Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Xiaobo Yan Date

Design Strategies for Studies Using Logistic Regression to Analyze Data on Pooled Samples

By

Xiaobo Yan
MSPH
Biostatistics and Bioinformatics

_____

Robert H. Lyles, Ph. D
Thesis Advisor

_____

Amita Manatunga, Ph. D
Reader

Design Strategies for Studies Using Logistic Regression to Analyze Data on Pooled Samples

By

Xiaobo Yan
MSPH
Biostatistics and Bioinformatics

B.E
East China University of Science and Technology
2016

Thesis Advisor: Robert H. Lyles, Ph. D

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2021

Abstract

Design Strategies for Studies Using Logistic Regression to Analyze Data on Pooled Samples

By Xiaobo Yan

We review logistic regression modeling to estimate the risk of potential factors' odds ratios and predict disease prevalence using pooled samples via the maximum likelihood (ML) approach. We determine the preferred methods to deal with either categorical variables or continuous variables. For categorical variables, random pooling within subsets stratified by the variables of interest yields the most accurate and most efficient estimate on both coefficient and prevalence. We take advantage of statistical software for continuous variables to pool samples with a prespecified number of pools by the k-means clustering algorithm to optimize the estimation performance. We also modify the k-means clustering function embedded in SAS to constrain the maximum pool size to consider laboratory operability and test limitation. We compare the estimates between incorporating perfect and imperfect testing (sensitivity and sensitivity) to demonstrate the necessity of adjustment ML for test bias. Both of our proposed strategies showed the most efficacy while keeping good performance accuracy for the Malaria data and simulated data. Further potential study on imperfect tests is also discussed at the end of the study.

KEY WORDS: Pooled testing, Prevalence, Logistic regression, K-means clustering, Malaria.

Design Strategies for Studies Using Logistic Regression to Analyze Data on Pooled Samples

By

Xiaobo Yan
MSPH
Biostatistics and Bioinformatics

B.E
East China University of Science and Technology
2016

Thesis Advisor: Robert H. Lyles, Ph. D

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2021

Acknowledgements

This thesis would not have been possible without the inspiration and support of some remarkable individuals — my thankfulness and appreciation to all of them for being part of this journey and making this thesis possible.

I owe my deepest gratitude to my advisor Prof. Robert H. Lyles. Without his enthusiasm, encouragement, support, and continuous optimism, this thesis would hardly have been completed. His guidance and suggestion have been such valuable input for this thesis. Since the day I first met him on Zoom, he has been supportive of discussing my thesis's potential topics. Ever since, Prof. Lyles has supported me by providing a research assistantship over the past year and academically and emotionally through the rough road to finish this thesis. He helped me conquer the coding problems and gave me the freedom I needed to move on towards completing this thesis.

I express my warmest gratitude to Prof. Amita Manatunga for being my thesis reader. Her insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I gratefully acknowledge the contributions of Meghna Desai, Ya-Ping Shi, and Monica Shah from CDC for generously sharing the Malaria data. It's so fulfilling that I can apply my methodologies to solve the real problem.

It is a pleasure to thank my boyfriend Pengfei Cai for the wonderful times we shared, especially when I felt down during this challenging time.

Finally, my sincere gratitude is to my family for their continuous and unparalleled love, help, and support. I am forever indebted to my parents for giving me the opportunities and experiences that have made me who I am. They selflessly encouraged me to explore new directions in life and seek my own destiny. This journey would not have been possible if not for them, and I dedicate this milestone to them.

Table of Contents

1 Introduction

In disease surveillance and epidemiological studies, pooled testing has been widely suggested as an alternative to testing individual samples for saving laboratory capacity and reducing required numbers of diagnostic tests (Chang-Xing Ma, 2011). Pooled testing is a screening approach that combines several samples into one test, and which can be traced back to the mid 20th century when used to diagnose syphilis (Dorfman, 1943). Usually, if the pooled sample is tested negative, all individual specimens in the pool are attributed negative results, which can save the cost of testing each member at a time. However, if the pooled test is positive, each sample in the pool must be re-tested to identify the pool's positive sample(s). Recent applications involve COVID-19 screening and diagnostic testing (U.S. FDA, 2020) and it also has been successfully applied to prevalence estimation (Brynildsrud, 2020). Pooling is not a design novel to COVID-19, and pooled testing has been employed for many years to avoid infections being spread through the blood supply, as was HIV, hepatitis B, hepatitis C, West Nile, and Zika (CDC, 2020). Pooling has also been demonstrated to be an effective strategy when regression modeling is a focus of study under the scenarios in which outcome or predictive variables are measured on pooled biospecimens with or without measurement errors (Clarice R. Weinberg, 1999; S. Vansteelandt, 2000; Enrique F Schisterman, 2010; Emily M. Mitchell, 2014; Dane R Van Domelen, 2018), and it has been suggested that pooling strategies minimize the loss of information compared to other partition strategies such as random sampling (Enrique F. Schisterman, 2008).

Some studies have developed statistical methodologies under regression settings to address various cases. For example, Weinberg and Umbach (Clarice R. Weinberg, 1999) showed how to fit logistic regression models when a continuous exposure variable is measured in pools,

while Vansteelandt, Goetghebeur and Verstraeten (S. Vansteelandt, 2000) focused on the case where the binary outcome is determined in pools. A refined method of allocating pools based on a k-means clustering algorithm (Emily M. Mitchell, 2014) took computational advantage of software that can "perfectly" generate a prespecified number of pools comprised of individuals who are relatively homogeneous with respect to characteristics that could be associated with outcome status. However, the embedded k-means clustering functions in commercial software such as SAS (FASTCLUS procedure, 2020) is not able to constrain the pool size which may cause trouble in terms of feasibility and test effectiveness. A severe concern with large pool sizes is that any positive sample may be sufficiently diluted to become undetectable by the test when screened in pools, introducing false-negative bias to the results; this phenomenon termed a dilution effect (Lawrence M. Wein, 1996). Taking account of this test efficiency issue, in this thesis we propose a novel modification to the k-means clustering algorithm to incorporate a pool size limitation by specifying the largest acceptable size prior to the pool allocation stage. In our motivating study, a dataset containing 4670 individual subpatent malaria test results and corresponding pooled testing results on subjects from West Kenya is used to demonstrate the accuracy and efficacy of these various pooling strategies for estimating odds ratios associated with specified factors, as well as disease prevalence. Several simulation studies were also implemented to evaluate maximum likelihood-based estimators based on logistic regression applied to poolwise test results, comparing the precision attained by the alternative pooling strategies and adding support to our findings in the Malaria analysis.

2 Methodology

2.1 Standard multiple logistic regression

This study focuses on the binary outcome with either positive or negative results, such as being diagnosed with a disease or disease-free. To predict the disease prevalence (the probability of being tested positive) and estimate the factors that characterize the prevalence via odds ratios, we propose a logistic regression model to describe the relationship between multiple predictors and the prevalence and estimate odds ratios and prevalence through the maximum likelihood approach.

Logistic regression (LR) is the most common the regression analysis to address the case in which the outcome is dichotomous. First being devised to describe population growth and the progress of chemical reactions (Cramer, 2002), the LR model has been widely adopted in bioassay and statistical fields over time. The outcome Y has two levels (1 = event, 0 = event-free). Conditional on covariates, we assume Y follows a Bernoulli distribution with the probability of Y = 1, indicated by p, ranging from 0 to 1. In an epidemiological setting, Y usually represents disease status. Thus, p describes the disease prevalence for subjects with specified covariate values. Unlike multiple linear regression, which fits straight line or hyperplane surfaces to continuous outcome data, the logistic regression model utilizes the logit function to transform the linear equation's output between 0 and 1. The logit function is defined as:

$$logit(p_i) = ln\left(\frac{p_i}{1-p_i}\right) \quad (1)$$

Based on the logit transformation of p, the generalized multiple logistic regression (MLR) model can be expressed as:

$$logit(p_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \quad (2)$$

where $X_{i1} \ldots X_{ip}$ are a set of predictor variables of interest for subject i.

