Published in final edited form as:

Stat Med. 2014 December 10; 33(28): 5028–5040. doi:10.1002/sim.6305.

A Highly Efficient Design Strategy for Regression with Outcome Pooling

Emily M. Mitchell^a, Robert H. Lyles^a, Amita K. Manatunga^a, Neil J. Perkins^b, and Enrique F. Schisterman^b

^aDepartment of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia 30322, U.S.A.

^bDivision of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, Maryland 20892, U.S.A.

Abstract

The potential for research involving biospecimens can be hindered by the prohibitive cost of performing laboratory assays on individual samples. To mitigate this cost, strategies such as randomly selecting a portion of specimens for analysis or randomly pooling specimens prior to performing laboratory assays may be employed. These techniques, while effective in reducing cost, are often accompanied by a considerable loss of statistical efficiency. We propose a novel pooling strategy based on the k -means clustering algorithm to reduce laboratory costs while maintaining a high level of statistical efficiency when predictor variables are measured on all subjects, but the outcome of interest is assessed in pools. We perform simulations motivated by the BioCycle study to compare this k -means pooling strategy with current pooling and selection techniques under simple and multiple linear regression models. While all of the methods considered produce unbiased estimates and confidence intervals with appropriate coverage, pooling under k -means clustering provides the most precise estimates, closely approximating results from the full data and losing minimal precision as the total number of pools decreases. The benefits of k -means clustering evident in the simulation study are then applied to an analysis of the BioCycle dataset. In conclusion, when the number of lab tests is limited by budget, pooling specimens based on k -means clustering prior to performing lab assays can be an effective way to save money with minimal information loss in a regression setting.

Keywords

Biomarkers; Design; k -means clustering; Pooling; Regression analysis

1. Introduction

Performing laboratory assays on biological specimens is an important aspect of epidemiological research, but the costs associated with large numbers of such tests are often

prohibitive. Potential cost benefits of pooling individual specimens (e.g., serum samples) prior to testing were recognized as far back as the 1940s, when screening military recruits for syphilis [1]. Since then, pooling has become a popular strategy for reducing cost, for instance, in testing donated blood for HIV [2] or determining prevalence of diseases such as HIV and HPV [3–5]. In epidemiological studies, pooling can provide a valuable tool in preserving irreplaceable resources by requiring the use of only a portion of each specimen for analysis [6] and similarly can help meet minimum volume requirements by combining multiple specimens into a single sample to be assayed. Pooling can also reduce information loss when estimating exposure means and variances in the presence of detection limits [7–10].

Pooling has been demonstrated to be an effective tool in studies requiring logistic regression modeling, when the outcome (binary) or an exposure (binary or continuous) is assessed on composite samples [6, 11–15]. While much of the current literature considers settings with pooling applied to an exposure variable of interest or a binary outcome in a logistic regression setting, we focus on pooling to assess a continuous outcome in linear regression. This scenario is particularly instructive, as it permits both a natural framework for analysis (via weighted least squares) and a clear roadmap for efficient pooling design considerations. We compare pooling and selection strategies for simple linear regression (SLR) and then extend these methods to multiple linear regression (MLR), utilizing a k -means clustering technique to identify an efficient pooling strategy that takes advantage of complete data on the variables to be used as predictors. We use simulations to test our methods, and we apply these methods to a regression analysis of 2005 – 2007 data on HDL cholesterol, serum estradiol, and other variables examined in the BioCycle Study [16].

2. Regression Formulation

For this study, we assume that the number of feasible lab assays (n) is limited by budget, so that physically combining individual biological specimens into pools is an attractive option. We first consider the statistical theory underlying several scenarios that may be encountered when working with data from pooled specimens. We then apply this theory to determine efficient strategies for assigning samples to pools, with an emphasis on maintaining high precision levels of regression coefficient estimates while restricting the number of lab assays performed.

2.1. Equal Aliquot Volumes

Consider the MLR model:

$$Y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \epsilon_{ij}, \mathbf{i} = 1 \dots n, \mathbf{j} = 1 \dots m_i,$$

where Y_{ij} is the outcome and $\mathbf{x}_{ij} = (1, x_{ij1}, \dots, x_{ijP})$ the row vector of covariates for the j^{th} subject in the i^{th} pool. P denotes the number of predictor variables, $\boldsymbol{\beta}$ is the column vector of coefficients, m_i is the number of specimens in pool i (i.e. pool size), and ϵ_{ij} represents the error term with mean 0 and variance σ^2 . Furthermore, let $N = \sum_{i=1}^n m_i$ denote the total sample size of collected specimens.

In practice, each specimen may contribute a different volume. When all specimens contribute equal aliquot volumes to a pool, we assume that the lab assay applied to that pooled sample yields the mean concentration over the individual specimens comprising the pool. The MLR model for the i^{th} pool then becomes:

$$Y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i, \quad i = 1 \dots n,$$

where $Y_i = \frac{1}{m_i} \sum_{j=1}^{m_i} Y_{ij}$ is the measured value for the i^{th} pool and $\epsilon_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \epsilon_{ij}$ is the error term for pool i , such that $E(\epsilon_i) = 0$ and $var(\epsilon_i) = \sigma^2/m_i$. $\mathbf{x}_i = \left(1, \bar{x}_{i \bullet 1}, \dots, \bar{x}_{i \bullet P}\right)$

represents the vector of predictors for pool i , where $\bar{x}_{i \bullet p} = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ijp}$ is the arithmetic mean of the p^{th} predictor across all specimens in pool i .

The MLR model for the entire set of pooled values is then:

$$\mathbf{Y}^* = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbf{E}(\boldsymbol{\epsilon}) = \mathbf{0}_n, \quad Var(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{V},$$

with $\mathbf{Y}_{n \times 1}^* = \{Y_i : i = 1 \dots n\}$, $\mathbf{X}_{n \times (P+1)}^* = \{\mathbf{x}_i : i = 1 \dots n\}$, and $\mathbf{V}_{n \times n} = diag(1/m_i)$, where the preceding $diag()$ notation indicates the diagonal matrix with (i, i) element equal to $1/m_i$ ($i = 1 \dots n$). This setting permits a classical application of weighted least squares (WLS) with weight matrix $\mathbf{V}^{-1} = diag(m_i)$; the WLS estimators of $\boldsymbol{\beta}$ and σ^2 are:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}^{*T} \mathbf{V}^{-1} \mathbf{X}^* \right)^{-1} \mathbf{X}^{*T} \mathbf{V}^{-1} \mathbf{Y}^* \\ \hat{\sigma}^2 &= \frac{\mathbf{Y}^{*T} \left(\mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{X}^* \left(\mathbf{X}^{*T} \mathbf{V}^{-1} \mathbf{X}^* \right)^{-1} \mathbf{X}^{*T} \mathbf{V}^{-1} \right) \mathbf{Y}^*}{n - (P+1)}, \end{aligned}$$

