Comparison of Methods for Clustered Data Analysis in a Non-Ideal Situation: Results from an Evaluation of Predictors of Yellow Fever Vaccine Refusal in the Global TravEpiNet (GTEN) Consortium

Sowmya R. Rao[1,2,*], Regina C. LaRocque[3,4], Emily S. Jentes[5], Stefan H.F. Hagmann[6,7], Edward T. Ryan[3,4], Pauline V. Han[5], David G. Kleinbaum[8] and Global TravEpiNet Consortium

[1]Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA; [2]Center for Healthcare Organization and Implementation Research (CHOIR), Bedford VA Medical Center, Bedford, MA; [3]Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA; [4]Department of Medicine, Harvard Medical School, Boston, MA; [5]Division of Global Migration and Quarantine, Centers for Disease Control and Prevention, Atlanta, GA; [6]Division of Pediatric Infectious Diseases, Bronx Lebanon Hospital Center, Bronx, NY, USA; [7]Department of Pediatrics, Albert-Einstein College of Medicine, Bronx, NY, USA; [8]Division of Healthcare Quality and Promotion, Centers for Disease Control and Prevention, Atlanta, GA, USA

Abstract: Not accounting for clustering in data from multiple centers might yield biased estimates and their standard errors, potentially leading to incorrect inferences. We fit 15 different models with different correlation structures and with/without adjustment for small clusters, including unadjusted logistic regression, Population-averaged models (Generalized Estimating Equations), Cluster-specific models (linear and non-linear with random intercept) and Survey data analysis methods to study the association of variables with the probability of declining yellow fever vaccine among patients seeking pre-travel health consultations at 18 US practices in the Global TravEpiNet Consortium from 1 January, 2009, to 6 June, 2012. Results varied by the method chosen. Generally, when the odds ratio estimates were similar, adjusting for clustering and the small number of clinics increased the standard errors. We chose the random intercept model with the Morel, Bokossa and Neerchal (MBN) adjustment to be the most preferable method for the GTEN dataset since this was one of the more conservative models that accounted for clustering, small sample sizes and also the random effect due to site. Investigators should not ignore clustering and consider the appropriate adjustments necessary for their studies.

Keywords: Clustering, cluster size, cluster imbalance, data analysis.

INTRODUCTION

Clinical trials, observational studies, and health services research give rise to clustered data when information is collected repeatedly on a single unit of analysis (e.g., clinics). Examples can be found in longitudinal studies with repeated measurements taken on the same individual (e.g., before and after an intervention, subjects followed over time), multi-center studies (data collected at various sites/clinics), or data with a hierarchical structure (e.g., patients nested within physicians nested within clinics). Such data are inherently correlated, and failing to account for this could bias estimates and their standard errors, potentially leading to incorrect inferences.

A number of approaches can account for clustering in data analysis [1-3]. Some require the covariance structure of the clusters to be specified, while others require the clustering variable to be entered in the model as a random effect. It is well known that if the number of clusters is large (≥30) and they are balanced in size, all the approaches are robust even when the covariance structure is misspecified. Although such approaches are being increasingly used to design studies, calculate statistical power, analyze data, and interpret results from studies involving clustering, careful attention is not always paid to the number or size of clusters and/or whether the clusters are balanced.

Investigators need greater guidance and awareness of approaches to address non-ideal situations, such as small numbers of clusters and/or imbalanced clusters, in multi-center studies. Although this manuscript is neither an exhaustive review of methods nor a tutorial, it demonstrates the effect of choosing different data analysis approaches. We discuss the various clustering adjustment methods and the effect of a small number of clusters and an imbalance in the size of those clusters on the estimates and their standard errors. We describe the methods briefly; an in-depth discussion of the technical details can be found elsewhere [1-13]. We compare numerical results obtained from 15 different

*Address correspondence to this author at the Department of Quantitative Health Sciences, University of Massachusetts Medical School, 368 Plantation Street, Worcester, MA 01605-2324, USA; Tel: 508-856-4046;
Fax: 508-856-8993; E-mail: sowmya.rao@umassmed.edu

