i.e., for estimates of the intercept and the main effects of any covariates used in constructing (y,c)-strata. For these potentially biased estimates, we propose to adapt bootstrap procedures (Efron and Tibshirani, 1993) to calculate adjusted standard errors (SEs).

Specifically, we have developed two bootstrap-based approaches to SE estimation, each of which begins as follows. Letting s denote the number of distinct covariate (**C**) strata, define $\mathbf{N} = (N_{01}, N_{02}, \dots, N_{0s}, N_{11}, N_{12}, \dots, N_{1s})$ as a random vector of stratum-specific sample sizes, of which $\mathbf{n} = (n_{01}, n_{02}, \dots, n_{0s}, n_{11}, n_{12}, \dots, n_{1s})$ is the realization observed in the data. For each of B replications (we took $B=50$ in our example and simulations), we first generate a new set of n_{yc} values from a multinomial ($G, \hat{\mathbf{p}}$) distribution, where $\hat{\mathbf{p}} = \mathbf{n}/G$ and G is the total number of individuals in the study. For each such replication, we recalculate the offsets defined in equation (4) using the original sample size for each new draw from the multinomial, and use these new offsets in fitting the poolwise logistic model to a new set of data using one of two approaches.

Here we detail the first approach, recommended when all covariates in **C** are categorical and no other covariates are included in the logistic regression so that equation (5) gives the structure of the corresponding poolwise model. Our strategy in this case is to apply a version of a parametric bootstrap designed to replicate a joint distribution for the individual-level (y,c) data and the pool-wise exposure data. We illustrate this strategy here in the setting of a single binary covariate (**C**), so that $s=2$ and there are a total of 4 (y,c) strata; however, it extends immediately to more general (y,c) stratification settings with multiple categorical **C**'s, assuming the n_{yc} 's are all of reasonable size.

After recalculating the offsets as described above, we randomly form pooling sets within each resulting (y,c) stratum while targeting the same distribution of pool sizes as in the corresponding stratum in the original study. For each such pooling set, we then generate a new poolwise exposure (X_i) based on the following multiple linear regression (MLR) model:

$$X_i = \hat{\psi}_0 g_i + \hat{\psi}_1 y_i + \hat{\psi}_2 c_i + \varepsilon_i,$$

where y_i and c_i each take the value 0 or g_i depending on the (y, c) stratum, g_i is the size of the i th regenerated pool, the distribution of the errors ε_i is $N(0, g_i \hat{\sigma}^2)$ and the values $\hat{\psi}_0$, $\hat{\psi}_1$, $\hat{\psi}_2$, and $\hat{\sigma}^2$ come from a weighted least squares (WLS) regression of the original observed poolwise exposures versus the original observed (g_i, y_i, c_i) with no intercept and weights equal to $1/g_i$. This WLS regression is highly efficient due to (y, c) -dependent pooling and loses no validity due to such pooling since y and c are conditioned upon in the model (Mitchell et al., 2014). The poolwise logistic model is then fit to each of the B resulting sets of data (g_i, y_i, x_i, c_i) , using the recalculated offsets corresponding to each set. Adjusted SEs to accompany the original data-based α and γ estimates are obtained as the square roots of the sample variances of the corresponding estimates across the B replications. Note that if any interaction(s) between X and one or more covariates in C are to be included in the logistic model of interest, then interaction(s) between Y and the same covariate(s) should be included in the above MLR model for X .

As mentioned previously, the above approach improves estimation of standard errors for coefficient estimates that are fundamentally changed by the use of offsets, namely, the intercept and the main effects of any covariates used in constructing (y, c) -strata. For the remaining estimates (e.g., β in (5) and (β, ψ) in (6)), we recommend reporting and basing confidence intervals (CIs) upon the unadjusted model-based SE from the fit of the set-based model. Use of bootstrap SEs for these estimates may be sensitive to deviations from the assumption of normal errors in the MLR model for X , and thus is not recommended (additional simulation results not shown). In contrast, the parametric bootstrap SEs for estimates fundamentally changed by the offsets continued to perform reasonably well in empirical experiments with right-skewed errors, as did the unadjusted SEs for estimates not fundamentally changed.

For the cases covered in Section 2.6 in which there are one or more continuous covariates in C and/or not all covariates are involved in the pool allocation process, we recommend an alternative nonparametric bootstrap procedure described in Appendix 2. We revisit our advice there regarding when to use the bootstrap, and we note in passing that similar bootstrap procedures could be implemented for y -pooling to obtain a corrected SE to accompany the intercept estimate.

3. Simulations

We conducted simulations in a manner mimicking cross-sectional sampling and used SAS software (SAS Institute, Inc., 2008) to generate data and to fit individual and set-based logistic models. We provide specifics regarding data generation in the footnotes to Tables 1–4. In each simulation run, the same individual observations were allocated randomly to sets using both (y, c) - and y -pooling strategies. For simplicity, if a count in a given stratum was not divisible by a desired pool size, a small number of individual observations were omitted at random to maintain equal pool sizes.