Since $ln\left(\frac{p_i}{1-p_i}\right)$ is the log odds that Y = 1 (given all X's), each parameter (β_j) in a first-order logistic model is interpreted as the log odds ratio for a 1-unit increase in X while keeping other X's constant, and its estimation can be fulfilled by the maximum likelihood method. To review, the log-likelihood of a MLR model with p predictors can be written as:

$$\ell(\beta_0, \beta_1, \ldots, \beta_p; y) = \sum_{i=1}^{n} \left(y_i ln(p_i) + (1 - y_i) ln(1 - p_i)\right) \quad (3)$$

By setting the (p+1) derivatives of the log-likelihood with respect to $\beta_0, \beta_1, \ldots, \beta_p$ equal to 0, we can numerically obtain the ML estimate of the vector of coefficients using standard software.

The method of estimating the variances and covariances of the estimated coefficients follows from well-developed theory of maximum likelihood estimation (Rao, 1973). According to this theory, the estimators are obtained from the matrix of second partial derivatives of the log-likelihood function (David W. Hosmer, 2013), which have the following general form,

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} X_{ij}^2 p_i (1 - p_i) \quad (4)$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} X_{ij} X_{ik} p_i (1 - p_i) \quad (5)$$

For j, k = 0, 1, 2…, p. Let the $(p + 1) \times (p + 1)$ matrix called the observed information matrix containing the negative of the terms given in equations (4) and (5) be denoted as $I(\boldsymbol{\beta})$. The variances and covariances of the estimated coefficients are obtained from the inverse of this matrix, which we denote as $Var(\boldsymbol{\beta}) = I^{-1}(\boldsymbol{\beta})$. It is difficult to write down the explicit

expression for the elements in this matrix except in special cases. We utilize software programs to obtain the observed information matrix and take square roots of the diagonal elements of the estimated variance-covariance matrix to estimate the standard errors associated with the collection of ($\widehat{\beta_0}, \widehat{\beta_1}, \ldots \widehat{\beta_p}$).

After fitting the model to data, the maximum likelihood estimates of p conditional on the covariate values can be estimated based on the regression model in model 2

$$\widehat{p_i} = \frac{1}{1 + e^{-\left(\widehat{\beta_0} + \sum_{j=1}^{p} \widehat{\beta_J} X_{ij}\right)}} \quad (6)$$

where $\widehat{\beta_J}$ is the maximum likelihood estimate calculated from equation 3.

2.2 Pooling strategies

In epidemiological studies requiring linear or logistic regression modeling, pooling has been demonstrated to be a valuable tool to constrain laboratory costs (Clarice R. Weinberg, 1999; Emily M. Mitchell, 2014). From a statistical standpoint, minimizing the loss of efficiency in estimating coefficients can be fulfilled by optimizing the pool design.

Vansteelandt, Goetghebeur, and Verstraeten (S. Vansteelandt, 2000) proposed a binary-outcome test on pools in the logistic regression setting. Coefficients corresponding to covariates and prevalence conditional on covariate values can be estimated via maximum likelihood (ML) theory that can adequately manage 1) multiple covariates, 2) different pool sizes, and 3) errors in test results (sensitivity and specificity). Vansteelandt's study illustrated that the pool design in terms of pool composition and size strongly impacted precision and cost-efficiency.

With regard to pool composition in multiple covariate settings, pools randomly composed on any given X would conceal X's effect and generate imprecise estimates (S. Vansteelandt, 2000). To generate X-homogeneous pools, VGV suggested to "… sort samples according to the

most important regressor X_j, next sorting the samples with equal X_j according to the second most important regressor X_k, and so on". This strategy can be perfectly applied to multiple categorical variables, and the sorting in terms of importance can be imposed based on the aims of the study. One can then perform random pooling within strata defined by single or multiple categorical covariates to assemble pools homogeneous on factors of interest.

However, when dealing with multiple continuous variables, creating strata by sorting covariates may not guarantee the homogeneity of all variables of interest. In Mitchell's study (Emily M. Mitchell, 2014), a novel application of k-means clustering was applied for allocation of specimens to pools in linear regression analysis with pooling to assess the continuous outcome. The k-means clustering algorithm classifies individual samples to a pre-specified number of clusters by maximizing the between-cluster weighted sum of squares (J.A. Hartigan, 1979). In other words, homogeneous groupings of individuals comprise the optimal pools. The main benefit of the k-means clustering is its flexibility when applied to the case in which multiple continuous predictors are of interest. The k-means algorithm provides an automated way for investigators to design the pool formulation. For example, k-means clustering can be applied to improve precision for estimating a single coefficient of primary interest, or to assign different weights to several coefficients. Figure 1 (Emily M. Mitchell, 2014) illustrates the pool allocation strategies of sorting and k-means clustering on one and two continuous variables, respectively. On the left panel, 40 observations (see figure's caption for the detailed generation procedure) are grouped into 5 clusters with an equal size of 8 after sorting by X_1, which can be viewed as the primary interest covariate (VGV). Thus, the five pools are relatively homogenous on X_1, while random concerning X_2. In contrast, the five pools on the right panel are

automatically generated by k-means clustering on both X_1 and X_2, promoting the efficient estimation of multiple covariates in regression settings.

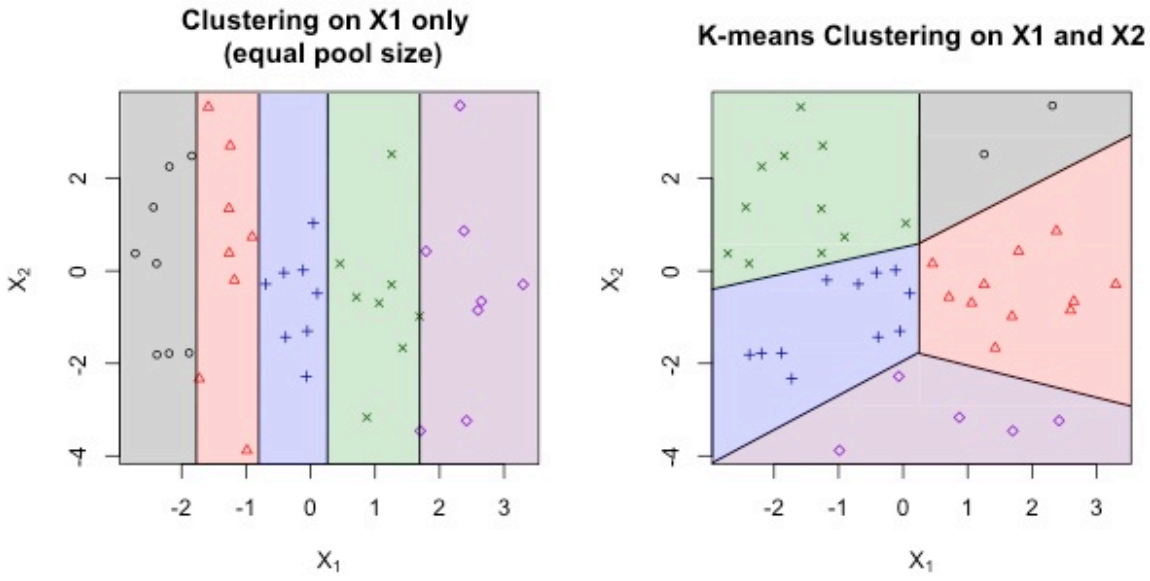

Figure 1 Contrasting pool allocations via sorting by a continuous covariate (X1) to form 5 pools of size 8 (left panel), vs. those based on k-means clustering applied to both continuous covariates (X1 and X2) to form 5 pools of varying size (right panel). The 40 (x_1, x_2) pairs were generated as follows: $X_1 \sim N(0,1)$, $X_2 = 0.25 - 0.5X_1 + 0.75Z$, $Z \sim N(0,1)$.

Although pool size has been a critical factor in the strategy performance, here we do not undertake a specific investigation on optimizing pool size. We note that satisfying a single user-prespecified pool size is easily achievable when sorting. However, we note that the built-in k-means functions available in standard software can only allow specifying the cluster number, while allowing cluster size to vary; this may be impractical in realistic lab assay settings. As such, one of our key objectives is to modify the original k-means function, to implement an approach we call “Controlled k-means”. The objective of this procedure is to control the maximum pool size at a level feasible for laboratory processing.