Table 1: Distribution of Patients by Site

Site Number	Number of Patients	Number of Patients Recommended Yellow Fever Vaccine
1	6782	1579
2	16	5
3	240	67
4	3718	921
5	967	257
6	558	378
7	1448	193
8	269	83
9	138	43
10	996	227
11	781	165
12	688	177
13	192	52
14	262	72
15	674	184
16	550	254
17	86	26
18	526	132
Total	18891	4815

models to study the association of variables with the probability of declining the yellow fever vaccine among patients who had pre-travel health consultations at 18 US practices in the Global TravEpiNet Consortium (GTEN) from 1 January, 2009, to 6 June, 2012. The number of patients at sites ranged from as small as 16 patients to as large as 6782 patients. The analysis sample consisted of 4815 patients who were recommended to receive the yellow fever vaccine. This restricted sample was distributed among the 18 practices in sizes of 5 to 1579 (Table **1**).

METHODS

Study Population

GTEN is a national consortium of travel medicine practices across the United States that care for international travellers seeking pre-travel health consultation. The consortium was established in 2009 and has been described in detail elsewhere [14]. The GTEN data collection protocol was approved by the institutional review board of all participating sites.

Data collected from patients include demographics (e.g., age, gender), travel information (e.g., date, duration, destination, purpose of travel, and type of accommodation), and health information (e.g., medical history, medication usage, immunizations, vaccines administered, and reasons for not administering vaccines if indicated). These data are collected by using a secure online tool.

Regarding country-specific yellow fever (YF) risk, certain countries are considered to be entirely endemic for YF, while others are deemed partially endemic. Even though the YF vaccine is recommended for patients traveling to these countries, some patients refuse the vaccine for themselves and/or for their children. We were interested in identifying the factors associated with YF vaccine refusal among patients recommended to receive the vaccine. The independent variables assessed included age (continuous), duration of travel (dichotomous; <29, ≥29 days), VFR status (whether the patient was visiting family and relatives in the region of origin of self or family in a low or low-middle income country as defined by the World Health

Organization's 2011 Human Development Index) [15-17]) (dichotomous), and the destination country endemic for YF (dichotomous). We obtained adjusted odds ratio estimates and associated standard errors from 15 different multivariable logistic regression models fit to the GTEN data, using SAS 9.3 (SAS Institute, Cary, NC) or SUDAAN 11.0.0 (RTI International, Cary, NC).

Analysis Methods

The correlation in our data arises from the subjects being clustered within clinics rather than due to repeated observations on the same subject. Cluster-specific models using random effects, population-averaged models using Generalized Estimating Equations (GEE), and survey data analysis methods are some of the popular methods to analyze clustered data. We applied these models with and without appropriate adjustments to evaluate the association of covariates to the probability of a patient's refusing the yellow fever vaccine.

1. Standard Logistic Regression Model

We define a binary outcome variable Y by Y_{ij}= 1 if the j^{th} subject within the i^{th} site declines the vaccine and Y_{ij}= 0 otherwise, where i=1,2,…,K and j=1,2,…, n_i. Let ***X*** be the vector of ***m*** covariate predictors and $\mathbf{X}_{ij}$ be the matrix of **X** values observed for the j^{th} subject within the i^{th} site.

Let $\pi_{ij} = P(Y_{ij} = 1)$,

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + X'_{ij}\beta, \quad i = 1,2,\ldots,K; j = 1,2,\ldots,n_i \qquad (1)$$

where, β_o is the intercept and β is the regression coefficient vector corresponding to the predictor vector **X**. This regression model was first fit using SAS's LOGISTIC procedure [18]. Although the data are defined within clusters, the use of the LOGISTIC procedure does not account for possible correlations within clusters, i.e., all observations within and between clusters are assumed to be independent.

2. Generalized Estimating Equations

The GEE model is similar to the standard logistic regression described in (1) above but accounts for the clustering by defining a covariance structure [1, 2]. GEE methodology applied in a Generalized Linear Model (GLM) setting requires a model with a link function (logit for binary response variable as in our case) and a covariance structure (working correlation matrix, e.g., independent, exchangeable/compound symmetry, unstructured, autoregressive, m-dependent) to describe the correlation of the measurements in a cluster. In these models, the parameters of the working covariance/correlation structure are estimated. SAS's GENMOD procedure (accounting for only fixed effects) can accommodate 6 types of covariance structures. The choice of correlation structure is typically not clear-cut and essentially requires consideration of a variety of factors (see Kleinbaum and Klein [4]). Nevertheless, Horton and Lipsitz [5] recommend the use of an unstructured matrix if the cluster size is small and balanced; an auto-regressive or m-dependent structure if the measurements are obtained over time and the correlation might be associated with time, and an exchangeable structure if there is no logical ordering for observations within a cluster. Also, GEE models allow for robust standard errors (described below under "multiplicative adjustments") to be computed to correct for possible misspecification of the working correlation structure.