Note that if $m_i = m$ for all pools, i.e. all pool sizes are equal, then $\mathbf{V} = (1/m)\mathbf{I}_n$, which reduces to the ordinary least squares (OLS) estimators:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}^{*T} \mathbf{X}^* \right)^{-1} \mathbf{X}^{*T} \mathbf{Y}^* \\ \hat{\sigma}^2 &= \frac{m \mathbf{Y}^{*T} \left(\mathbf{I}_n - \mathbf{X}^* \left(\mathbf{X}^{*T} \mathbf{X}^* \right)^{-1} \mathbf{X}^{*T} \right) \mathbf{Y}^*}{n - (P+1)}, \end{aligned}$$

The above weighted and unweighted estimators for the vector $\boldsymbol{\beta}$, along with the corresponding estimated variance-covariance matrices of $\hat{\boldsymbol{\beta}}$, in each case, are strictly unbiased as follows from OLS and WLS theory.

2.2. Unequal Aliquot Volumes

When aliquot volumes are not uniform across specimens, we assume these differing aliquot volumes are known and yield a pooled measurement that is a weighted average of the biomarker level of each specimen constituting that pool. In these situations, appropriate

adjustments must be made to the weight matrix in the WLS analysis. Let a_{ij} be the number of units (e.g., mL) contributed by the j^{th} member of pool i . It is then reasonable to assume

that the measurement for pool i is the weighted average $Y_i = \left(\sum_{j=1}^{m_i} a_{ij} \right)^{-1} \sum_{j=1}^{m_i} a_{ij} Y_{ij}$, and the MLR model for the i^{th} pool is then

$$Y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i, \quad i = 1 \dots n,$$

where, now, $\mathbf{x}_i = \left(\sum_{j=1}^{m_i} a_{ij} \right)^{-1} \sum_{j=1}^{m_i} a_{ij} \mathbf{x}_{ij}$ is the vector of predictors for the i^{th} pool consisting of the weighted averages of each predictor across all specimens in pool i . The

random error for pool i is denoted by $\epsilon_i = \left(\sum_{j=1}^{m_i} a_{ij} \right)^{-1} \sum_{j=1}^{m_i} a_{ij} \epsilon_{ij}$, such that $E(\epsilon_i) = 0$ and

$Var(\epsilon_i) = \sigma^2 v_i$, with $v_i = \left(\sum_{j=1}^{m_i} a_{ij} \right)^{-2} \sum_{j=1}^{m_i} a_{ij}^2$. Standard WLS regression can still be performed, this time with weight matrix $\mathbf{V}^{-1} = \text{diag}(1/v_i)$. Failure to include the appropriate weights in the regression analysis would be expected to result in a loss of efficiency. Flawed inference due to invalid estimation of regression coefficient standard errors could also occur, unless robust standard errors were applied. Furthermore, if aliquot information is excluded from calculation of each pool's covariate vector (i.e. if the unweighted means are used), coefficient estimates may also be biased.

3. Pooling and Selection Methods

For this study, assume data on the vector of predictors ($\mathbf{X}$) has been collected on N subjects, but funding permits only n lab assays ($n < N$) for assessment of the response Y . The simplest way to reduce the number of assays is to randomly select n specimens for inclusion in the analysis. Another strategy is to randomly allocate each of the N specimens into one of n equal-sized pools, so that all (or essentially all) specimens are included in the analysis. Given that the predictor ($\mathbf{X}$) data are available on each subject prior to pooling, substantial gains in efficiency relative to these random strategies are possible when this information is applied to the selection or pooling process.

Based on the linear regression models presented in Section 2, an optimal pooling or selection strategy with respect to efficiently estimating the coefficient vector $\boldsymbol{\beta}$ will minimize

$var(\hat{\beta}_p)$ for $p = (1, \dots, P)$. In the case of simple linear regression (SLR) with only one predictor, this simplifies to minimizing

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n w_i (x_i - \bar{x}_w)^2}, \quad (1)$$

where w_i is the weight corresponding to observation (or pool) i and

$\bar{x}_w = \left(\sum w_i \right)^{-1} \left(\sum w_i x_i \right)$ is the weighted grand mean.

3.1. Smart Selection

For any strict selection strategy applied to the data (i.e. using only individual specimens), $w_i = 1$ for all i in (1). For the most efficient selection strategy, minimizing (1) equates to

choosing the n observations with covariate (x_i) values farthest from their mean $\left(\bar{x}\right)$. This can be achieved by first choosing the n observations with covariates furthest from the grand mean, then testing several variations by substituting nearby observations and recalculating the sum of squares term. The specimens corresponding to the covariate values with the largest sum of squares are chosen for analysis. We refer to this strategy as “smart selection”.

3.2. Smart Pooling

Although smart selection is the most efficient selection-only strategy for SLR, one major disadvantage of this method is the complete omission of some biospecimens (generally, those closest to the overall mean) from the analysis. A potential improvement is based on a similar idea, but utilizes pooling instead of selection to limit the total number of lab assays performed.

For pools with equal volume aliquots, $w_i = m_i$ is the number of samples in pool i (pool size).

The components in the denominator of (1) are defined as $x_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij}$ and

$\bar{x}_w = \frac{1}{N} \sum_{i=1}^n m_i x_i = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^{m_i} x_{ij}$. Minimizing the variance of $\hat{\beta}_1$ is then synonymous with maximizing the between-pool weighted sum of squares. When pool sizes are equal, this can be achieved by ordering the data by x and forming pools sequentially, so that pool i contains the set of observations $\{(y, x_{(j)}): (i-1)m < j \leq im\}$ where $(y, x_{(j)})$ is the observation associated with the j^{th} order statistic of x (see Appendix for proof). Ma *et al.* [17] refer to this pooling strategy as a ‘D-optimal design’ and demonstrate its ability to improve estimate efficiency when compared to random pooling or selection strategies. For the remainder of this paper, we will refer to this pooling strategy as “smart pooling”.