3.1 Binary covariate

Table 1 summarizes simulations to compare pooling and analytic strategies, when model (1) contains a single binary covariate. The methods considered include Method 1: y-pooling (2 pooling strata) and fitting the set-based model in (5) using the offsets in (4) after dropping the “c” subscripts; Method 2: (y,c)-pooling (4 pooling strata) and fitting model (5) with the offsets given in (4). Simulation conditions are detailed in the footnote to Table 1, with 2000 runs generating data for 1000 individual subjects in each simulation. 500 pools of size 2 were targeted in the first simulation, 250 of size 4 in the second, and 500 pools of mixed sizes (125 of size 4, 125 of size 2, and 250 individual samples) in the third.

The key finding from Table 1 is that (y,c)-pooling produces more precise coefficient estimates than does y-pooling, as evidenced by the reduced empirical standard deviations displayed in parentheses. This gain is especially evident for the coefficient γ corresponding to the covariate (C) utilized in the pooling process, for which very little efficiency is lost relative to the complete data analysis based on individual assays (cf. top row). More detailed results are shown in conjunction with (y,c)-pooling, in order to illustrate key points about standard errors made in Section 2.7. In particular, the numbers in parentheses in Table 1 correspond to empirical SD's. Those in brackets reflect mean unadjusted SEs derived directly from fitting the set-based model, while those in braces are averages of adjusted SEs via the parametric bootstrap approach of Section 2.7. As expected, without adjustment the estimated SEs are much smaller than the empirical SDs. SEs associated with the estimated exposure coefficient ($\hat{\beta}$) are the exception, as recalculating the offsets has little effect on $\hat{\beta}$ (and no effect when all pool sizes are equal). Mean adjusted SEs come close to matching empirical SDs in each case, and Wald-type CIs based on these SEs approximately achieve nominal 95% coverage. For β , we calculated these CIs using unadjusted SEs.

Table 2 clarifies precision gains obtained through (y,c)-pooling for cross-sectional scenarios examined in our simulations. The primary entries in the table are the empirical relative efficiencies of (y,c)- vs. y-pooling for estimating each model parameter (α , β , γ), as well as for estimating prevalence among those with $X=0$ and $C=1$. Numbers in parentheses provide the percentage of simulation runs in which the standard error accompanying the estimated exposure effect parameter (β) was smaller under (y,c)- as opposed to y-pooling. Under each pooling design, the simulations also examined the impact of the magnitude of marginal association ('none', 'moderate', or 'high'; see Table 2 footnotes) between exposure (X) and the binary covariate (C) used in (y,c)-pooling. In every case (for a total of 36 parameters estimated), precision under (y,c)-pooling was better than under y-only pooling (i.e., relative efficiency > 1). In addition, there is a tendency for larger relative efficiency benefits with larger pool sizes (i.e., for pools of size 4 as opposed to size 2 or a combination of sizes 1, 2 and 4). With few exceptions, relative efficiency increased at least slightly within any given pooling design as the marginal X-C association increased.

Table 3 summarizes a repeat of the simulation from Table 1 involving pools of size 4, but with a much smaller overall sample size ($N=300$). We continue to observe reasonable performance in terms of bias and empirical SDs for estimated model coefficients. In particular, note the heightened precision benefit apparent with (y,c)- as opposed to y-pooling

with respect to estimation of all three model parameters under smaller sample conditions. Table 3 also demonstrates considerable improvements in bias as well as precision obtained by applying a small-sample bias correction approach (Firth, 1993) when fitting the set-based model (2) with the offsets in (4). This approach is readily available using an option in the SAS LOGISTIC procedure (SAS Institute, Inc., 2008), and has particular appeal here since pooling reduces effective sample size. Corresponding estimated SEs were obtained as described in Section 2.7, where we also implemented the Firth bias correction when fitting the model at each bootstrap replication.

We report further empirical observations in the Web-based Appendix (Tables A, B, and C). Table A summarizes simulations similar to those in Table 1 for pools of size 4, except we varied the magnitudes of the assumed β and γ coefficients to see whether they were related to the precision gains possible via (y,c)- as opposed to y-pooling. Table B, analogous to Table 2 in the main text, provides empirical relative efficiency information to accompany Table A. When β or γ was large in absolute value (1 and -1 , respectively), (y,c)-pooling was markedly more efficient than y-pooling. When both coefficients are small (0.25 and -0.25 , respectively), y-pooling was nearly as efficient as (y,c)-pooling. However, in the latter case, y-pooling yielded estimates nearly as precise as the no-pooling estimates, implying little room for improvement. Nevertheless, precision continued to be at least as good under (y,c)- as opposed to y-pooling for estimating every parameter in every scenario studied. This is noteworthy given other reasons to favor (y,c)-pooling, e.g., to retain flexibility to account for effect modification by one or more variables in C.

3.2 Using a continuous covariate or ignoring a covariate in the pooling process

The simulations summarized in Table C of the Web-based Appendix are designed to illustrate scenarios discussed in Section 2.6 [see model (6)]. In both scenarios, a continuous covariate Z is included in the logistic model of interest but is not utilized for pool allocation. We utilized a binary covariate C in model (6) for pooling in the first scenario. In the second, we dichotomized a continuous covariate C at its median to allocate individuals to pools, but modeled C as continuous.