We considered the following covariance structures in our analyses using PROC GENMOD in SAS:

(i) Independent

$$\begin{pmatrix} 1 & 0\ldots & 0 \\ 0 & 1\ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0\cdots & 1 \end{pmatrix}$$

(ii) Exchangeable (compound symmetry) – correlation is constant over time.

$$\begin{pmatrix} 1 & \rho\cdots & \rho \\ \rho & 1\ldots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \rho\cdots & 1 \end{pmatrix}$$

(iii) Unstructured – every correlation term is different.

$$\begin{pmatrix} 1 & \rho_{1,2}\cdots & \rho_{1,n} \\ \rho_{1,2} & 1\ldots & \rho_{2,n} \\ \vdots & \ddots & \vdots \\ \rho_{1,n} & \rho_{2,n}\cdots & 1 \end{pmatrix}$$

3. Random Effects Models

Random effects models are used in situations when it is of interest to allow one's model to incorporate heterogeneity among responses within different clusters, and/or when data have a hierarchical structure (e.g., patients nested within providers nested within clinics). These models, known as Generalized Linear Mixed Models (GLMM), are a generalization of the GEE approach to include random effects as predictors along with the fixed effects [3, 6]. As with fixed effect models, random effects models do not assume observations within a cluster to be independent and adjust for the amount of correlation due to the clustering effect. In these models, the parameters of the covariance structure of the random effects are also estimated, in addition to the parameters of the covariance structure of the errors associated with the fixed effects part of the model. In addition, robust standard errors can be computed for random effects models to correct for misclassification of the working correlation structure.

Again, we define a binary outcome variable Y by Y_{ij}= 1 if the j^{th} subject within the i^{th} site declines the vaccine and Y_{ij}= 0 otherwise, where i=1,2,…,K and j=1,2,…, n_i. Let ***X*** be the vector of ***m*** covariate predictors and $\mathbf{X}_{ij}$, considered to be the matrix of the *fixed effects,* be the matrix of **X** values observed for the j^{th} subject within the i^{th} site. Letting $\pi_{ij} = P(Y_{ij} = 1)$, and $\boldsymbol{u_j}$ be the *random effect* due to the clinic (the effect of patients clustering within clinics on the outcome variable),

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + X'_{ij}\beta + u_j,$$

$$i = 1,2,\ldots,K; j = 1,2,\ldots,n_i; u_j \sim N(0,\sigma_u^2)$$

We fit both linear and non-linear mixed models with random intercepts, using PROC GLIMMIX and PROC NLMIXED, respectively, in SAS.

4. Survey Logistic Regression Models

Models designed to analyze data from surveys account for the complex designs of the survey and adjust for the clustering of data within units (e.g., sites, clinics). We obtained robust standard errors and variances computed by using Taylor Linearization methods (a form of GEE) using PROC SURVEYLOGISTIC with the option *varmethod=Taylor* (in SAS) and PROC RLOGIST with the option *semethod=zeger* (in SUDAAN [19]) [1].

Having a small number of clusters limits the use of large sample tests and confidence intervals based on an approximate normal distribution. The estimation of the variance-covariance matrix (also known as the "sandwich" estimator) was derived by Binder (1983) [7], and Liang and Zeger (1986) [1], using a Taylor series expansion. Liang and Zeger [1] further derived the same for the GEE model for various response variables from the exponential family. These adjustments are available in SAS.