3.3. k-means Clustering

Further improvements in efficiency can be achieved over smart pooling when pool sizes are permitted to vary, and an optimal solution can be targeted through a k -means clustering algorithm. Given a set of observations and the desired number of clusters, this algorithm assigns each observation to a cluster such that the between-cluster weighted sum of squares is maximized [18]. While the k -means clustering algorithm is traditionally applied to assist with classification analysis, we exploit this existing technique in order to improve estimate precision in regression settings. This algorithm is readily available in standard software packages such as the *kmeans* function in R or the FASTCLUS procedure in SAS.

In the SLR case, the clusters so identified by the k -means algorithm comprise the optimal pools by virtue of minimizing (1), since the between-cluster sum of squares (now the between-pool sum of squares) has been maximized under the k -means algorithm. While k -means clustering provides only small efficiency gains over smart pooling in an SLR scenario, the main advantage of k -means pooling emerges from its flexibility as a tool for

efficient allocation of subjects to pools based on an arbitrary number of predictor variables, such as in multiple linear regression (MLR). In an MLR setting, we seek to minimize

$var(\hat{\beta}_p)$ for $p = 1, \dots, P$, where

$$\begin{aligned} var(\hat{\beta}_p) &= \left\{ p^{th} \text{element of } diag\left(\sigma^2(\mathbf{X}^{*T}\mathbf{V}^{-1}\mathbf{X}^*)^{-1}\right) \right\} \\ &= \sigma^2 \left\{ \left(1 - r_{\mathbf{x}_p|\mathbf{x}_{(-p)}}^2\right) \sum_{i=1}^n w_i \left(x_{i,p} - \bar{x}_p\right)^2 \right\}^{-1} \end{aligned} \quad (2)$$

$\bar{x}_p$ is the weighted mean of $\mathbf{x}_p = (x_{1,p}, \dots, x_{n,p})$, the vector of the p^{th} covariate values for each observation, and $r_{\mathbf{x}_p|\mathbf{x}_{(-p)}}^2$ is the squared coefficient of multiple determination from the weighted regression of $\mathbf{x}_p$ on the other covariates. Of course, simultaneously maximizing efficiency for all regression coefficients is challenging, since a near-optimal pooling strategy for one can be far from optimal for others. The k -means clustering algorithm is particularly

helpful in this case, since it aims to maximize $\sum_{i=1}^n w_i \left(x_{i,p} - \bar{x}_p\right)^2$, the between-cluster sum of squares, for all p [18]. In concept, this is a generalization of the D-optimal design for SLR, which seeks to maximize the determinant of the $\mathbf{X}'\mathbf{X}$ matrix [17]. Delaigle and Hall [19] promote a similar pooling strategy, where variables are separated into bins and observations with $\mathbf{X}$ values in the same bins are pooled. While this binning strategy also strives to improve efficiency, the k -means algorithm provides a more flexible and accessible method for pooling observations with similar covariates values, particularly when cutoff values for bins are not obvious or pre-specified.

When a particular coefficient, say β_p , is of primary interest, it may be tempting to include only the corresponding covariate vector, $\mathbf{x}_p$, in the clustering procedure. Doing so, however, may actually reduce the efficiency of $\hat{\beta}_p$. To see this, consider an example with two independent covariates (X_1 and X_2) and 100 observations, where we seek to use the observed covariate data to create 4 distinct pools that minimize the variance of $\hat{\beta}_1$. This goal is synonymous with maximizing the between-pool sum of squares (BSS) for

X_1 , $\sum_{i=1}^n w_i \left(x_{i,1} - \bar{x}_1\right)^2$, while minimizing $r_{X_1|X_2}^2$. In the Figure 1 example, although the BSS for X_1 is larger when only X_1 is included in the clustering procedure (left panel), the reduction in $r_{X_1|X_2}^2$ from adding X_2 to the clustering procedure (right panel) is large enough to produce an overall reduction in the variance formula for the coefficient estimate of X_1 . Thus, even when covariates are independent (and particularly when they are correlated), k -means clustering applied to all covariates tends to improve efficiency for every coefficient estimate, especially when all are viewed as equally important.

A more effective way to improve estimate precision for a particular coefficient is through weighted k -means clustering. Generally, all variables (both continuous and binary) should be standardized in order to ensure similar influence on the clustering procedure. When applicable, the covariate of interest can then be multiplied by a weight constant (> 1), which will amplify its influence on the algorithm. The efficiency gain for the weighted variable's

coefficient estimate will nearly always result in a decrease in efficiency for all other coefficients, so it is important to carefully assess the goals of the regression analysis prior to performing weighted k -means clustering.

It is worth noting that when the correct regression weights are applied, the WLS coefficient estimates and their estimated variances remain unbiased for all of the pooling and selection methods considered here, including smart selection and pooling as well as pooling based on k -means clustering. This follows from well-known missing data theory, since these strategies create pools based only on fully observed covariate values that are conditioned upon in the analysis [20].

4. Simulation Study

The BioCycle Study, conducted from 2005 to 2007, followed premenopausal women from Western New York State for one or two complete menstrual cycles [16]. Regularly menstruating women not currently taking oral contraceptives were eligible for participation. 259 women between the ages of 18 and 44 completed the study. Data collected during the study included age (years) and BMI (kg/m^2), as well as serum estradiol (pg/mL), vitamin E ($\mu\text{g/mL}$), and HDL levels (mmol/L), which were measured on the 22nd day of a participant's menstrual cycle [16]. Participant BMI was right-skewed, where values ranged from 16.1 to 35.0 with an average body mass index of 24.2. Vitamin E and estradiol levels were also right-skewed, with average values of 10.1 and 136.4, respectively. Age appeared to be approximately normal, with an average participant age of 27 and standard deviation of 8.2. To facilitate comparative analyses, observations containing missing data were removed, and the first 240 of the remaining 242 complete cases were included in the final dataset. In our study, we treat HDL level as the outcome and perform artificial pooling on this variable to test the various pooling strategies. The remaining variables were treated as fully known and their values were used to facilitate the k -means and smart pooling and selection processes.

Results from the regression analysis on the complete BioCycle study dataset (Section 5) motivated the following simulations to test the pooling and selection strategies discussed in Section 3. Since BMI, vitamin E, and estradiol were right-skewed in the dataset, 240 observations were generated under a multivariate normal distribution with mean vector and covariance matrix matching the log-transformed versions of these covariates, along with a variable representing the untransformed age covariate. Standard deviations for these simulated variables ranged from 0.16 to 0.64, with covariances between -0.023 and 0.423 . The simulated versions of $\log(\text{BMI})$ and $\log(\text{vitamin E})$ were exponentiated to resemble their counterparts in the original dataset, while $\log(\text{estradiol})$ retained its log transformation for the same reason. The outcome (HDL) was then generated via SLR or MLR, based on the parameter estimates from regression analyses on the full BioCycle dataset.