To conduct the "Controlled k means" algorithm for allocating subjects' specimens to pools, we begin by specifying the desired number of pools and the maximum allowable pool size. We then implement a standard k-means clustering function targeting the specified number of pools. The process of controlling the maximum pool size is fulfilled by iterating standard k-means clustering and dividing large clusters in a recursive manner, as presented in Figure 2. Original pools that are of the specified maximum size are set aside to be retained in the final pool allocation, while original pools larger than the specified maximum size are divided into one or more pools of that maximum size that are likewise set aside. The subjects left over are collected together with the subjects who comprised the original pools that were smaller than the specified maximum size. These individuals are then subjected to the next implementation of the k-means clustering function, targeting the original number of pools minus the number set aside, and we repeat all steps above recursively until we obtain the desired number of pools (the majority will typically be of the maximum size that was specified).

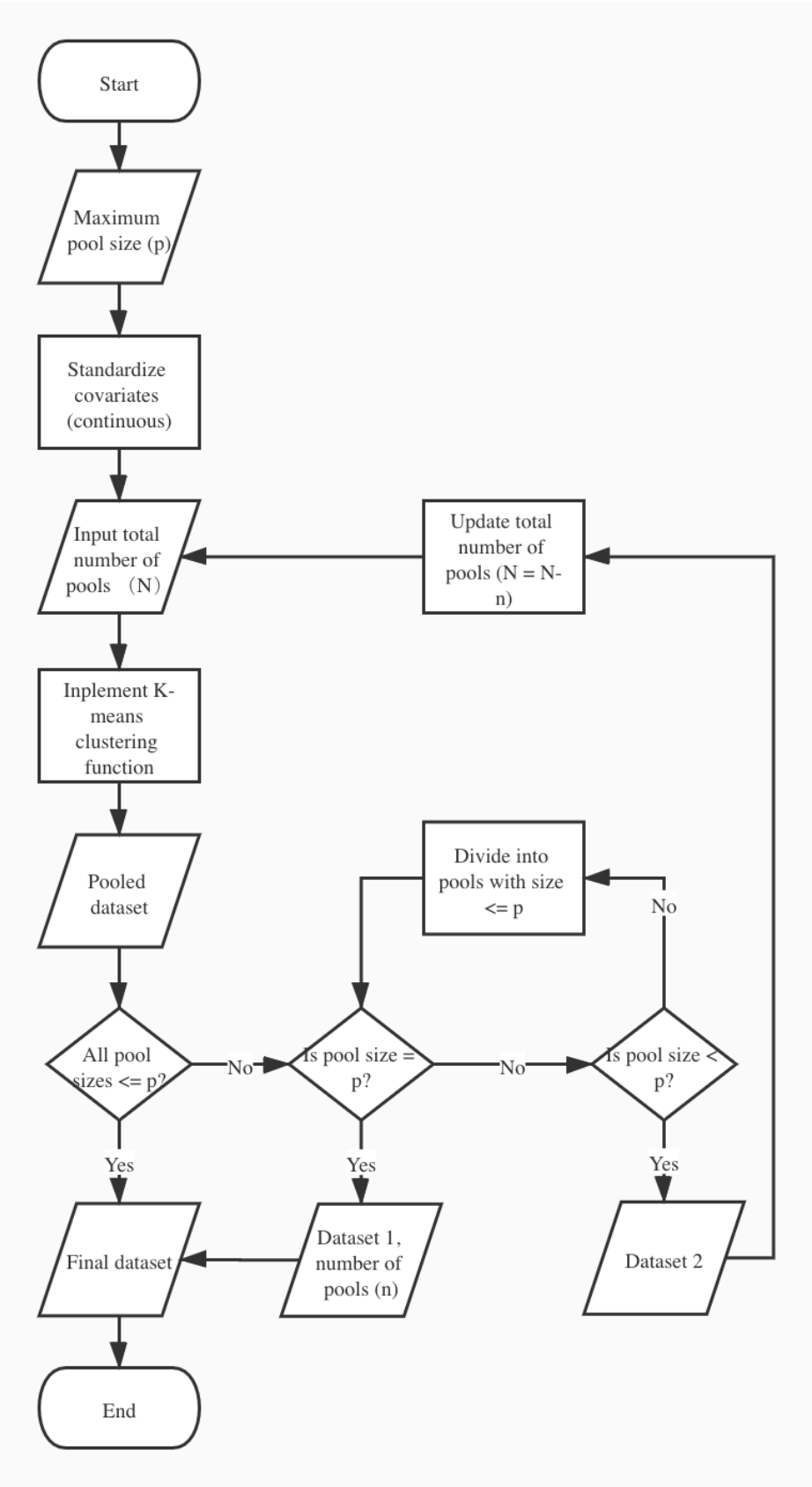

Figure 2 Flowchart of controlled K-means clustering process

2.3 Logistic regression in the pooling setting

Unlike testing on individuals, we can only observe pool-wise outcomes instead of individual results. In that case, the logistic regression modeling needs to be modified accordingly. Given a perfect test that does not introduce any bias, a pooled sample's positive result is interpreted as indicating at least one sample in the pool being positive. Intuitively, the negative result on pools indicates that all components are negative within the pool. Let Y_i denote the result of the i^{th} pool, where $Y_i = 0$ if the pool tests negative and $Y_i = 1$ if the pool tests positive. So, the observed data likelihood is given by

$$L = \prod_{i=1}^{k} (1 - \pi_{i0})^{y_i} \pi_{i0}^{(1-y_i)} \tag{7}$$

where

$$\pi_{i0} = \Pr\left(Pool\ i\ is\ negative\right) = \prod_{j=1}^{n_i} \left[1 + \exp\left(\beta_0 + \sum_{t=1}^{T} \beta_T x_{ijt}\right)\right]^{-1} \tag{8}$$

Considering the test's bias on the pool level, equation (8) can be modified to account for imperfect poolwise tests by introducing sensitivity (Se) and specificity (Sp). In the current work, we assume that both are known in advance and independent of pool size. The assumption also implies that the pooled test's sensitivity and specificity are approximately the same as those for an individual test (David W. Cowling, 1999). As shown by Vansteelandt et al. (2000), the modified probability of pool i being negative becomes:

$$\pi_{i0}^{*} = (1 - \mathrm{Se}) + (\mathrm{Se} + \mathrm{Sp} - 1) \times \prod_{j=1}^{n_i} \left[1 + \exp\left(\beta_0 + \sum_{t=1}^{T} \beta_T x_{ijt}\right)\right]^{-1} \tag{9}$$

We can still utilize maximum likelihood theory to numerically estimate odds ratios and prevalence under the modified logistic regression model, just as we can for poolwise logistic regression without misclassification and for the standard MLR model (Section 2.1).

3 Results

3.1 Motivational study

Malaria is a mosquito-borne disease caused by a parasite. Although preventable and curable, 409,000 people died of Malaria worldwide in 2019 (Parasites - Malaria, 2021), and 94% of cases and deaths were from the African region (Malaria, 2020). Subpatent malaria infections, which harbor infections at lower parasite densities, represent those below the detection threshold of microscopy or malaria rapid diagnostic testing (RDT)To identify subpatent infections, resource-intensive molecular tools such as polymerase chain reaction (PCR) are required. A previous sub-study (Aaron M Samuels, 2020; Meghna R. Desai, 2020; Shah) aimed to estimate the prevalence of PCR positive infections among RDT negative samples using a pooled testing strategy and provides a practical dataset upon which to apply our pooling strategies in a logistic regression setting. The dataset contains 4,670 RDT negative participants, together with the PCR test results obtained on all the individuals and on pools of size 5 (934 pools). According to WHO (Malaria, 2020), children aged under five years are the most vulnerable group affected by malaria. and the original pool allocations were randomly generated within each of three areas. To characterize the pooling test characteristics, we therefore focus on the participant's age and residence’s areas with different transmission intensity as the factors of key interest.