Multiplicative Adjustments

The classical sandwich estimator and the associated covariance matrix are biased if the number of independent sampling units is small. In "*Over dispersion Models in* SAS" [8], Morel and Neerchal briefly describe the adjustments available to correct for the small sample bias. Several investigators have proposed multiplicative adjustments to correct for this small sample bias by making adjustments to the middle term of the sandwich estimator: first-order Taylor series approximation on the residuals involved in the computation of the middle term [9, 10], using the F-distribution (or *t-distribution*) instead of the chi-square (or normal) with an adjustment to the degrees of freedom [10], an adjustment to provide confidence intervals to attain nominal coverage probabilities [11] or an approximate F-test (or *t-test*) to allow for simultaneous testing of more than one parameter. The general form of the "sandwich" estimator is given by:

$$\widehat{Var}(\hat{\beta}) = c\ x\ \hat{\Omega}\left(\sum_{i=1}^{m} A_i \widehat{D'_i} \hat{\Sigma}_i^{-1} F'_i e_i e'_i F_i \hat{\Sigma}_i^{-1} \hat{D}_i A_i\right)\hat{\Omega} \qquad (2)$$

where,

$$E(Y) = \mu; Var(Y) = \Sigma; D = \frac{\partial\mu}{\partial\beta}$$

$\Omega = (D'\Sigma^{-1}D)^{-}$ is the generalized inverse

$$e_i = (y_i - \widehat{\mu_l})$$

m is the number of independent sampling units

This estimator is biased if ***m*** is small. The following table from *"Over dispersion Models in SAS"* [8], (page 325) displays the multiplicative small sample bias corrections to the sandwich estimator available with PROC GLIMMIX in SAS.

Table 2: Multiplicative Small Sample Bias Corrections to the Sandwich Estimator

ADJUSTMENT	C	A_i	F_i	REFERENCE
CLASSICAL	*1*	*I*	*I*	*Liang and Zeger (1986); Zeger and Liang (1986)*
FIRORES	*1*	*I*	$(1-H_i')^{-1}$	*Mancl and DeRouen (2001)*
FIROEEQ (r)	*1*	$Diag\left\{\left(1-\min\{r,[Q]_{jj}\}\right)^{\frac{-1}{2}}\right\}$	*I*	*Fay and Graubard (2001)*

where,

$$H_i = D_i \Omega D_i' \Sigma_i^{-1}$$

$Q = D_i' \widehat{\Sigma}_i^{-1} D_i \widehat{\Omega}$; *0<=r<1* = is a constant defined by the user to provide an upper bound on the correction factor (default *r* = 0.75);

The classical adjustment computes the estimator defined in equation (2) above. We can obtain the same by setting *r*=0 in the FIROEEQ adjustment. These adjustments can be made by specifying them in the option *empirical=(classical, firores, fioroeeq)* in the procedure statement of PROC GLIMMIX.

Additive Adjustments

Morel developed an adjustment to the Taylor estimator of the covariance matrix obtained from a logistic regression model to analyze data from complex surveys [12]. Morel, Bokossa, and Neerchal (MBN) further developed an additive small sample bias correction that reduces as the number of clusters increases and disappears completely when the number of clusters is large [13]. This addition is applied to the entire estimator and not just to the middle term, as is the case with all the multiplicative adjustments. The MBN adjustment to the random effects models is shown below:

$$\widehat{Var}(\widehat{\beta}) = c\ x\ \widehat{\Omega}\left(\sum_{i=1}^{m} \widehat{D_i'}\widehat{\Sigma}_i^{-1} e_i e_i' \widehat{\Sigma}_i^{-1} \widehat{D}_i\right)\widehat{\Omega} + \delta_m \phi \widehat{\Omega} \qquad (3)$$

where,

$c = \dfrac{(m^*-1)}{(m^*-k)}\dfrac{m}{(m-1)}, m^*$ number of observations and *m* number of clusters

$\widehat{\delta_m} = \left\{\begin{matrix} \frac{k}{(m-k)} & if\ m > (d+1)k \\ \frac{1}{d} & otherwise \end{matrix}\right\}, d \geq 1(default\ d = 2),$ and

$$\widehat{\phi} = \max\left\{r, trace\left[\widehat{\Omega}\left(\sum_{i=1}^{m} \widehat{D_i'}\widehat{\Sigma}_i^{-1} e_i e_i' \widehat{\Sigma}_i^{-1} \widehat{D_i}\right)\right] / k\right\},$$

$0 \leq r \leq 1(default\ r = 1)$

The Morel adjustment can be made by using the option *vadjust=morel* in the model statement of PROC SURVEYLOGISTIC (in SAS). The adjustment developed by Morel, Bokossa, and Neerchal can be made by using the option *empirical=mbn* in the procedure statement of PROC GLIMMIX (in SAS).