4.1. SLR: Equal Aliquots

10,000 simulations were performed in R for each scenario (R code for all of the pool allocations and analyses discussed in the paper is available from the first author). For the first study, we simulate a simple linear regression with equal aliquot volumes. For this

simulation, the predictor and outcome variables were generated to mimic the results from regressing HDL levels on log(estradiol) from the BioCycle Study, such that $Y \sim N(21.3 - 0.83X_1, 4.95^2)$. SLR was performed under each of the pooling and selection methods discussed in Section 3, as well as random pooling and random selection. The *kmeans* function in R version 2.15.0 was used to define *k*-means clusters. Due to the complex nature of the algorithm, the standard *k*-means functions in SAS and R identify a locally optimal solution for the desired number of clusters (pools). Efforts have been made to find more globally optimal solutions through the prudent choice of initial cluster centers (e.g. leader algorithms or incremental selection) or by repeating the procedure at a number of random starts and choosing the best clustering [21, 22]. Simulations (not shown) comparing such global *k*-means approaches with standard *k*-means suggest that, on average, they yield very comparable coefficient estimates in the linear regression setting studied here. Since standard *k*-means is more accessible and computationally efficient, all *k*-means clustering results were produced using this method unless otherwise noted.

For random selection, only the first n observations from the simulated dataset were retained, while random pooling assigned each group of N/n sequential observations to the same pool. Smart pooling was performed similarly to random pooling, except that the simulated data was ordered by X_1 prior to assigning pools. Smart selection was conducted by calculating

the squared distance $\left(x_{i1} - \bar{x}_1\right)^2$ for all i , then choosing the n observations with the largest squared distance values. Variations on this initial selection were then tested, and the selection with the largest sum of squares term was chosen for analysis.

For all simulations (both SLR and MLR), performance of each method is assessed through bias, relative efficiency, and 95% confidence interval coverage for coefficient estimates, where relative efficiency is defined as the ratio of the empirical standard deviation across the 10,000 simulations of the $\hat{\beta}_p$ estimates calculated from the full data regression to that of $\hat{\beta}_p$ under the specified method. To calculate confidence intervals, the additional assumption of normality of errors is applied, so that confidence intervals are calculated as:

$\hat{\beta}_p \pm t_{0.975, df} \hat{SE}$, where $\hat{SE}$ is the estimated standard error of $\hat{\beta}_p$, and $t_{0.975, df}$ is the critical value of the t-distribution with $df = n - (P + 1)$ degrees of freedom.

Since all methods provide unbiased estimates of β_1 , Table 1 displays the relative efficiency and confidence interval coverage for the SLR results. While all methods provide appropriate confidence interval coverage (~95%) for $\hat{\beta}_1$, estimates calculated under *k*-means pooling are the most precise for every number of assays, retaining over 99% efficiency even at 1/15th the original sample size. The smart pooling method performs similarly to *k*-means, although with slightly less precision in all situations, likely due to its additional restriction of equal-sized pools. As expected, the random selection and pooling strategies are the least efficient methods, displaying considerable precision loss with decreased number of assays. While both random strategies appear to perform similarly with respect to precision, smart pooling noticeably outperforms smart selection, providing estimates from the simulation with the smallest number of assays ($n = 16$) that are more precise than estimates from the simulation with the largest number of assays under smart selection ($n = 120$).

Another advantage of using k -means or smart pooling over smart selection is the facilitation of regression diagnostics. It is often beneficial to check the assumption of a linear association between the outcome and predictor variable to determine whether a transformation on X is required. For example, if a quadratic instead of a linear relationship exists between X and Y , this relationship is often preserved in k -means or smart pooling, but may not be as apparent in smart selection. Figure 2 demonstrates a situation where these covariate-informed pooling strategies give a clearer picture of the true relationship between X and Y than smart selection. For this illustration, 240 observations were generated, with $X \sim \text{Unif}(-2, 2)$ and $Y \sim N(X^2, 1)$. 60 observations were selected for smart selection, and 60 pools were formed under smart and k -means pooling.

The quadratic relationship between X and Y , which is easily identified in the scatterplot from the full data, is preserved in both the smart and k -means pooling scatterplots. This relationship is not as apparent in the smart selection scatterplot due to elimination of observations near the center of the X distribution, and the need to test for a quadratic form of the predictor variable is less likely to be identified under this method.

4.2. SLR: Unequal Aliquots

When pools consist of aliquots with unequal volumes, it is important to include aliquot size in both the calculation of the weighted average of the covariate vector as well as the construction of the regression weights (specified in Section 2.2). Failure to do so would likely result in precision loss, inappropriate standard error estimates, or even biased coefficient estimates.

If aliquot volumes are known a priori, estimate precision can be improved by using these values to inform the clustering procedure, particularly in simple linear regression. By including aliquot volume as if it were another subject-specific covariate in the k -means clustering algorithm, pools are formed from specimens with similar aliquot sizes. This strategy can improve precision since w_i , the weight contribution for pool i in (1), tends to deflate when pool i is formed from specimens with differing aliquot sizes, thus increasing the overall variance for $\hat{\beta}_1$. By forming pools from specimens with similar aliquot sizes, potential precision loss is mitigated. This improvement in efficiency is most noticeable in simple linear regression, particularly when the predictor is binary, but becomes less apparent with the inclusion of additional covariates in the regression formulation. Additionally, when aliquot volumes are correlated with a predictor variable, using aliquot sizes to inform pools is less helpful because the k -means clustering procedure will form similar pools regardless of aliquot volume. Fortunately, the effect of including aliquot information in the pooling strategy with respect to relative efficiency gains can be assessed by using fully known covariate and aliquot information to evaluate the denominators of equations (1) or (2). Thus, the value of assigning pools with or without aliquot information can be determined prior to physically forming the pools, so long as aliquot volumes are known beforehand.

Table 2 illustrates the potential efficiency gains when aliquot volumes are included in the clustering algorithm, as well as the consequences of misrepresenting differing aliquot volumes in the regression analysis. A potential mistake when dealing with unequal-sized

aliquots might be to ignore or inaccurately record the aliquot sizes. If aliquot information is excluded from the analysis or if aliquoting errors are large or pervasive, the coefficient estimates as well as the estimated standard errors may be biased. Results from this erroneous method are also included in Table 2 as an illustration of the potential consequence of failing to adequately account for differing aliquot volumes. In this simulation, the bias associated with $\hat{\beta}_1$ (not shown here) under models that completely ignore aliquot volumes is minor, although this will not always be the case in real-life situations. While aliquot volumes are assumed to have no effect on the selection strategies, the SLR results under these methods are reiterated in this table for comparison purposes.