In Table C, we found little difference in precision between (y,c)-pooling and y-pooling when estimating the coefficient (ψ) corresponding to a covariate (Z) that was not utilized for pool allocation. However, (y,c)-pooling improved precision for estimating other model coefficients (especially the covariate main effect γ), regardless of whether C was binary or continuous. SEs estimated using the nonparametric bootstrap procedure (Appendix 2) came close on average to the empirical SDs and yielded near-nominal CI coverage. Note that we used unadjusted SEs obtained from the fit of the set-based model (6) to construct CIs for the β and ψ coefficient estimates, and these coverages are compatible with nominal rates.

4. Example

4.1 Collaborative Perinatal Project substudy data

The Collaborative Perinatal Project (CPP) recruited pregnant women in the U.S. between 1959 and 1974 to examine associations between exposures and pregnancy outcomes (Hardy,

2003). In a nested case-control study of stored serum from the CPP, cytokines were measured in participants who experienced a spontaneous abortion (SA), along with controls (Whitcomb et al. 2007). Covariates include demographics such as age, race, and smoking status. We analyze race (C_1 ; black vs. Caucasian), smoking status (C_2 ; yes vs. no), SA (Y) and the cytokine monocyte chemotactic protein 1 (MCP1; X). Controls were matched to cases by gestational age at specimen collection, but that variable showed no appreciable correlation with exposure (MCP1) levels. We thus carried out a standard, rather than conditional, logistic regression analysis (Breslow and Day, 1980).

In the ancillary study, data on SA, race, smoking and MCP1 were available for 672 individual women. Of these, 307/672 (46%) had a SA. Considering known risk factors, 316/672 (47%) smoked and 191/672 (28%) were black. For an illustrative data analysis, we implemented artificial y- and (y,c)-pooling by summing individual-level cytokine assay measurements within hypothetical pooling sets so that results can be compared across the two strategies.

Exploratory data analysis using the data with individual MCP1 measurements suggested a possible interaction between race and MCP1 exposure with a marginally significant association between MCP1 and SA among Caucasian mothers. In contrast, there was little evidence of an association with exposure in the overall model without the race-by-MCP1 interaction. Consequently, we consider the following two models:

- a. $\text{logit}\{\text{Pr}(\text{SA}=1)\} = \alpha + \beta\text{MCP1} + \gamma_1\text{SMOKE} + \gamma_2\text{RACE}$
- b. $\text{logit}\{\text{Pr}(\text{SA}=1)\} = \alpha + \beta\text{MCP1} + \gamma_1\text{SMOKE} + \gamma_2\text{RACE} + \delta(\text{RACE} \times \text{MCP1})$

4.2 Artificial pooling process

We conducted artificial y-pooling within SA strata only and artificial (y,c)-pooling within (SA, smoking, race), i.e., (Y, C_1, C_2), strata. Our strategy was to form pooling sets of size 4 within each pooling stratum, resorting to singles (pooling sets of size 1) when necessary. Table 4 shows the resulting numbers of pooling sets and singles within each of the 8 (Y, C_1, C_2) strata. To avoid infinite offsets, we included 4 singles in Stratum (0,0,1) even though the number of subjects (56) in that stratum is divisible by 4. This allocation was necessary because Stratum (1,0,1), which contributes to the same offsets, contained some singles.

To ensure the same number of pooling sets under both pooling strategies, we conducted random y-pooling as follows. We randomly created 75 pooling sets of 4 along with 7 singles for $Y=1$ ($n=307$) and 89 pooling sets of 4 along with 9 singles for $Y=0$ ($n=365$). Thus, we had a total of 180 hypothetical assays under either pooling strategy, potentially offering over 70% cost savings compared to the 672 assays required to measure MCP1 on each individual. The sum of the MCP1 measurements for the four individuals in a pool was taken as the assay value for the “pooled” specimen (measured value times pool size).

4.3 Results

The top half of Table 5 shows results for Model a) based on the individual data (672 assays) and based on the 180 pooling sets produced under the two strategies described above. We

used the offsets given in expression (4) when fitting the set-based logistic models for y- and (y,c)-pooling. We fit all models using the LOGISTIC procedure in SAS (SAS Institute, Inc., 2008). We report unadjusted standard errors (SEs) obtained directly from LOGISTIC output for all parameters estimated under no-pooling and y-pooling, and for the exposure (β) estimate under (y,c)-pooling. In the latter case, we favor the unadjusted SE due to some evidence of right-skewness in MCP1 values (see Section 2.7). Given that both covariates (C_1 and C_2) were binary and utilized for pool allocation, we computed adjusted SEs to accompany the α , γ_1 , and γ_2 estimates under (y,c)-pooling using the parametric bootstrap approach described in Section 2.7. We provide no SE for the estimated intercept (α) for y-pooling, as the parameter is of little relevance in this case-control study and the offsets change in repeated sampling so that the unadjusted SE is underestimated. One could employ a similar bootstrap approach to rectify this issue in practice.