The models, software procedures, and the software used for this analysis are displayed in Table **3**.

RESULTS

As mentioned above, a total of 4815 patients were recommended to receive the yellow fever vaccine. Of these, 247 (5%) declined the vaccine when it was offered to them. This restricted sample was distributed in 18 clusters of sizes 5 to 1579.

Odds ratios with 95% confidence intervals and p-values obtained from the different regression models to evaluate the association of the variables with the probability of refusing the yellow fever vaccine are displayed in Table **4**. The odds ratios and standard error estimates varied by the method chosen. Generally, when the odds ratio estimates were similar, models (#1-3) that did not adjust for clustering or for the small samples had lower standard errors (or smaller width of confidence intervals) and the standard errors increased by adjusting for the clustering (models 4-10), which further increased when an additional adjustment was added for the small number of clinics (models 11-15). Among the models that adjusted for the clustering and for small samples, the MOREL adjustment in SAS SURVEYLOGISTIC procedure gave the most conservative standard errors (widest confidence intervals) for all variables except age (model 12 had the most conservative result), and the

Table 3: Models, Procedures and Software Applied to the Data

Model Number	TYPE OF MODEL	PROCEDURE	SOFTWARE
	No adjustment for Clustering or Small Samples		
1	Unadjusted Logistic Regression	LOGISTIC	SAS
2	Linear - No Random Intercept	GLIMMIX	SAS
3	Non-Linear – No Random Intercept	NLMIXED	SAS
	Adjustment for Clustering and No Adjustment for Small Samples		
	Generalized Estimating Equations – Different Covariance Structures		
4	Independent	GENMOD	SAS
5	Exchangeable	GENMOD	SAS
6	Unstructured	GENMOD	SAS
	Random Intercept		
7	Linear	GLIMMIX	SAS
8	Non-Linear	NLMIXED	SAS
	Survey Methods		
9	Logistic Regression	SURVEYLOGISTIC	SAS
10	Logistic Regression – Independent Covariance Structure	RLOGIST	SUDAAN
	Adjustment for Small Samples		
	Random Intercept		
11	CLASSICAL	GLIMMIX	SAS
12	FIRORES	GLIMMIX	SAS
13	FIROEEQ	GLIMMIX	SAS
14	MBN	GLIMMIX	SAS
	SAS Survey Method		
15	MOREL	SURVEYLOGISTIC	SAS

random intercept model with the MBN was a close second. The variance of the random effect was highly significant in all these models (p<0.0001).

Except for the VFR variable, which remained significant under all approaches, the significance of the other variables varied (as expected) with the chosen approach. Overall, they remained significant or insignificant under most approaches, except for the following: (i) Age was statistically significant at the two-sided alpha level of 0.05 in the unadjusted models but was no longer significant when we adjusted for the clustering and the small sample size; (ii) travel duration was insignificant in all models except for GEE with *unstructured covariance structure*, and *random intercept models* with Classical, FIRORES and FIROEEQ adjustments; (iii) traveling to Africa was insignificant only in 2 models – GEE with *unstructured covariance structure* and the *SAS Survey Method* with the MOREL adjustment; and (iv) traveling to South America was significant only in the *random intercept model* fit with the GLIMMIX procedure in SAS.

DISCUSSION

Clustered observations often arise in research studies. Treating clustered observations as independent could bias the estimates and their standard errors. Several methods exist to adjust for clustering and they all work well, even when the covariance structures are misspecified, when the number of clusters is large (n≥30), and the clusters are well-balanced. Small number of clusters limits the use of large sample tests and/or confidence intervals based on the normal approximation and could also possibly inflate the Type I error rates [8]. The imbalance within each cluster does not allow full flexibility in the choice of correlation structures due to nonconvergence with certain correlation structures.