Incorporating aliquot volumes into the k -means clustering procedure provides estimates of β_1 that are nearly fully efficient for the larger number of pools, while omitting this information from the pooling strategy, as in smart pooling, results in relative efficiency levels that are outperformed by smart selection at the largest number of assays simulated. This reduced efficiency under smart pooling, however, is much less sensitive to decreasing the number of assays and easily beats smart selection in each of the remaining scenarios. In all cases, correctly incorporating aliquot information into the calculation of pooled covariates and weights improves estimate precision.

4.3. Multiple Linear Regression

Since k -means is clearly the most efficient pooling method out of those tested in an SLR setting, we now assess its performance in MLR with equal-sized aliquots. 10,000 simulations were conducted to mimic the BioCycle dataset, with $N = 240$ and $(X_1, \log(X_2), \log(X_3), X_4) \sim N_4(\mu_X, \Sigma)$ where μ_X and Σ were chosen to match the sample mean and covariance matrix from the BioCycle dataset. The outcome (Y) was then generated conditional on the simulated covariate values such that $Y \sim N(\mu_Y, 4.64^2)$, where $\mu_Y = 13.8 - 0.94X_1 + 0.08X_2 + 0.04X_3 + 0.21X_4$.

In this simulation study, standard k -means is compared to weighted k -means, smart pooling, and smart selection on X_1 , where X_1 may be considered the main variable of interest. Results from random pooling are also supplied for comparison purposes, where observations were randomly combined into n equal-sized pools. Random selection was also performed, but these results were omitted from Table 3 since they proved indistinguishable from random pooling. For smart pooling and selection, only X_1 was included in the pooling or selection procedure. For standard k -means, all covariates were standardized to have a mean of 0 and standard deviation of 1 prior to clustering, to ensure that each would contribute a similar impact on the final clusters. Covariates were also standardized for weighted k -means, but X_1 was then multiplied by a weight of 2. The value of two was chosen for the weights in this analysis as it provides a reasonable balance toward slightly improving the precision of $\hat{\beta}_1$ without considerably penalizing the precision of the other coefficient estimates. In general, small to moderate weights are preferred since weighted k -means becomes indistinguishable from smart pooling when weights become too large; the influence of the remaining variables on weighted k -means clustering is essentially nullified. To facilitate interpretation of β , all covariates were returned to their original (unstandardized) forms for the regression analysis.

Relative efficiency and 95% confidence interval coverage for these simulations are displayed in Table 3. Again, bias is omitted from the results display since all coefficient estimates are unbiased. All methods at all scenarios provide nominal 95% confidence interval coverage, confirming the validity of the estimates as well as their estimated standard errors under each pooling method. While the precision of each coefficient estimate is proportional to the total number of assays, the relative efficiency of these estimates varies between methods. For β_2 , β_3 , and β_4 , standard k -means provides the most precise estimates for all scenarios, maintaining over 98% efficiency at half the number of lab tests required, and continuing to maintain over 84% efficiency at only 16 lab tests. Weighted k -means performs similarly, but with slightly less precision for these unweighted variables. While smart pooling again outperforms smart selection with respect to the precision of $\hat{\beta}_1$, both of these methods perform quite poorly for estimating the remaining coefficients, exhibiting precision similar to that of random pooling since these variables were excluded from the clustering and selection procedures.

Weighted k -means provides the most precise estimates of β_1 , our coefficient of interest, maintaining over 90% efficiency at the smallest number of pools. Regardless of the correlation between covariates, further simulations (not shown) indicate that weighted k -means regularly outperforms both smart strategies as well as standard k -means for efficiently estimating β_1 , and standard k -means performs best for the remaining coefficients. It is important to note that the performance of weighted k -means depends on the magnitude of the chosen weights, with larger weights for one variable often corresponding to worse precision for the unweighted variables. The magnitude of this cumulative precision loss is often much more considerable than the precision gained in the weighted variable. Thus, even when only one variable is of interest in an MLR setting, smart pooling performs poorly, and weighted k -means requires careful consideration of potential repercussions. As an efficient global strategy, we generally recommend standard k -means pooling. All of our empirical studies suggest that this method yields outstanding efficiency gains over random pooling and provides the best overall solution, particularly when all variables are considered equally important.

5. Data Analysis

Based on the simulations testing the various pooling and selection strategies, standard k -means clustering, on average, provides the most precise estimates in simple linear regression and maintains high levels of precision in MLR even as the number of pools decreases. Now, we apply k -means clustering to the BioCycle study dataset, to illustrate the performance of this pooling strategy compared to results from regression on the full dataset.

For this data analysis, MLR was performed on various numbers of pools formed from a global k -means clustering algorithm [22] on all four predictor variables, where pools were artificially formed assuming no processing or measurement error. The results are displayed in Table 4. Coefficient estimates and standard errors from halving the total number of pools (assays) approximate the results from the full sample extremely closely for all coefficients, both in terms of point estimates and standard errors. Although more discrepancy is observed, as expected, for smaller numbers of pools, remarkably little efficiency is lost even based on

as few as 16 pools. By using the fully observed covariate data to strategically inform the pooling procedure, investigators can closely approximate MLR results from the full data while drastically reducing laboratory costs.

6. Discussion

When the number of lab tests that can be performed is limited by budget, pooling specimens based on k -means clustering applied to individual-level covariates can be an effective way to reduce cost with minimal information loss in a linear regression setting. In MLR, this method performs far better than the current cost-reduction strategies of smart or random pooling or selection, and has the additional facility of incorporating aliquot sizes, when appropriate, to promote efficiency in an SLR setting. By utilizing all of the covariate data to inform pooling, standard k -means clustering provides an accessible method to identify a specified number of pools commensurate with available resources for performing lab assays, while providing an excellent overall solution aimed at favorable precision for each coefficient estimate. In MLR, weighted k -means can be useful if the precision of one coefficient is deemed particularly important, but should be used only after careful consideration of the potential precision reduction of the remaining coefficient estimates.