The individual data and both pooled analyses agreed well in terms of point estimates and interpretation. Specifically, standard Wald tests suggested a trend toward a deleterious effect of smoking on the odds of SA, and a highly significant positive association between SA status and black ethnicity. The estimated SEs suggest minimal efficiency loss despite the reduction from 672 to 180 hypothetical MCP1 assays. As in the simulations, we observe potentially improved precision with (y,c)- compared to y-pooling. In this example, the individual-data coefficient estimates were also generally closer to those obtained under (y,c)-pooling than to those obtained under y-pooling, in accord with our expectations; however, standard errors are large.

The results under Model b) using the same pooling sets as with Model a) again showed a close general agreement between the coefficient and SE estimates based on the individual data (672 assays) and the pooled data analyses (180 hypothetical assays) (Table 5). Considering possible trends, Wald tests suggested a marginally significant association between MCP1 exposure and SA among white mothers ($p=0.08$ and $p=0.07$ for individual and pooled data, respectively) and a marginally significant interaction between race and MCP1 exposure ($p=0.10$ and $p=0.08$). For the pooled data analysis, we report adjusted SEs based on the bootstrap approach in Section 2.7 in conjunction with the estimated coefficients (α , γ_1 , γ_2) associated with variables used to construct the pooling strata. We report unadjusted SEs for the estimated coefficients (β , δ), given the right-skewness in MCP1 levels. We provide no results for model b) for y-pooling, as (y,c)-pooling is essential for estimation of exposure-by-covariate interactions.

5. Discussion

For an expensive assay, the savings associated with pooling can be substantial. Pooling can also help to conserve irreplaceable biospecimens while enabling inclusion of study subjects who have little sample remaining and would otherwise have to be excluded and could potentially be informatively missing. Weinberg and Umbach (1999) proposed a set-based logistic regression analysis when using pooling to assess continuous biomarker levels in case-control studies. They focused on y-pooling, but suggested (y,c)-pooling for studying interactions. The current article focuses on (y,c)-pooling (with C potentially multivariate) and offers analytic extensions relevant to this important logistic regression scenario, with the

goal of enhancing the potential benefits of pooling in epidemiologic studies. We found that (y,c)-pooling, followed by fitting a set-based logistic regression model with appropriate offsets and corrected standard errors, can ensure validity of estimated model coefficients and corresponding inferences.

We explored potential precision benefits of (y,c)-pooling for estimation of coefficients (β and γ) corresponding to main effects of the pooled exposure X and covariates C . Precision gains associated with (y,c)-pooling would extend the benefits of the method beyond the assessment of possible interaction effects between X and C that was recognized by WU. Tables 1–3 in the main text and Tables A–C in the Web-based Appendix demonstrate how efficiency can be improved by (y,c)-pooling. Interestingly, we find that precision benefits under (y,c)- as opposed to y-pooling for estimating the exposure parameter (β), while smaller, persisted even when X and C were marginally independent (Table 2). It is worth noting that if (y,c)-pooling was conducted at the design phase but one ultimately decided to exclude a covariate in C from the logistic model (e.g., on the assumption that it was not a confounder), then we recommend ignoring that covariate when calculating the offsets in (4). Failing to do so could lead to bias in the estimate of β . Limited simulation studies suggest that β may be more precisely estimated by leaving such a variable out of the model (and offset calculation). However, it is generally safest to include all pool allocation covariates in the model and offset definition unless one is willing to assume that one or more are non-confounders.

We also observed that (y,c)-pooling requires the use of offsets based on probabilities whose estimates are subject to sampling variation. Because the model applied does not capture that variation, unadjusted model-based standard errors for estimated coefficients tend to be too small. We provide bootstrap-based adjusted standard errors for this purpose, with parametric (Section 2.7) or non-parametric (Appendix 2) resampling recommended depending on whether or not all covariates are binary and utilized in the pool allocation process. These approaches work well based on confidence interval coverage in our empirical studies evaluating the performance of set-based logistic regression coefficients under (y,c)-pooling (Tables 1, 3, A, and C). We also make clear that (y,c)-pooling accommodates continuous covariates via discretization at the pool allocation stage, while still allowing their inclusion in the model as continuous (e.g., Table C).

Our simulations demonstrate the importance of the offsets to be used when fitting the set-based model subsequent to (y,c)-pooling. We noted that the WU offsets (as corrected in 2014) can be applied to sampling designs other than case-control, such as a cross-sectional study designed to assess prevalence odds ratios. As illustrated in the CPP example, the coefficient estimation methods we have described readily accommodate X by C interaction terms. The caveat is that such interactions must involve categorical covariates and those covariates must be among those defining (y,c)-pooling strata. For simulation results supporting the use of a propensity probability-based definition of offsets for the purpose of modeling X by C interaction, see Lyles and Mitchell (2013).