In Monica's study (Shah), the PCR test was conducted in Asembo (n = 1735), Gem (n = 2145), and Karemo (n = 790) between 2013 to 2015 in western Kenya. The percentages of positive subpatent parasitemia by individual PCR tests are 10.89% (Asembo), 14.13% (Gem),

and 18.35% (Karemo). The average age among the 4670 individuals was 28.19 years old, ranging from 0.08 to 93.59. In the current study, we consider treating age as both a continuous variable and a categorical variable (less than 5 years old, 5 to 15 years old and greater than 15 years old) to apply k-means clustering and sorting strategies, respectively. We initially fit the following two logistic regression models to the malaria data:

$$logit(p) = -1.525 + 0.001 \times Age - 0.604 \times Area1 - 0.307 \times Area2 \quad (10)$$

$$logit(p) = -1.424 - 0.759 \times Age1 + 0.572 \times Age2 - 0.603 \times Area1 - 0.293 \times Area2 \quad (11)$$

where "Age" in model 10 represented the continuous age, "Area1" stands for Asembo and "Area2" for Gem (with Karemo as the referent group). In model 11, "Age1" represented age less than 5 years old and "Age2" indicated age between 5 to 15 years old. The coefficients from models 10 and 11 were estimated from complete individual tests (n = 4670) containing all information from the study sample, so in that sense we regard them as the "ideal" coefficients. We note that model selection efforts leading to the two models above demonstrated no significant interaction between age and area.

3.1.1 Age as a categorical variable

Table 1 Coefficient estimates of logistic regression model when treating age as categorical variable

Estimates (SE)	Intercept	Age < 5	Age 5-15	Asembo	Gem
Complete sample (n = 4670)	-1.424 (0.095)	-0.759 (0.151)	-0.057 (0.104)	-0.603 (0.121)	-0.293 (0.112)
Random sampling (n = 934) (1)	-1.370 (0.204)	-0.817 (0.351)	-0.079 (0.240)	-0.859 (0.278)	-0.323 (0.243)
Random pooling (n = 934) (2)	-1.548 (0.213)	-0.381 (0.461)	-0.382 (0.242)	-0.748 (0.319)	-0.241 (0.284)
Random pooling on area strata with perfect pooled test result (n = 934) (3)	-1.309 (0.130)	-1.386 (0.811)	-0.281 (0.303)	-0.676 (0.148)	-0.362 (0.140)
Random pooling on area strata with actual pooled test result (n = 934) (4)	-1.871 (0.143)	-1.367 (0.954)	-0.418 (0.369)	-0.382 (0.163)	-0.183 (0.155)
Random pooling on area strata with actual pooled test result (adjusted for Se*, Sp*) (n = 934) (5)	-1.374 (0.186)	-2.051 (1.972)	-0.495 (0.449)	-0.466 (0.205)	-0.232 (0.197)
Random pooling on 9 (age, area) strata (n = 937**) (6)	**-1.405 (0.119)**	**-0.788 (0.170)**	**0.001 (0.127)**	**-0.653 (0.147)**	**-0.340 (0.138)**

* SE: the proportion of pools containing at least one true positive individual that tested positive; SP: the proportion of pools containing no true positive individuals that returned a negative pooled test result.
** Since the sample sizes of 6 (age, area) strata were not divisible by 5, the reminder (< 5) of each stratum was classified as a pool.

The estimates for each coefficient and corresponding standard errors from applying different sampling or pooling strategies are listed in Table 1. Those strategies considered include:

1) Random sampling: randomly selecting 934 individual samples (consistent with the number of pools with size 5) out of 4670 samples, using individual test result as binary outcome.
2) Random pooling: randomly dividing 4670 samples into 934 pools of size 5, using the poolwise result as a binary outcome. In this case, the poolwise result is inferred based on the known individual test results (which are assumed to be perfect).

3) Random pooling within area strata with perfect pool test result (henceforth referred to as "R.P. on area (perfect)"): randomly pooling samples into pools of size 5 within each area stratum, using perfect pooling test results as the binary outcome. In this case the poolwise test results are again perfect (no bias, as in (2) above).
4) Random pooling within area strata with actual pool test result (henceforth referred to as "R.P. on area (actual)"): This is the analysis of the actual randomly pooled samples into pools of size 5 within each area stratum as performed by the malaria study investigators. We expect bias here, since sensitivity (74.3% overall) and specificity (98.3% overall) were not perfect when applying the PCR assay to pools.
5) Random pooling within area strata with actual pool test result adjusted for Se, Sp (henceforth referred to as "R.P. on area (adjusted)"): randomly pooled samples into pools of size 5 within each area stratum were analyzed using actual pooling test results as the binary outcome but adjusting to account for the poolwise sensitivity (74.3%) and specificity (98.3%) in ML estimation (equation 9).
6) Random pooling on 9 (Age, Area) strata (henceforth referred to as "9 strata"): randomly pooling samples into pools of size 5 within 9 (age, area) strata, using perfect individual test results to infer the correct poolwise binary outcomes. In this case the poolwise test is perfect (no bias).

In Table 1 with the exception of methods 4 and 5, either individual test results or pooled test results were regarded as perfect. In other words, there was no misclassification in the test results. To distinguish from the actual pooled test result inherently contained in the malaria dataset (hereinafter called "actual pooled test results"), we manipulated "perfect pooled test results". The "perfect pooled test result" was assigned positive if at least one positive individual

existed in the pool and negative otherwise. Compared to random pooling, both the accuracy and efficiency of estimates delivered by random sampling were better. When we applied random pooling on each area stratum, all individuals in a pool were homogeneous on the area. The standard errors of the two area-covariates (Asembo, Gem) decreased significantly, becoming smaller than those obtained by random sampling. Intuitively, homogenous pools strengthened the information for a given covariate effect on the outcome, delivering more precise estimates. This conclusion could be further proved by random pooling within samples stratified by the two covariate factors (age and area) simultaneously. The standard errors of all four covariates (two age and two area coefficients) were smaller than those from either random pooling or random sampling. Random pooling on 9 (age, area) strata yielded the most efficient estimates by mining the precision benefits from the pooling strategy of producing covariate-homogeneous pools, and most closely approached the precision obtained in the complete sample individual-level analysis.

The malaria dataset provided a rare opportunity to dive into the imperfect test setting. It contained both individual test results and pooled (random pooling on area strata of size 5) test results. In this study, the individual test results were considered perfect. There is evidence, however, that the pooled test results were subject to some misclassification errors. The inconsistency between certain individual test results and the pooled test results containing those individual specimens was the source of bias in the regression analysis of the actual pool-wise test outcomes, as well as the source of information about the poolwisie sensitivity and specificity. In this case, sensitivity was defined as the proportion of pools testing positive given that they contained at least one true positive individual. Specificity was the proportion of pools testing negative given that they contained no true positive individuals. Overall, the estimated poolwise sensitivity and specificity in the malaria study were 74.3% and 98.3%, respectively. We inserted

these estimates into equation 9 to obtain the adjusted probability of a negative pool in the ML calculations. Comparing the estimates of regression coefficients between the "perfect" results and "actual" results, although area-homogeneous pools took some advantage of precision, the impact on the accuracy of estimates introduced by the “actual” test bias was non-neglectable and also contributed to imprecise estimates of disease prevalence. However, when we modified the likelihood function to account for the estimated sensitivity and specificity (here treating those values as known), we can see that both areas' coefficient estimates were pulled closer to the "true" values as opposed to the unadjusted ones.