While the current study focuses on analysis and design considerations when a linear regression outcome is assayed on pooled samples, the case of pooling on exposure also merits mention. In this setting, the same pool-wise regression models and WLS analyses discussed in Section 2 still apply, although the potential efficiency to be preserved by informed pooling will generally be much less than in the case of pooling on the response. This is due to the fact that not only is the pooled predictor (presumably of great interest) unavailable to inform the pools, but inclusion of the observed outcome into a pooling strategy, such as k -means or smart pooling, will introduce bias into a standard WLS analysis. While clustering based on the remaining covariates will at least benefit the precision of their coefficient estimates, additional research is required to identify any methods that can utilize the fully known outcome data to provide unbiased and efficient estimates of the pooled exposure coefficient. Another issue requiring further consideration concerns potential benefits of adjusting the measured analyte of interest by an additional analyte (e.g. serum lipids, creatinine). The resulting ratio of the pooled measurements would not necessarily be equivalent to the mean of the individual-level ratios, which could complicate analysis. Further investigation on this subject is needed to take full advantage of possible efficiency gains from pooling design as well as analyte adjustment techniques.

The simulations in this analysis were designed to mimic the BioCycle study dataset, and pooling strategies were performed with the goal of reducing laboratory cost. Many alterations on these assumptions, however, are likely to occur in real-life pooling scenarios. For instance, simulations and analyses in this study were performed assuming no pooling or measurement error, while, in practice, these potential error components could influence the measured value of the pools and resulting inference. This topic has been explored by Schisterman *et al.* [9, 10] in the context of estimating the marginal mean and variance of a biomarker, and adaptation of such methods to the regression setting with outcome pooling is recommended to ensure proper analysis when such errors are present. In other cases, the

main motivation behind pooling may not be financial. As mentioned in Section 1, pooling can be a useful tool to reduce the number of non-detects in the presence of a limit of detection [7, 8, 10]. In such cases, additional investigation of optimal pooling designs may be necessary in order to maintain a high level of precision while minimizing the number of pools that fall below the detection limit.

Ideally, the final regression models would be known prior to performing any actual pooling, based on a priori biological knowledge or suspected pathway causal diagrams. In such cases, exploration of the best pooling or selection strategy for a particular regression model could be investigated beforehand, since the potential precision of the estimated coefficients depends only on the fully observed covariates. Consequently, the most efficient pooling strategy applied to individual covariate values is expected to be efficient for any additional outcome that might be measured on the pooled samples. Empirical studies suggest that k -means pooling, for instance, can promote high levels of efficiency in a logistic regression setting when pooling is performed on a binary outcome (as described by Vansteelandt *et al.* [23]), particularly when multiple continuous predictors are to be included in the model. In addition, recent work has demonstrated the efficacy of k -means clustering in improving efficiency in regression settings where a skewed outcome is subject to pooling [24]. Thus, pooling designs can retain their potential statistical efficiency advantages for the analysis of multiple outcomes and models, so long as the same covariates are to be considered. Such foresight of the final structure of the regression model, however, may not always be possible. Upon examination of the fitted model, it may be desirable to add or remove predictor variables or apply a transformation (e.g. log) to one or more covariates. In other cases, the pooled measurements treated as an outcome in one model may be useful as an exposure in a different model. In these subsequent scenarios, the original pooling strategy is unlikely to provide optimal efficiency. One way to improve precision might be to employ a two-stage hybrid pooling design, where a selection of individual specimens are assayed individually, and pools are formed only after determining the most likely structure of the main model of interest. Additional investigation into the best way to implement this method, however, is required.

While any kind of pooling or selection strategy will almost always be accompanied by some degree of information loss, application of a k -means clustering strategy can considerably mitigate this loss, and will almost certainly provide more precise regression estimates than random pooling or selection. In this scenario, any potential precision loss associated with a pooling or selection method for a given model can be assessed using only the fully known covariate values, so that the advantages and disadvantages of pooling can be thoroughly evaluated beforehand. This characteristic may prove particularly useful in a cost-benefit analysis, when determining the optimal number of pools to balance statistical precision and lab expense. Ultimately, the decision of whether and how to pool should be carefully considered on a case by case basis prior to performing any actual pooling, in order to ensure maximum benefit from the chosen method.

Acknowledgments

This research was supported by grants from the National Institute of Nursing Research (1RC4NR012527-01), the National Institute of Environmental Health Sciences (5R01ES012458-07), the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, and by the Long-Range Research Initiative of the American Chemistry Council. Additional partial support from the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR000454 is also acknowledged.

Contract/grant sponsor: National Institute of Nursing Research and the National Institute of Environmental Health Sciences

Appendix

Consider a smart pooling strategy of n pools, each with a pool size of m such that pool i is of

the form $P_i = \{(y, x_{(j)}) : (i-1)m < j \leq im\}$ and $\bar{x}_i = \frac{1}{m} \sum_{j=(i-1)m+1}^{im} x_{(j)}$. Then the sum of

squares term in the denominator of (1) is $\sum_{i=1}^n \left(\bar{x}_i - \bar{x} \right)^2$, where

$\bar{x} = \frac{1}{n} \sum_{i=1}^n \bar{x}_i = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m x_{ij}$. We will show that any deviation from this pooling strategy will decrease the sum of squares, thus increasing the variance term in (1).

Consider two pools of size m , denoted P_r and P_s , $r < s$, such that $P_r = \{(y, x_{(j)}) : (r-1)m < j \leq rm\}$ and $P_s = \{(y, x_{(j)}) : (s-1)m < j \leq sm\}$, where $x_{(j)}$ is the j^{th} order statistic. Let $\bar{x}_r$ and $\bar{x}_s$ denote the mean values for these pools, respectively. Now suppose that we switch two elements in these pools, say $x_{r'}$ from pool P_r and $x_{s'}$ from pool P_s . Then the measured value for pool P_r will increase, since every element in the original pool P_s is greater than every element in the original pool P_r . Similarly, the measured value for pool P_s will decrease. Let δ_r denote this change in the measured value of P_r , and let δ_s denote the change in the measured value of P_s , such that

$$\delta = \delta_r = -\delta_s = \frac{1}{m} (x_{s'} - x_{r'})$$

The sum of squares component for the original pools is:

$$\left(\bar{x}_r - \bar{x} \right)^2 + \left(\bar{x}_s - \bar{x} \right)^2$$

And for the new pools is

$$\left(\bar{x}_r + \delta_r - \bar{x} \right)^2 + \left(\bar{x}_s + \delta_s - \bar{x} \right)^2 = \left(\bar{x}_r + \delta - \bar{x} \right)^2 + \left(\bar{x}_s - \delta - \bar{x} \right)^2$$