We expect set-based analyses that use (y,c)-pooling to behave well in situations where the overall and stratum-specific sample sizes (n_{yc}) are reasonably large. To ensure this, one may

be best served by targeting pooling sets of small to moderate size and avoiding overly fine stratification by using only covariates of high interest for pool allocation, such as known confounders or potentially important effect modifiers such as race or sex. Although (y,c)-pooling allows the flexibility of a range of pool sizes (g), we suspect that using approximately the same predominant pool size and maintaining a similar pool size distribution within each (y,c)-stratum is desirable. An investigator should consider the stratum-specific sample sizes (n_{yc}) when planning what pooling sizes to use. One useful possibility is a “hybrid” design where some individual assays together with larger pool sizes are included within each stratum (Schisterman et al., 2010). Given that empirical studies such as ours can only consider a range of sample size, pool sizes, and true coefficient settings, simulation experiments targeted toward the application at hand may be useful at the design stage before physical pooling of specimens. Such a simulation experiment would likely be most useful when the investigator has at hand all of the covariate data and is at the point of selecting the overall partition to be used in forming pooling sets. To this end, versions of the simulation programs used to produce the results in Tables 1–4 under (y,c)-dependent pooling are available from the authors by request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institute of Environmental Health Sciences (Z01-ES040006), the National Institute of Nursing Research (1RC4NR012527-01), the National Institute of Environmental Health Sciences (5R01ES012458-07), and the National Center for Advancing Translational Sciences (UL1TR000454).

REFERENCES

- Brookmeyer R. Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence. *Biometrics*. 1999; 55:608–612. [PubMed: 11318222]
- Dorfman R. The detection of defective members of a large population. *Annals of Mathematical Statistics*. 1943; 14:436–440.
- Efron, B.; Tibshirani, RJ. *An Introduction to the Bootstrap*. New York: Chapman and Hall; 1993.
- Emmanuel JC, Bassett MT, Smith HJ, Jacobs JA. Pooling of sera for human immunodeficiency virus (HIV) testing: an economical method for use in developing countries. *Journal of Clinical Pathology*. 1988; 41:582–585. [PubMed: 3164325]
- Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993; 80:27–38.
- Hardy JB. The Collaborative Perinatal Project: lessons and legacy *Annals of Epidemiology*. 2003; 13:303–311.
- Hartigan, JA. *Clustering Algorithms*. New York: Wiley; 1975.
- Kline RL, Brothers TA, Brookmeyer R, Zeger S, Quinn TC. Evaluation of human immunodeficiency virus seroprevalence in population surveys using pooled sera. *Journal of Clinical Microbiology*. 1989; 27:1449–1452. [PubMed: 2768435]
- Lan S, Hsieh C, Yen Y. Pooling strategies for screening blood in areas with low prevalence of HIV. *Biometrical Journal*. 1993; 35:553–565.
- Lyles, RH.; Mitchell, EM. Technical Report # 13-02. Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University; 2013. On efficient use of logistic regression to analyze exposure assay data on pooled biospecimens.

- Mitchell EM, Lyles RH, Manatunga AK, Perkins NJ, Schisterman EF. A highly efficient design strategy for regression with outcome pooling. *Statistics in Medicine*. 2014; 33:5028–5040. [PubMed: 25220822]
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.
- Saha-Chaudhuri P, Weinberg CR. Specimen pooling for efficient use of biospecimens in studies of time to a common event. *American Journal of Epidemiology*. 2013; 178:126–135. [PubMed: 23821316]
- SAS Institute, Inc. SAS/STAT 9.2 User's Guide. 2008
- SAS Institute, Inc. 2013 <http://support.sas.com/kb/22/601.html>.
- Schisterman EF, Vexler A, Mumford SF, Perkins NJ. Hybrid pooled-unpooled design for cost-efficient measurement of biomarkers. *Statistics in Medicine*. 2010; 29:597–613. [PubMed: 20049693]
- Schisterman EF, Vexler A. To pool or not to pool, from whether to when: applications of pooling to biospecimens subject to a limit of detection. *Pediatric and Perinatal Epidemiology*. 2008; 22:486–496. [PubMed: 18782255]
- Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics*. 1999; 55:718–726. [PubMed: 11314998]
- Weinberg CR, Umbach DM. Correction to “Using pooled exposure assessment to improve efficiency in case-control studies”. *Biometrics*. 2014
- Whitcomb BW, Schisterman EF, Klebanoff MA, Baumgarten M, Rhoten-Vlasak A, Luo X, Chegini N. Circulating chemokine levels and miscarriage. *American Journal of Epidemiology*. 2007; 166:323–331. [PubMed: 17504778]

Appendix 1

Justifying offsets in eqn. (4) when pooling strata are defined within categories of a continuous covariate

Suppose that one or more continuous covariates are to be included in a logistic model for pooled specimens, and that one seeks improved efficiency via (y,c)-pooling with strata defined in part by categorized versions of those continuous covariates. To justify the use of the offsets in eqn. (4) in this setting, we follow the original WU (1999) derivation of the poolwise logistic model.