Table 2 Prevalence estimates of 9 Malaria subsets stratified on age and area categories by different strategies

Prevalence (SE)	Asembo (p = 0.109)			Gem (p = 0.141)			Karemo (p = 0.184)		
Age	< 5	5 – 15	> 15	< 5	5 – 15	> 15	< 5	5 – 15	> 15
Complete sample (n = 4670)	0.058 (0.009)	0.123 (0.012)	0.117 (0.009)	0.078 (0.011)	0.160 (0.013)	0.152 (0.009)	0.101 (0.015)	0.203 (0.020)	0.194 (0.015)
Random sampling (n = 934)	0.045 (0.016)	0.091 (0.022)	0.097 (0.018)	0.075 (0.024)	0.145 (0.029)	0.155 (0.021)	0.101 (0.035)	0.190 (0.042)	0.203 (0.033)
Random pooling (n = 934)	0.064 (0.024)	0.129 (0.030)	0.091 (0.018)	0.103 (0.034)	0.197 (0.032)	0.143 (0.020)	0.127 (0.049)	0.238 (0.052)	0.175 (0.031)
R.P. on area (perfect) (n = 934)	0.033 (0.025)	0.094 (0.022)	0.121 (0.013)	0.045 (0.033)	0.124 (0.028)	0.158 (0.015)	0.063 (0.046)	0.169 (0.040)	0.213 (0.022)
9 strata (n = 937*)	**0.055 (0.009)**	**0.113 (0.013)**	**0.113 (0.010)**	**0.074 (0.011)**	**0.149 (0.015)**	**0.149 (0.011)**	**0.100 (0.017)**	**0.197 (0.025)**	**0.197 (0.019)**

* Since the sample sizes of 6 (age, area) strata were not divisible by 5, the reminder (< 5) of each stratum was classified as a pool.

The estimates of subpatent malaria prevalence were obtained by taking advantage of the SAS NLMIXED procedure (NLMIXED Procedure, 2020). Remarkably, the prevalence of a particular subset population is also readily estimable in the poolwise logistic regression setting. Since there are three levels of age and three areas, we have 9 (3×3) (age, area) strata in total, for which ML estimates of prevalence could be calculated from the poolwise logistic regression

model we fitted. We summarized the sub-prevalence of 9 strata attained by different strategies (Table 2) for comparison purposes. Compared to using complete samples, random pooling within the nine strata yields almost fully efficient estimates for prevalence.

Table 3 Prevalence estimates of 3 areas by complete sample and 4 strategies

Prevalence (SE)	Asembo (p = 0.109)	Gem (p = 0.141)	Karemo (p = 0.184)
Complete sample (n = 4670)	0.109 (0.007)	0.141 (0.008)	0.184 (0.014)
Random sampling (n = 934)	0.087 (0.015)	0.138 (0.017)	0.192 (0.030)
Random pooling (n = 934)	0.093 (0.017)	0.148 (0.016)	0.182 (0.030)
R.P. on area (perfect) (n = 934)	0.102 (0.008)	0.132 (0.008)	0.187 (0.017)
9 strata (n = 937*)	**0.103 (0.008)**	**0.134 (0.009)**	**0.183 (0.017)**

* Since the sample sizes of 6 (age, area) strata were not divisible by 5, the reminder (< 5) of each stratum was classified as a pool.

The prevalence of subpatent malaria in each area was estimated based on the individual test results in the malaria dataset. The estimates for the three areas and the standard errors were obtained from the following logistic regression model

$$logit(p) = \beta_0 + \beta_1 \times Asembo + \beta_1 \times Gem \tag{12}$$

which only contained reference coded indicator variables for two areas. The prevalence of each area was estimated by inserting the MLEs (equation 6) of the model coefficients into the following equations after fitting the model to data:

$$p(Asembo) = \frac{1}{1 + e^{-(\beta_0+\beta_1)}}$$

$$p(Gem) = \frac{1}{1 + e^{-(\beta_0+\beta_2)}}$$

$$p(Karemo) = \frac{1}{1+e^{-\beta_0}}.$$

The results were listed in Table 3. Either random sampling or random pooling were strategies that create a loss of accuracy and efficiency in the estimates. Especially for the high-risk area (Karemo), the large significant standard error reflects substantial uncertainty in the estimates.

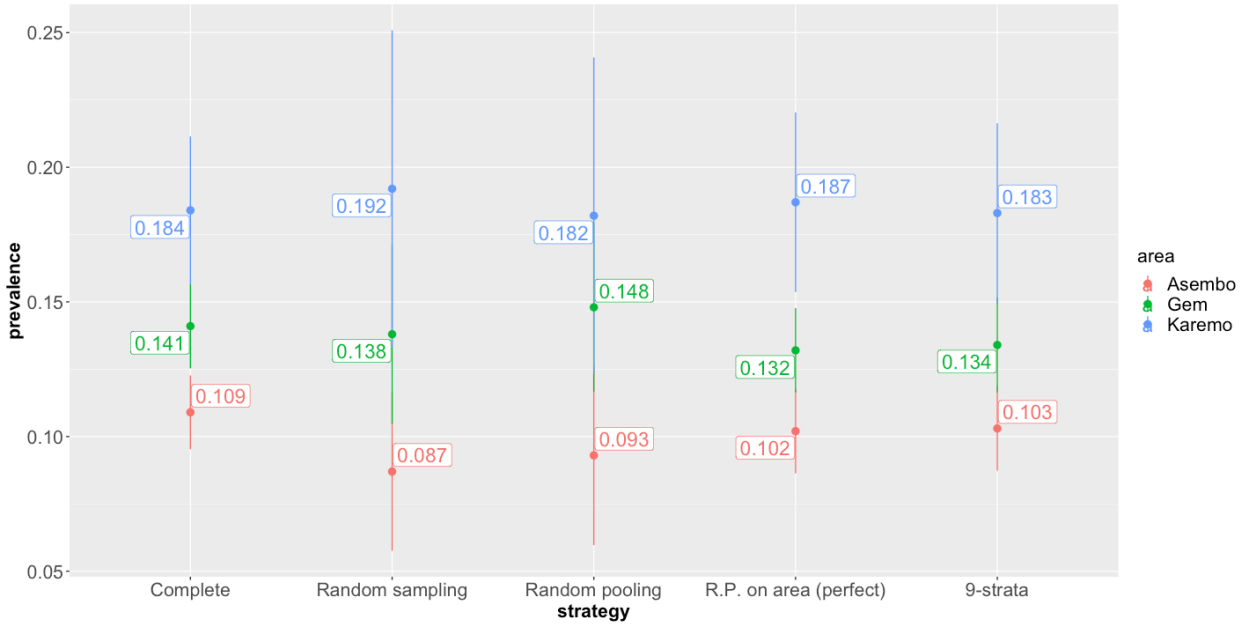

Figure 3 Comparison of point and 95% confidence interval estimates by complete sample and 4 strategies

Figure 3 displays both point estimates and 95% confidence intervals for the prevalence in the three areas derived from different sampling or pooling strategies. It was also noticeable that in the low-risk area (Asembo), the point estimates by random pooling on the area or random pooling within the 9 (age, area) strata were much more accurate than those obtained from random pooling or sampling. Concerning prevalence estimation, random pooling performances on the area and random pooling on 9-strata are pretty similar and far better than the arbitrary pooling or sampling, which generated heterogeneous pools.

3.1.2 Age as a continuous variable

Table 4 Coefficient estimates of logistic regression model when treating age as a continuous variable

Estimate (SD)	Intercept	Age	Asembo	Gem
Complete sample (n = 4670)	-1.525 (0.111)	0.001 (0.002)	-0.604 (0.120)	-0.307 (0.111)
Random sampling (n = 934) (1)	-1.552 (0.246)	0.003 (0.005)	-0.894 (0.277)	-0.368 (0.243)
Random pooling (n = 934) (2)	-1.622 (0.252)	0.005 (0.004)	-1.008 (0.350)	-0.219 (0.262)
Random pooling on area strata with perfect pool test result (n = 934) (3)	-1.612 (0.190)	0.004 (0.005)	-0.691 (0.147)	-0.398 (0.139)
Uncontrolled k-means on area strata (n = 934) (4)	-1.603 (0.153)	0.001 (0.002)	-0.482 (0.158)	-0.201 (0.151)
Controlled k-means on area strata with maximum pool size = 5* (n = 934) (5)	**-1.580 (0.136)**	**0.001 (0.002)**	**-0.507 (0.144)**	**-0.303 (0.138)**
Controlled k-means on area strata with maximum pool size = 6 (n = 934) (6)	-1.831 (0.143)	0.003 (0.002)	-0.468 (0.150)	-0.255 (0.142)
Controlled k-means on area strata with maximum pool size = 10 (n = 934) (7)	-1.691 (0.149)	0.004 (0.002)	-0.533 (0.155)	-0.264 (0.148)

* 934 pools of 5 are generated under this setting.