Then

$$\begin{aligned}
(\bar{x}_r + \delta - \bar{x})^2 + (\bar{x}_s - \delta - \bar{x})^2 &= (\bar{x}_r - \bar{x})^2 + (\bar{x}_s - \bar{x})^2 + 2\delta(\bar{x}_r - \bar{x}) - 2\delta(\bar{x}_s - \bar{x}) + 2\delta^2 \\
&= (\bar{x}_r - \bar{x})^2 + (\bar{x}_s - \bar{x})^2 + 2\delta(\bar{x}_r - \bar{x}_s + \delta) \\
&= (\bar{x}_r - \bar{x})^2 + (\bar{x}_s - \bar{x})^2 + \frac{2}{m^2}(x_{s'} - x_{r'}) \left[(m\bar{x}_r - x_{r'}) - (m\bar{x}_s - x_{s'}) \right] \\
&= (\bar{x}_r - \bar{x})^2 + (\bar{x}_s - \bar{x})^2 - \frac{2}{m^2}(x_{s'} - x_{r'}) \left(\sum_{s \neq s'} x_s - \sum_{r \neq r'} x_r \right) \\
&< (\bar{x}_r - \bar{x})^2 + (\bar{x}_s - \bar{x})^2
\end{aligned}$$

since $\frac{2}{m^2} > 0$, and $x_{s'}x_s > x_{r'}x_r$. Thus, deviating from the smart pooling strategy results in a decrease of the total sum of squares and a subsequent precision loss. This result generalizes to any change in the smart pooling strategy (maintaining equal-sized pools), since any alteration of the pools can be broken down into pairwise switches.

References

1. Dorfman R. The detection of defective members of large populations. *Annals of Mathematical Statistics*. 1943; 14:436–400.
2. 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]
3. Lan S, Hsieh C, Yen Y. Pooling Strategies for Screening Blood in Areas with Low Prevalence of HIV. *Biometrical Journal*. 1993; 35:553–565.
4. Brookmeyer R. Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence. *Biometrics*. 1999; 55:608–612. [PubMed: 11318222]
5. Weinberg C. HPV screening for cervical cancer in rural India. *New England Journal of Medicine*. 2009; 361:305–306. [PubMed: 19610159]
6. Weinberg C, Umbach D. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics*. 1999; 55:718–726. [PubMed: 11314998]
7. Mumford SL, Schisterman EF, Vexler A, Liu A. Pooling biospecimens and limits of detection: effects on ROC curve analysis. *Biostatistics*. 2006; 7:585–598. [PubMed: 16531470]
8. Schisterman F, 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]
9. Schisterman EF, Vexler A, Mumford SL, Perkins NJ. Hybrid pooled-unpooled design for cost-efficient measurement of biomarkers. *Statistics in Medicine*. 2010; 20:598–613.
10. Schisterman EF, Vexler A, Ye A, Perkins NJ. A combined efficient design for biomarker data subject to a limit of detection due to measuring instrument sensitivity. *Annals of Applied Statistics*. 2011; 5:2651–2667.
11. Vansteelandt S, Goetghebeur E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*. 2000; 56:1126–1133. [PubMed: 11129470]
12. Chen P, Tebbbs JM, Bilder CR. Group testing regression models with fixed and random effects. *Biometrics*. 2009; 65:1270–1278. [PubMed: 19210734]
13. Zhang Z, Albert PS. Binary regression analysis with pooled exposure measurements: a regression calibration approach. *Biometrics*. 2011; 67:636–645. [PubMed: 20662830]

14. Zhang Z, Liu A, Lyles RH, Mukherjee B. Logistic regression analysis of biomarker data subject to pooling and dichotomization. *Statistics in Medicine*. 2012; 31:2473–2484. [PubMed: 21953741]
15. Lyles RH, Tang L, Lin J, Zhang Z, Mukherjee B. Likelihood-based methods for regression analysis with binary exposure status assessed by pooling. *Statistics in Medicine*. 2012; 31:2485–2497. [PubMed: 22415630]
16. Schliep KC, Schisterman EF, Mumford SL, Pollack AZ, Zhang C, Yi A, Stanford JB, Hammoud AO, Porucznik CA, Wactawski-Wende J. Caffeinated beverage intake and reproductive hormones among premenopausal women in the BioCycle Study. *American Journal of Clinical Nutrition*. 2012; 95:488–497. [PubMed: 22237060]
17. Ma C-X, Vexler A, Schisterman EF, Tian L. Cost-efficient designs based on linearly associated biomarkers. *Journal of Applied Statistics*. 2011; 38:2739–2750.
18. Hartigan, JA. *Clustering Algorithms*. Wiley; New York: 1975.
19. Delaigle A, Hall P. Nonparametric regression with homogeneous group testing data. *The Annals of Statistics*. 2012; 40:131–158.
20. Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2nd Edition. Wiley; New York: 2002.
21. Jain AK, Murty MM, Flynn PJ. Data Clustering: A Review. *ACM Computing Surveys*. 1999; 31:264–323.
22. Likas A, Vlassis N, Verbeek JJ. The global k -means clustering algorithm. *Pattern Recognition*. 2003; 36:451–461.
23. Vansteelandt S, Goetghebeur E, Verstraeten T. Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics*. 2000; 56(4):1126–1133. [PubMed: 11129470]
24. Mitchell EM, Lyles RH, Manatunga AK, Danaher M, Perkins NJ, Schisterman EF. Regression for skewed biomarker outcomes subject to pooling. *Biometrics*. 2014; 70:202–211. [PubMed: 24521420]

Figure 1.

k -means clustering on independent variables. 100 observations were grouped into four clusters, indicated by different symbols. The left panel displays clusters based only on X_1 . The right panel shows clusters based on both X_1 and X_2 . Between sum of squares (BSS) values for X_1 and $r^2_{X_1|X_2}$ values corresponding to the resulting 4 pools are displayed for each clustering strategy.

Figure 2. Scatterplots of X and Y for the full data, smart selection, smart pooling, and *k*-means pooling methods.

Table 1

Relative efficiency and 95% confidence interval (CI) coverage of $\hat{\beta}_1$ from SLR simulation on $N = 240$ observations with equal volume aliquots. n = number of pools formed or observations selected for analysis, and relative efficiency = $SD\left(\hat{\beta}_1^f\right) / SD\left(\hat{\beta}_1\right)$, where $\hat{\beta}_1^f$ is the parameter estimate under the full data regression. Regression parameters were simulated to mimic the SLR results of HDL levels on log(estradiol) from the BioCycle study.