We consider a single continuous covariate C , but the argument generalizes immediately to multiple covariates. To form pooling sets, one categorizes C ; let c^* denote an arbitrary category (sub-interval of the range of C). Let $X_{\bullet}$ and $C_{\bullet}$ denote the sums of the X (the variable assessed on the specimen pool) and C values, respectively, for the g subjects in a pooling set, and let Y represent the binary outcome. Consider the joint distribution of $X_{\bullet}$ and $C_{\bullet}$ for g individuals selected from a stratum defined by $Y=1$ and $C \in c^*$. The density is a convolution:

$$\begin{aligned}
& h(\mathbf{X}_\bullet, \mathbf{C}_\bullet | g \text{ random individuals with } Y \\
& \quad = 1 \text{ and } \mathbf{C} \in \mathbf{c}^*) \\
& = \int \dots \int \left[\prod_{j=1}^{g-1} f(\mathbf{X}_j, \mathbf{C}_j | Y_j = 1, \mathbf{C}_j \in \mathbf{c}^*) \right] f(\mathbf{X}_\bullet \\
& \quad - \sum_{j=1}^{g-1} \mathbf{X}_j, \mathbf{C}_\bullet \\
& \quad - \sum_{j=1}^{g-1} \mathbf{C}_j | Y_g \\
& \quad = 1, \mathbf{C}_g \in \mathbf{c}^*) d\mathbf{X}_1 \dots d\mathbf{X}_{g-1} d\mathbf{C}_1 \dots d\mathbf{C}_{g-1}
\end{aligned}$$

Because $\Pr(Y_j | \mathbf{X}_j, \mathbf{C}_j, \mathbf{C}_j \in \mathbf{c}^*) = \Pr(Y_j | \mathbf{X}_j, \mathbf{C}_j)$, we can write

$$f(\mathbf{X}_j, \mathbf{C}_j | Y_j = 1, \mathbf{C}_j \in \mathbf{c}^*) = \frac{\Pr(Y_j = 1 | \mathbf{X}_j, \mathbf{C}_j) f(\mathbf{X}_j, \mathbf{C}_j | \mathbf{C}_j \in \mathbf{c}^*)}{\Pr(Y_j = 1 | \mathbf{C}_j \in \mathbf{c}^*)}$$

Now mimicking the argument on pg. 719 of WU while replacing their $\Pr(D | E)$ with our $\Pr(Y | \mathbf{X}, \mathbf{C})$ and taking the logistic model for individuals as $\text{logit}[\Pr(Y | \mathbf{X}, \mathbf{C})] = \alpha + \beta \mathbf{X} + \gamma \mathbf{C}$, we can then write the expression for $h(\cdot)$ as:

$$\frac{\exp(g\alpha + \beta \mathbf{X}_\bullet + \gamma \mathbf{C}_\bullet)}{\Pr(Y = 1 | \mathbf{C} \in \mathbf{c}^*)^g} \times q(\mathbf{X}, \mathbf{C}),$$

where $q(\mathbf{X}, \mathbf{C})$ is a complicated integral expression as on pg. 719 of WU. The analogous expression for controls becomes

$$h(\mathbf{X}_\bullet, \mathbf{C}_\bullet | g \text{ random individuals with } Y = 0 \text{ and } \mathbf{C} \in \mathbf{c}^*) = \frac{1}{\Pr(Y = 0 | \mathbf{C} \in \mathbf{c}^*)^g} \times q(\mathbf{X}, \mathbf{C})$$

Continuing the derivation as in WU (1999), we arrive at a model for the set-based ratio:

$$\begin{aligned}
& \frac{\Pr(\text{case set} | \mathbf{X}_\bullet, \mathbf{C}_\bullet, g, \forall \mathbf{C}_j \in \mathbf{c}^*)}{\Pr(\text{control set} | \mathbf{X}_\bullet, \mathbf{C}_\bullet, g, \forall \mathbf{C}_j \in \mathbf{c}^*)} \\
& = \exp(g\alpha \\
& \quad + \beta \mathbf{X}_\bullet + \gamma \mathbf{C}_\bullet + g \times \ln \left(\frac{\Pr(Y = 0 | \mathbf{C} \in \mathbf{c}^*)}{\Pr(Y = 1 | \mathbf{C} \in \mathbf{c}^*)} \right) + \ln \left(\frac{\Pr(\text{case set} | g, \forall \mathbf{C}_j \in \mathbf{c}^*)}{\Pr(\text{control set} | g, \forall \mathbf{C}_j \in \mathbf{c}^*)} \right)
\end{aligned}$$

To fit this model, we use the estimated offset in eqn. (4) with a minor redefinition of n_{yc} and n_{yeg} ; they now represent numbers of subjects in strata constructed from the continuous covariate $\mathbf{C}$.