The estimate of each coefficient and its standard error from different sampling or pooling strategies were listed in Table 4. Those strategies considered include:

1) Random sampling: randomly selecting 934 individual samples (consistent with the number of pools with size 5) out of 4670 samples, using individual test result as binary outcome.
2) Random pooling: randomly dividing 4670 samples into 934 pools of size 5, using pooling test result as binary outcome. In this case, the poolwise result is inferred based on the known individual test results (which are assumed to be perfect).
3) Random pooling within area strata with perfect pool test result (henceforth referred to as "R.P. on area (perfect)"): randomly pooling samples into pools of size 5 within each area

stratum, using perfect pooling test results as the binary outcome and assuming the test is perfect (no misclassification bias).

4) Uncontrolled k-means within area strata (henceforth referred to as "Uncontrolled K-means"): implementing k-means clustering embedded in SAS 9.4 (FASTCLUS procedure, 2020) within each area stratum.
5) Controlled k-means on area strata with maximum pool size of 5 (henceforth referred to as "Controlled K-means (5)"): implementing the proposed controlled k-means approach with maximum pool size of 5 within each area stratum.
6) Controlled k-means on area strata with maximum pool size of 6 (henceforth referred to as "Controlled K-means (6)"): implementing the proposed controlled k-means approach with maximum pool size of 6 within each area stratum.
7) Controlled k-means on area strata with maximum pool size of 10 (henceforth referred to as "Controlled K-means (10)"): implementing the proposed controlled k-means approach with maximum pool size of 10 within each area stratum.

Similar to the findings conveyed by Table 1, random pooling on samples stratified by area still gains more precise estimates of the two area coefficients compared to completely random pooling. When we applied k-means on age within each area stratum, it was apparent that both accuracy and efficiency of the age coefficient increased significantly. Pooling samples by k-means clustering on a continuous variable provided an invaluable efficiency for the coefficient estimate, consistent in spirit with the findings of Mitchell et al. (2014). However, one of the main drawbacks of the original k-means implemented in SAS is that the maximum pool size is uncontrolled. Large pools could be unwieldy for the lab and could significantly decrease the test sensitivity (Maryza Graham, 2020), leading to biased conclusions. In the malaria dataset, the

pool size based on 4670 samples based on uncontrolled k-means specifying 934 pools ranged from 1 to 29. The proposed controlled k-means algorithm, in contrast, allows users to determine both the total number of pools and the maximum pool size to address this issue.

For the comparison, we fixed the total pool number as 934, which was consistent with the pool number in the random pooling strategies. When specifying the maximum pool size at 5, we obtained 934 pools of equal sizes of 5; this was because five is the average pool size for 4670 samples to generate 934 pools (4670/934 = 5). In other words, the algorithm could not create a pool with a size less than 5. Otherwise, another pool would contain more than five individuals. For practical application, the maximum pool size should be no less than the average pool size. We did not see a notable difference in point estimates and standard errors when comparing the estimates between uncontrolled k-means and the controlled versions. This indicates that the additional constraint on pool size did not harm the efficiency of estimated regression coefficients, while in practice it is likely to have great benefits in terms of feasibility in the lab and reducing the risk of misclassification bias.

3.2 Simulation

We used simulations to verify our proposed pooling strategies' performance on both categorical and continuous variables in logistic regression settings. We generated 2000 replications for both continuous age and categorical age variables, each generating 4670 individual observations under model 10 (continuous age) and model 11 (age categorized into three levels). In both simulations, all variables were generated independently for simplification.

3.2.1 Age as a categorical variable

In the categorized age setting, "Area" was generated from a tabled probability distribution (0.37, 0.46, 0.17), representing the proportion of Asembo (low risk area), Gem (mid risk area), and Karemo (high risk area) in the Malaria dataset. Two dummy variables were created for Asembo and Gem, respectively. "Age" was generated from a tabled probability distribution (0.16, 0.21, 0.63), representing the proportion of less than five years old, 5 to 15 years old, and greater than 15 years old in the Malaria dataset. Two dummy variables were created for less than five years old and 5 to 15 years old. The true coefficient vector was (-1.4, -0.8, 0.06, -0.6, -0.3).

The simulation results to compare different pooling strategies were summarized in Table 5. Also, the imperfect test and adjusted test results were included as well. To manipulate the biased pooled test result, we generated a new variable accounting for sensitivity (74.3%) and specificity (98.3%). If the "perfect" pooled test result is positive (at least one individual is positive), the "imperfect" result will be assigned a random value (1/0) from a binary distribution with a probability of 0.743. If the "perfect" test result is negative (no individual is positive), the "imperfect" result will be assigned a random value (0/1) from a binary distribution with a probability of 0.983.

Compared to completely random pooling, random sampling delivered more efficient estimates (smaller MSE). However, applying random pooling on samples stratified by area first, the two area factors' coefficients were much more accurate than with random sampling. This conclusion is consistent with the Malaria dataset. Similarly, the 9-strata pooling strategy yielded the most efficient and precise estimates of coefficients for the four predictors, demonstrating that allocating samples to predictor-homogeneous pools maximizes the precision benefits in logistic regression settings. Regarding imperfect test settings, if we did not adjust the data likelihood

function accordingly for test bias (incorporating Se and Sp), we found the covariate estimates deviated noticeably on average from the true values. However, this deviation could be reduced by plugging in test sensitivity and specificity in the likelihood function to estimate the coefficients. Those estimates/SD/MSE marked in red look implausible and will be further discussed in the Discussion section.

Table 5 Simulation estimates of logistic regression model when treating age as categorical variable

Mean estimate (SD) [MSE]	Intercept $\beta_0 = -1.4$	Age < 5 $\beta_1 = -0.8$	Age 5 – 15 $\beta_2 = 0.06$	Low-risk area $\beta_3 = -0.6$	High-risk area $\beta_4 = -0.3$
Complete sample (n = 4670)	-1.403 (0.095) [0.097]	-0.810 (0.150) [0.151]	0.060 (0.099) [0.103]	-0.597 (0.119) [0.120]	-0.301 (0.112) [0.111]
Random sampling (n = 934)	-1.413 (0.223) [0.219]	-0.837 (0.360) [0.347]	0.056 (0.235) [0.231]	-0.600 (0.273) [0.271]	-0.294 (0.252) [0.251]
Random pooling (n = 934)	-1.408 (0.219) [0.216]	-1.079 (1.601) [12.813]	0.037 (0.265) [0.259]	-0.600 (0.316) [0.306]	-0.296 (0.269) [0.267]
R.P. on area (perfect) (n = 934)	-1.396 (0.139) [0.141]	-1.080 (1.592) [10.713]	0.045 (0.252) [0.262]	-0.602 (0.142) [0.145]	-0.305 (0.138) [0.137]
R.P. on area (actual) (n = 934)	-1.894 (0.153) [0.154]	-0.893 (1.536) [9.913]	0.026 (0.295) [0.296]	-0.489 (0.159) [0.159]	-0.242 (0.150) [0.149]
R.P. on area (adjusted) (n = 934)	-1.395 (0.200) [0.202]	-1.436 (2.634) [34.329]	0.030 (0.295) [0.296]	-0.611 (0.204) [0.204]	-0.313 (0.197) [0.196]
9 strata (n*)	**-1.400 (0.121) [0.122]**	**-0.808 (0.167) [0.166]**	**0.064 (0.123) [0.124]**	**-0.600 (0.144) [0.146]**	**-0.304 (0.139) [0.138]**

* The pool numbers are varying because the number of strata which is divisible by 5 in each replication are different.