Method	Relative Efficiency (95% CI Coverage)			
	$n = 120$	$n = 60$	$n = 30$	$n = 16$
<i>k</i> -means	1.000 (95.1)	0.999 (95.0)	0.996 (95.0)	0.993 (95.1)
Smart Pooling	0.999 (95.2)	0.998 (95.1)	0.995 (95.1)	0.988 (95.3)
Smart Selection	0.962 (94.9)	0.838 (94.7)	0.696 (95.0)	0.574 (95.1)
Random Pooling	0.703 (95.2)	0.491 (95.0)	0.337 (95.0)	0.231 (95.0)
Random Selection	0.704 (95.2)	0.489 (95.0)	0.341 (95.2)	0.237 (95.3)

Table 2

Relative efficiency with respect to the full data and 95% confidence interval (CI) coverage of $\hat{\beta}_1$ from SLR simulations with unequal aliquot volumes.

Method	Relative Efficiency (95% CI Coverage)			
	$n = 120$	$n = 60$	$n = 30$	$n = 16$
Selection Strategies				
Smart Selection	0.962 (94.9)	0.838 (94.7)	0.696 (95.0)	0.574 (95.1)
Random Selection	0.704 (95.2)	0.489 (95.0)	0.341 (95.2)	0.237 (95.3)
Correctly-Specified Aliquots [†]				
<i>k</i> -means	0.999 (94.9)	0.995 (95.1)	0.986 (95.1)	0.950 (95.1)
Smart Pooling	0.957 (95.2)	0.930 (95.3)	0.919 (95.2)	0.908 (95.1)
Random Pooling	0.704 (95.2)	0.490 (94.8)	0.338 (95.0)	0.232 (95.0)
Ignoring Aliquots				
<i>k</i> -means	0.950 (94.9)	0.929 (95.1)	0.918 (95.2)	0.911 (94.9)
Smart Pooling	0.950 (95.2)	0.926 (95.2)	0.917 (95.2)	0.906 (95.3)
Random Pooling	0.669 (95.1)	0.453 (95.0)	0.311 (95.1)	0.210 (94.7)

[†] Aliquot volumes randomly selected from the set $(\frac{1}{2}, \frac{1}{2}, \frac{3}{4}, 1)$ informed the *k*-means clustering strategy as well as the appropriate weights for all pooling strategies under the heading “Correctly-Specified Aliquots.” “Ignoring Aliquots” presents regression results when aliquot volumes are ignored in both the pool allocation and analysis steps.

Table 3

Relative efficiency and 95% confidence interval (CI) coverage for estimated regression coefficients from MLR simulations at various numbers of assays.

Method	Relative Efficiency (95% CI Coverage)			
	β_1	β_2	β_3	β_4
<i>n</i> = 120				
Standard <i>k</i> -means [†]	0.983 (95.1)	0.985 (95.0)	0.983 (95.1)	0.981 (94.9)
Weighted <i>k</i> -means	0.993 (95.2)	0.978 (94.9)	0.980 (95.1)	0.977 (95.0)
Smart Pooling	0.988 (95.2)	0.695 (94.8)	0.699 (95.1)	0.692 (94.8)
Smart Selection	0.953 (95.1)	0.698 (95.1)	0.688 (94.9)	0.697 (95.1)
Random Pooling	0.696 (94.7)	0.706 (95.1)	0.697 (95.0)	0.705 (95.3)
<i>n</i> = 60				
Standard <i>k</i> -means	0.957 (95.0)	0.955 (95.0)	0.959 (95.1)	0.952 (94.8)
Weighted <i>k</i> -means	0.982 (94.8)	0.936 (95.0)	0.939 (95.2)	0.939 (95.1)
Smart Pooling	0.962 (95.0)	0.482 (95.1)	0.480 (94.9)	0.479 (94.9)
Smart Selection	0.818 (94.7)	0.483 (95.2)	0.467 (94.6)	0.480 (95.2)
Random Pooling	0.475 (94.7)	0.479 (94.7)	0.483 (95.5)	0.483 (95.2)
<i>n</i> = 30				
Standard <i>k</i> -means	0.909 (95.2)	0.903 (94.8)	0.914 (95.1)	0.903 (95.3)
Weighted <i>k</i> -means	0.965 (94.8)	0.866 (94.7)	0.870 (95.0)	0.857 (94.7)
Smart Pooling	0.908 (95.0)	0.321 (95.0)	0.318 (95.0)	0.323 (95.1)
Smart Selection	0.657 (95.0)	0.324 (95.2)	0.303 (95.0)	0.319 (95.0)
Random Pooling	0.319 (95.1)	0.321 (95.3)	0.327 (95.7)	0.324 (95.2)
<i>n</i> = 16				
Standard <i>k</i> -means	0.847 (94.8)	0.842 (94.8)	0.851 (95.0)	0.846 (95.0)
Weighted <i>k</i> -means	0.937 (94.8)	0.764 (94.6)	0.795 (95.2)	0.764 (94.9)
Smart Pooling	0.802 (94.8)	0.208 (94.8)	0.206 (95.1)	0.207 (94.6)
Smart Selection	0.496 (94.7)	0.207 (95.2)	0.186 (95.1)	0.205 (95.0)
Random Pooling	0.207 (95.1)	0.208 (95.2)	0.209 (95.5)	0.207 (95.1)

[†]Total sample size is *N* = 240. All covariates were included in the Standard and Weighted *k*-means procedures; for weighted *k*-means, only *X*₁ was weighted. Smart Pooling and Smart Selection included only *X*₁ in the pooling process.

Table 4

Point estimates and standard errors (SE) from MLR on the BioCycle study dataset after k -means pooling under various numbers of pools. The first row displays regression estimates for the full dataset (unpooled), while the remaining rows provide estimates from pools formed under a global version of k -means clustering.

<i>n</i>	Estimate (SE)			
	log(estradiol)	BMI	Vitamin E	Age
240	−0.938 (0.471)	0.079 (0.078)	0.041 (0.037)	0.207 (0.037)
120	−0.879 (0.472)	0.071 (0.078)	0.045 (0.037)	0.209 (0.037)
60	−0.868 (0.494)	0.057 (0.081)	0.042 (0.038)	0.205 (0.039)
30	−0.729 (0.558)	0.121 (0.092)	0.031 (0.042)	0.194 (0.043)
16	−1.297 (0.540)	0.149 (0.085)	0.042 (0.039)	0.183 (0.040)