Appendix 2

Bootstrap Approach to Estimate Standard Errors with One or More Continuous Covariates in C and/or Not All Covariates Involved in Pooling

To handle the scenarios discussed in Section 2.6, we base our approach on nonparametric bootstrapping as suggested by writing the joint distribution of $(X, \mathbf{Z}, Y, \mathbf{C})$ as $f(X, \mathbf{Z} \mid Y, \mathbf{C}) \times p(Y, \mathbf{C})$, where $\mathbf{Z}$ is the vector of covariates that are not used in forming the pooling strata and $\mathbf{C}$ is the vector of covariates that are used, together with Y , in forming the pooling strata. The second term is accommodated by regenerating the $n_{y\mathbf{c}}$'s as in the proposed parametric bootstrap approach described in Section 2.7. For each such regenerated set of $n_{y\mathbf{c}}$ values, we define corresponding regenerated $n_{y\mathbf{c},b}$ values ($b=1, \dots, B$) targeting the same proportions ($n_{y\mathbf{c},b}/n_{y\mathbf{c}}$) as in the original data, rounding as necessary to mimic those proportions and then rounding $n_{y\mathbf{c},b}$ to the nearest integer divisible by g . We then select a total of $n_{y\mathbf{c},b}/g$ pools with replacement from among the original pooling sets of size g in stratum $(y, \mathbf{c})$. The complete set of pooling sets so selected at the b^{th} replication is then analyzed via model (6), applying the b^{th} set of regenerated offsets. Adjusted standard errors (SEs) to accompany α and γ estimates from the original data are obtained as the square roots of the sample variances of the corresponding estimates across the B replicates.

The approach described in Section 2.7 was applied to obtain adjusted SEs for the example and for simulations summarized in Tables 1–3. The nonparametric bootstrap approach described here was applied for simulations summarized in Table C of the Web-based Appendix. While we report mean bootstrap-based SEs in all reported simulations, we used unadjusted SE estimates for CI calculations stemming from the fit of a set-based model to the original data when those estimates corresponded to predictors that were not used in the pool allocation process or to $X \times C$ interaction terms. Specifically, we favor unadjusted standard errors accompanying the β coefficient estimates in models (5) and (6), the ψ estimates in model (6), and estimates of coefficients corresponding to $X \times C$ interaction terms in such models (see Section 2.7).

Table 1 Simulation results to compare pooling strategies with a single binary covariate^{*,†}

Pooling Design	Pooling Strategy	$\alpha = -0.5$	$\beta = 1$	$\gamma = -1$
N = 1000	No pooling	-0.502 (0.108)	1.008 (0.092)	-1.006 (0.156)
~500 assays; Pools of size 2	Method 1	-0.500 (0.117)	1.018 (0.115)	-1.014 (0.181)
	Y only			
	Y and C	-0.505 (0.111) [0.084] {0.105} <93.8%>	1.016 (0.107) [0.106] {0.106} <95.3%>	-1.003 (0.157) [0.122] {0.161} <94.8%>
~250 assays; Pools of size 4	Method 1	-0.495 (0.137)	1.037 (0.158)	-1.047 (0.255)
	Y only			
	Y and C	-0.512 (0.116) [0.078] {0.112} <94.4%>	1.031 (0.143) [0.139] {0.143} <95.7%>	-1.002 (0.168) [0.103] {0.167} <94.7%>
~500 assays; Pools of size 1, 2, and 4	Method 1	-0.506 (0.116)	1.018 (0.116)	-1.006 (0.196)
	Y only			
	Y and C	-0.509 (0.107) [0.081] {0.108} <95.2%>	1.016 (0.110) [0.112] {0.112} <95.7%>	-0.998 (0.167) [0.116] {0.164} <93.6%>

* 2000 simulated studies, each generating 1000 individual observations under model (1) with one binary covariate $C \sim \text{Bernoulli}(0.5)$ utilized in (y,c)-pooling and $X \sim 0.25 - 0.5C + \epsilon_X$, where ϵ_X 's are $\sim \text{i.i.d. } N(0,1)$. **Method 1** y-pooling and fitting set-based model (5), using WU offsets (dropping c subscripts) in (4); **Method 2** (y,c)-pooling and fitting model (5) with WU offsets in (4).

Estimates are means across simulations. Numbers in parentheses are empirical SDs across simulations, those in brackets are mean unadjusted SEs, and those in braces are mean adjusted SEs based on parametric bootstrap approach (Section 2.7). Reported coverages are for Wald CIs utilizing adjusted SEs in the case of α , γ , and unadjusted SEs in the case of β .

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Empirical Relative Efficiencies for (Y,C)- vs. Y-Pooling with Varying Levels of Marginal X-C Association^{*,†}

Pooling Design	X-C Association [‡]	$\alpha = -0.5$	$\beta = 1$	$\gamma = -1$	Prevalence [¶]
~500 assays; Pools of size 2	None (59.7%)	1.16	1.08	1.24	1.10
	Moderate (72.7%)	1.11	1.14	1.33	1.21
	High (83.8%)	1.16	1.17	1.33	1.19
~250 assays; Pools of size 4	None (63.6%)	1.58	1.21	2.18	1.55
	Moderate (75.4%)	1.41	1.23	2.29	1.77
	High (85.2%)	1.68	1.58	2.63	1.78
~500 assays; Pools of size 1, 2, and 4	None (63.6%)	1.18	1.06	1.31	1.15
	Moderate (76.9%)	1.18	1.11	1.38	1.21
	High (87.8%)	1.23	1.10	1.43	1.25

^{*} 2000 simulated studies, each generating 1000 individual observations under model (1) with one binary covariate $C \sim \text{Bernoulli}(0.5)$ utilized in (y,c)-pooling and $X=0.25 - \psi C + e_X$, where e_X 's are $\sim \text{i.i.d. } N(0,1)$.