Similarly, the estimated prevalence for the 9 (area, age) strata and empirical standard deviations and MSEs from 2000 simulations were summarized in Table 6. Estimation yielded by random pooling within the 9-strata was most efficient in all subsets and came closest to competing with complete sampling. The result was consistent with that obtained using the actual Malaria data. In contrast with Table 5, we did not observe any doubtful estimates for prevalence.

Table 6 Simulated prevalence estimates of 9 subsets stratified on age and area categories by different strategies

Prevalence (SD) [MSE]	Low risk area			Mid risk area			High risk area		
Age	< 5	5 – 15	> 15	< 5	5 – 15	> 15	< 5	5 – 15	> 15
Complete sample (n = 4670)	0.057 (0.008) [0.008]	0.126 (0.012) [0.012]	0.120 (0.009) [0.009]	0.076 (0.010) [0.010]	0.162 (0.013) [0.013]	0.154 (0.009) [0.009]	0.100 (0.015) [0.015]	0.208 (0.020) [0.020]	0.198 (0.015) [0.015]
Random sampling (n = 934)	0.058 (0.019) [0.019]	0.126 (0.027) [0.027]	0.119 (0.019) [0.019]	0.076 (0.024) [0.023]	0.163 (0.030) [0.030]	0.155 (0.020) [0.020]	0.100 (0.033) [0.032]	0.208 (0.045) [0.044]	0.198 (0.035) [0.034]
Random pooling (n = 934)	0.057 (0.026) [0.025]	0.125 (0.030) [0.029]	0.120 (0.021) [0.020]	0.075 (0.033) [0.032]	0.161 (0.031) [0.031]	0.155 (0.020) [0.019]	0.099 (0.044) [0.043]	0.206 (0.046) [0.045]	0.199 (0.034) [0.034]
R.P. on area (perfect) (n = 934)	0.057 (0.025) [0.025]	0.126 (0.023) [0.023]	0.120 (0.012) [0.012]	0.075 (0.032) [0.031]	0.162 (0.027) [0.028]	0.155 (0.014) [0.014]	0.099 (0.042) [0.041]	0.208 (0.036) [0.037]	0.199 (0.022) [0.022]
9 strata (n*)	**0.058 (0.009) [0.009]**	**0.127 (0.014) [0.014]**	**0.120 (0.010) [0.010]**	**0.076 (0.011) [0.011]**	**0.163 (0.016) [0.016]**	**0.154 (0.011) [0.011]**	**0.100 (0.017) [0.017]**	**0.209 (0.025) [0.025]**	**0.198 (0.019) [0.019]**

* The pool numbers are varying because the number of strata which is divisible by 5 in each replication are different.

3.2.2 Age as a continuous variable

The area's settings were the same in the continuous age model, but we generated age ~ N (28.19, 22.32^2), where 28.19 was the mean and 22.32 was the standard deviation based on the 4670 individual samples in the Malaria dataset. The true coefficient vector was (-1.5, 0.001, -0.6, -0.3). The mean estimate of coefficients with empirical standard deviation and MSE were listed in Table 7. Besides the precision benefits gained from the stratification by area, k-means clustering significantly increased the precision of estimates for the age coefficient. Even though the maximum pool size of 6 was specified (controlled k-means), it did not reduce the efficiency of the age coefficient estimate relative to uncontrolled k-means and maintained good performance in terms of the area factor coefficients. The simulation results show that controlled

k-means yielded the most efficient estimation of coefficients for both continuous and categorical variables in this logistic regression modeling setting.

Table 7 Simulation estimates of logistic regression model when treating age as a continuous variable

Mean estimate (SD) [MSE]	Intercept $\beta_0 = -1.5$	Age $\beta_1 = 0.001$	Low risk area $\beta_2 = -0.6$	Mid risk area $\beta_3 = -0.3$
Complete sample (n = 4670)	-1.502 (0.109) [0.106]	0.001 (0.002) [0.002]	-0.597 (0.116) [0.119]	-0.299 (0.110) [0.110]
Random sampling (n = 934)	-1.515 (0.245) [0.240]	0.001 (0.004) [0.004]	-0.589 (0.277) [0.269]	-0.297 (0.251) [0.248]
Random pooling (n = 934)	-1.513 (0.249) [0.247]	0.001 (0.005) [0.005]	-0.604 (0.306) [0.308]	-0.299 (0.263) [0.268]
R.P. on area (perfect) (n = 934)	-1.508 (0.186) [0.184]	0.001 (0.005) [0.005]	-0.602 (0.143) [0.144]	-0.301 (0.137) [0.136]
Uncontrolled K-means (n = 934)	-1.499 (0.142) [0.141]	0.001 (0.002) [0.002]	-0.599 (0.154) [0.157]	-0.300 (0.150) [0.149]
Controlled K-means (6) (n = 934)	**-1.497 (0.138) [0.134]**	**0.001 (0.002) [0.002]**	**-0.601 (0.146) [0.148]**	**-0.301 (0.141) [0.140]**

4 Discussion

4.1 Imperfect tests

In an epidemiological study, the test's bias (sensitivity and specificity) is not neglectable because of its impact on estimate precision. In table 1, all covariate estimates based on actual test results (method 4) depart further from the ideal estimates compared to those based on perfect test results (method 3). The assumptions we made for the imperfect test include (1) both Se and Sp are known in advance and independent of pool size, (2) the sensitivity and specificity of the pooled test are approximately the same as it is for an individual test (David W. Cowling, 1999). In the Malaria data, we applied Se and Sp calculated from pooled test results to all pools regardless of size. The accuracy of estimates improved when we adjusted the data likelihood for the "rough" Se, Sp (method 5 in Table 1), indicating that accounting for the test bias estimating process benefits the precision. A former study (Claudia Muñoz-Zanzi, 2006) has shed light on

the fact that Se and Sp would be influenced by the pool size and the number of negative individual samples in the pool, providing an idea for a more precise estimation of Se, Sp if more information about the pools is known in advance.

4.2 Overall prevalence

In this study, we focused on estimating the Malaria prevalence in each of the three areas via the logistic regression model that only contained area-covariates. From figure 1, we could see that the random pooling within 9-strata performed as well as random pooling within each area on both point estimates and interval estimates. One advantage of random pooling within 9-strata is that it yields more precise and more efficient estimates within each subgroup stratified by the covariates specified (Table 1), which might benefit those interested in a disease's subgroup prevalence. To estimate the overall prevalence (across all 3 areas) and its variance based on the area-specific prevalence estimates, a straightforward approach weighs each of the three area-specific estimates by the proportion (w) of the total population residing in that area. The overall prevalence and its variance become:

$$\hat{P}_{overall} = w_{Asembo}\hat{P}_{Asembo} + w_{Gem}\hat{P}_{Gem} + w_{Karemo}\hat{P}_{Karemo}$$

$$Var(\hat{P}_{overall}) = w_{Asembo}^2 SE(\hat{P}_{Asembo})^2 + w_{Gem}^2 SE(\hat{P}_{Gem})^2 + w_{Karemo}^2 SE(\hat{P}_{Karemo})^2$$

4.3 Investigating implausible simulation results

In the simulation study, we noticed that some estimates/SD/MSE attained by random pooling and random pooling on area-strata (based on three ways to deal with the outcome: assuming perfect test, using actual test result, and adjusted with Se/Sp) seemed implausible (highlighted in red in Table 5). After checking the code and ensuring no flaws in the simulation process, we believe this was due to the coefficient setting since we use the estimates from the logistic regression model based on complete samples to imitate the "true" situation. Then we

took a more in-depth look at how the implausible results were generated. We extracted all datasets and listed the coefficient estimates together with the standard errors from 2000 replicates and sorted by the magnitude of the estimated coefficient for age less than 5. The ten most extreme observations yielded by random pooling and the three random pooling on area strata strategies were shown in Table 8. While the actual coefficient was set as -0.8, it was clear that the biased estimates and large SDs and MSEs in Table 5 were due to those simulations yielding extreme coefficient estimates for age less than 5 with high standard errors.

Table 8 Ten most extreme values from R.P. and three R.P. on area strata with different setting with the outcomes

Estimate for Age < 5 (SD)	Random pooling	R.P. on area (perfect)	R.P. on area (actual)	R.P. on area (adjusted)
		$\beta_1 = -0.8$		
1	-16.46 (1798.18)	-16.24 (1823.13)	-15.76 (1222.64)	-17.00 (2176.21)
2	-16.05 (2054.93)	-16.22 (1112.22)	-15.74 (1457.17)	-16.80 (2584.98)
3	-15.94 (1694.31)	-15.89 (1516.55)	-14.86 (578.42)	-16.63 (2137.67)
4	-15.56 (1488.84)	-15.58 (1042.90)	-14.72 (606.82)	-16.54 (2086.57)
5	-15.223 (1844.38)	-15.52 (777.98)	-14.36 (787.95)	-16.47 (2098.59)
6	-14.99 (860.09)	-15.45 (941.96)	-14.20 (1143.62)	-16.12 (1800.71)
7	-14.90 (1116.26)	-15.15 (671.12)	-14.09 (1294.26)	-15.42 (681.09)
8	-14.68 (716.98)	-14.97 (1010.98)	-13.97 (723.84)	-15.23 (1012.26)
9	-14.38 (892.59)	-14.959 (764.78)	-13.95 (1087.36)	-15.18 (855.60)
10	-14.38 (583.18)	-14.82 (698.87)	-13.94 (608.11)	-15.14 (712.43)

To double confirm the simulation process was reasonable, we checked the medians of the 2000 estimates for age less than 5 from the four strategies and recalculated the mean, SD, and MSE after removing the most extreme 5% (100 replications). The results were summarized in Table 9. Since the median is less susceptible to extreme values than the mean, we can see that the medians of all four strategies were much closer to the true coefficient (-0.8) compared to the

means. Alternatively, removing the 5% most extreme estimates yielded results comparable to the medians of complete simulation. In all, the unreasonable estimates in

Table 5 were due to the extremes among the 2000 simulations. Regardless, we note that the recommended strategy of random pooling on 9 (age, area) strata yielded the best performance in terms of both accuracy and precision in Table 5 and was not subject to any such extreme results.

Table 9 Medians of coefficient for age < 5 from 2000 simulations and the Mean/SD/MSE of coefficient for age < 5 from 1900 simulations (after removing most 5% extremes)

Simulation number	Statistics	Random pooling	R.P. on area (perfect)	R.P. on area (actual)	R.P. on area (adjusted)
2000	Median	-0.818	-0.814	-0.651	-0.815
1900	Mean (SD) [MSE]	-0.835 (0.470) [0.538]	-0.834 (0.480) [0.553]	-0.659 (0.491) [0.561]	-0.891 (0.741) [0.918]

5 References

[1] Samuels, A. M., Odero, N. A., Odongo, W., Otieno, K., Were, V., Shi, Y. P., Sang, T., Williamson, J., Wiegand, R., Hamel, M. J., Kachur, S. P., Slutsker, L., Lindblade, K. A., Kariuki, S. K., & Desai, M. R. (2020). Impact of Community-Based Mass Testing and Treatment on Malaria Infection Prevalence in a High-Transmission Area of Western Kenya: A Cluster Randomized Controlled Trial. *Clinical Infectious Diseases*.

[2] Brynildsrud, O. (2020). COVID-19 prevalence estimation by random sampling in population - optimal sample pooling under varying assumptions about true prevalence. *BMC Medical Research Methodology*, 196 - 203.

[3] *CDC*. (2020, Mar. 18). Retrieved from Blood Safety Basics: https://www.cdc.gov/bloodsafety/basics.html

[4] Chang, X. M., Albert, V., Enrique, F. S. & Lili, T. (2011). Cost-efficient designs based on linearly associated biomarkers. *Journal of Applied Statistics, 38*, 2739-2750.

[5] Clarice, R. W., & David, M. U. (1999). Using Pooled Exposure Assessment to Improve Efficiency in Case-Control Studies. *Biometrics*, 718 - 726.

[6] Claudia, M., Mark, T., Sharon, H., & Wesley, J. (2006). Factors affecting sensitivity and specificity of pooled-sample testing for diagnosis of low prevalence infections. *Prev Vet Med*, 309 - 322.

[7] Cramer, J. S. (2002). The origins of logistic regression. *Tinbergen Institute Working Paper*.

[8] Dane, R. V. D., Emily, M. M., Neil, J. P., Enrique, F. S., Amita, K. M., Yijian, H., & Robert, H. L. (2018). Logistic regression with a continuous exposure measured in pools and subject to errors. *Stat Med*, 4007 - 4021.

[9] David, W. C., Ian, A. G., & Wesley, O. J. (1999). Comparison of methods for estimation of individual-level prevalence based on pooled samples. *Preventive Veterinary Medicicine*, 211 - 255.

[10] David, W. H., Stanley, L., & Rodney, X. S. (2013). *Applied Logistic Regression.* Hoboken, New Jersey: John Wiley & Sons, Inc.

[11] Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics*, 436 - 440.

[12] Emily, M. M., Robert, H. L., Amita K. M., Neil J. P., & Enrique F. S. (2014). A highly efficient design strategy for regression with outcome pooling. *Statistics in Medicine*, 5028 - 5040.

[13] Enrique, F. S., Albert, V., Sunni, L. M., & Neil J. P. (2010). Hybrid pooled-unpooled design for cost-efficient measurement of biomarkers. *Stat Med*, 579 - 613.

[14] Enrique, F. S., & Albert, Vexler (2008). To pool or not to pool, from whether to when: applications of pooling to biospecimens subject to a limit of detection. *Paediatr Perinat Epidemiol*, 486 - 496.

[15] *FASTCLUS procedure*. (2020, Oct. 8). Retrieved from SAS Help center: https://documentation.sas.com/doc/en/statug/15.2/statug_fastclus_toc.htm

[16] Hartigan, J., & Wong, M. (1979). A K-Means Clustering Algorithm. *Applied Statistics*, 100 - 108.

[17] Lawrence, M. W., & Stefanos, A. Z. (1996). Pooled Testing for HIV Screening: Capturing the Dilution Effect. *Operations Research*, 543 - 569.

[18] *Malaria*. (2020, Nov. 30). Retrieved from WHO: https://www.who.int/news-room/fact-sheets/detail/malaria

[19] Maryza, G., Eloise, W., Nicole, I., Eka, B., Tebuka, T., Julian, D., Mike, C., Chantel, L., Benjamin, P. H., & Deborah, A. W. (2020). Sample pooling on the Cepheid Xpert® Xpress SARS-CoV-2 assay. *Diagnostic microbiology and infectious disease*, https://doi.org/10.1016/j.diagmicrobio.2020.115238.

[20] Meghna, R. D., Aaron, M. S., Wycliffe, O., John, W., Nobert, A. O. Kephas, O., Ya, P. S., Stephen, P. K., Mary, J. H., Simon, K., & Kim, A. L. (2020). Impact of Intermittent Mass Testing and Treatment on Incidence of Malaria Infection in a High Transmission Area of Western Kenya. *The American Journal of Tropical Medicine and Hygiene*, 369 - 377.

[21] *NLMIXED Procedure*. (2020, Oct. 28). Retrieved from SAS Help center: https://documentation.sas.com/doc/en/statug/15.2/statug_nlmixed_toc.htm

[22] *Parasites - Malaria*. (2021, Feb. 17). Retrieved from CDC: https://www.cdc.gov/parasites/malaria/index.html

[23] Rao, C. R. (1973). *Linear Statistical Inference and its Application.* New York: Wiley Inc.

[24] Vansteelandt, S., Goetghebeur, E., & Verstraeten, T. (2000). Regression Models for Disease Prevalence with Diagnostic Tests on Pools of Serum Samples. *Biometrics*, 1126 - 1133.

[25] Shah, M. P. Unpublished data.

[26] *U.S. FDA*. (2020, Aug. 24). Retrieved from Pooled Sample Testing and Screening Testing for COVID-19: https://www.fda.gov/medical-devices/coronavirus-covid-19-and-medical-devices/pooled-sample-testing-and-screening-testing-covid-19