Analysis under y-pooling conducted by fitting model (5), with WU offsets (dropping ϵ subscripts) in (4).

Analysis under (y,c)-pooling conducted by fitting model (5) with WU offsets in (4).

[†] Main table entries are simulation-based relative efficiencies, calculated as squared ratio of empirical SD for estimates under y-pooling to empirical SD under (y,c)-pooling.

[‡] Numbers in parentheses are percentage of runs for which standard error accompanying estimated exposure (β) coefficient was smaller under (y,c)-pooling. X-C association implemented as follows: **None** ($\psi=0$), **Moderate** ($\psi=0.5$), **High** ($\psi=1$).

[¶] Prevalence parameter estimated = $\Pr(Y=1 \mid X=0, C=1) = 0.182$

Table 3

Simulation results to compare pooling strategies with a single binary covariate and small overall sample size^{*,†}

Pooling Strategy		$\alpha = -0.5$	$\beta = 1$	$\gamma = -1$
N = 300	No pooling	-0.513 (0.194)	1.016 (0.175)	-1.013 (0.295)
	Y only	Method 1	1.204 (0.860)	-1.209 (0.994)
	Y and C	Method 2	1.136 (0.378)	-0.996 (0.315)
	Y and C	Method 2 [‡]	1.004 (0.287)	-0.997 (0.298)
~75 assays; Pools of size 4		[0.144]	[0.256]	[0.192]
		{0.216}	{0.288}	{0.323}
		<94.1%>	<97.0%>	<95.6%>

^{*} 2000 simulated studies, each generating 300 individual observations under model (1) with one binary covariate $C \sim \text{Bernoulli}(0.5)$ utilized in (y,c)-pooling and $X=0.25 - 0.5C + e_X$, where e_X 's are $\sim \text{i.i.d. } N(0,1)$. **Method 1** y-pooling and fitting set-based model (5), using WU offsets (dropping c subscripts) in (4); **Method 2** (y,c)-pooling and fitting model (5) with WU offsets in (4).

[†] Estimates are means across simulations. Numbers in parentheses are empirical SDs across simulations, those in brackets are mean unadjusted SEs, and those in braces are mean adjusted SEs based on parametric bootstrap approach (Section 2.7). Reported coverages are for Wald CIs utilizing adjusted SEs in the case of α and γ , and unadjusted SEs in the case of β .

[‡] Method 2, applying Firth (1993) bias adjustment (see Section 3.1).

Table 4

Summary of (Y, C₁, C₂) -pooling strategy applied to CPP example data (Y=SA status, C₁=smoking, C₂=race)

(Y, C ₁ , C ₂) Stratum	# of Subjects (n _{yc})	# Pools of 4 (m _{yc4})	# of Singles (m _{yc1})
(1,1,1)	44	11	0
(1,1,0)	110	27	2
(1,0,1)	59	14	3
(1,0,0)	94	23	2
(0,1,1)	32	8	0
(0,1,0)	130	32	2
(0,0,1)	56	13	4
(0,0,0)	147	36	3
Assay totals	672	164	16

Table 5

Parameter estimates (standard errors) when fitting Models a) and b) to CPP data for 180 artificial pools (164 pools of size 4, 16 singles)

Parameters (Model a)	No pooling (672 assays)	Y- pooling* (180 “assays”)	(Y, C ₁ , C ₂)- pooling [†] (180 “assays”)
Intercept (α)	-0.522 (0.136)	-0.565 (--)	-0.529 (0.145)
MCP1 exposure (β)	0.316 (0.238)	0.367 (0.257)	0.343 (0.249)
Smoking (γ_1)	0.277 (0.157)	0.308 (0.161)	0.278 (0.156)
Race: Black (γ_2)	0.516 (0.174)	0.570 (0.188)	0.518 (0.184)
Parameters (Model b)			
Intercept (α)	-0.564 (0.139)	--	-0.575 (0.154)
MCP1 exposure (β)	0.472 (0.265)	--	0.512 (0.277)
Smoking (γ_1)	0.285 (0.158)	--	0.289 (0.168)
Race: Black (γ_2)	0.735 (0.219)	--	0.775 (0.231)
Race*MCP1 interaction (δ)	-1.207 (0.729)	--	-1.420 (0.811)

* SE's for y-pooling are unadjusted (see Section 2.7); y-pooling is not applicable when models include interactions with a pooled exposure variable.

[†] SE's for (y, c₁, c₂)-pooling (with exception of SEs corresponding to $\hat{\beta}$ and $\hat{\delta}$) adjusted using parametric bootstrap approach (Section 2.7).