Table 4: Odds Ratios, 95% Confidence Intervals, and p-values from 15 different regression models without and with adjustments for clustering and small number of clusters to evaluate the association of covariates to the probability of refusing Yellow Fever Vaccine

Model Number	Type of Model	Age OR[b] (95% CI[c]) p-value	Duration OR (95% CI) p-value	VFR[a] OR (95% CI) p-value	Traveling to Africa OR (95% CI) p-value	Traveling to South America OR (95% CI) p-value
	No adjustment for Clustering or Small Samples					
1	Unadjusted Logistic Regression	1.008(1.001-1.015) 0.0333	1.160(0.856-1.571) 0.3376	5.069(3.712-6.922) <0.0001	0.472(0.259-0.860) 0.0141	0.601(0.323-1.120) 0.1089
2	GLIMMIX – No Random Intercept	1.008(1.001-1.015) 0.0334	1.160(0.856-1.571) 0.3377	5.069(3.711-6.923) <0.0001	0.472(0.259-0.860) 0.0142	0.601(0.323-1.120) 0.1090
3	NLMIXED – No Random Intercept	1.008(1.001-1.015) 0.0334	1.160(0.808-1.512) 0.3376	5.069(3.489-6.649) <0.0001	0.472(0.188-0.755) 0.0142	0.601(0.227-0.975) 0.1089
	Adjustment for Clustering and No Adjustment for Small Samples					
	Generalized Estimating Equations – Different Covariance Structures					
4	Independent	1.008(0.995-1.021) 0.2397	1.160(0.814-1.654) 0.4121	5.069(3.164-8.120) <0.0001	0.472(0.283-0.785) 0.0038	0.601(0.316-1.144) 0.1210
5	Exchangeable	1.003(0.987-1.019) 0.7293	0.863(0.723-1.030) 0.1018	2.293(1.557-3.377) <0.0001	0.423(0.285-0.629) <0.0001	0.649(0.397-1.062) 0.0852
6	Unstructured	0.997(0.991-1.002) 0.2700	0.860(0.747-0.991) 0.0373	1.880(1.313-2.693) 0.0006	0.880(0.730-1.060) 0.1781	1.145(0.885-1.482) 0.3029
	Random Intercept[d]					
7	GLIMMIX	1.003(0.995-1.012) 0.3970	0.793(0.557-1.127) 0.1955	2.863(1.929-4.249) <0.0001	0.255(0.126-0.513) 0.0001	0.483(0.237-0.983) 0.0446
8	NLMIXED	1.004(0.995-1.012) 0.4094	0.792(0.491-1.092) 0.2114	2.882(1.653-4.110) <0.0001	0.253(0.061-0.445) 0.0014	0.481(0.112-0.851) 0.0606
	Survey Methods					
9	SAS SURVEYLOGISTIC	1.008(0.994-1.021) 0.2534	1.160(0.805-1.671) 0.4256	5.069(3.121-8.234) <0.0001	0.472(0.279-0.797) 0.0050	0.601(0.310-1.166) 0.1321
10	SUDAAN LOGISTIC – Independent Covariance Structure	1.008(0.993-1.023) 0.2532	1.160(0.783-1.719) 0.4254	5.069(3.008-8.542) <0.0001	0.472(0.268-0.829) 0.0050	0.601(0.295-1.226) 0.1319
	Adjustment for Clustering and for Small Samples					
	Random Intercept[d]					
11	CLASSICAL	1.003(0.983-1.024) 0.7384	0.793(0.643-0.976) 0.0290	2.863(1.834-4.469) <0.0001	0.255(0.137-0.474) <0.0001	0.483(0.222-1.048) 0.0655
12	FIRORES	1.003(0.975-1.033) 0.8130	0.793(0.630-0.997) 0.0466	2.863(1.552-5.279) 0.0008	0.255(0.129-0.502) <0.0001	0.483(0.192-1.212) 0.1208
13	FIROEEQ	1.003(0.980-1.027) 0.7712	0.793(0.637-0.986) 0.0367	2.863(1.711-4.789) <0.0001	0.255(0.131-0.495) <0.0001	0.483(0.207-1.127) 0.0921
14	***MBN***	***1.003(0.981-1.026) 0.7617***	***0.793(0.526-1.194) 0.2658***	***2.863(1.567-5.230) 0.0006***	***0.255(0.099-0.654) 0.0045***	***0.483(0.166-1.399) 0.1797***
	SAS Survey Method					
15	MOREL	1.008(0.987-1.029) 0.4572	1.160 (0.546-2.465) 0.6995	5.069 (2.204-11.660) 0.0001	0.472(0.116-1.925) 0.2949	0.601(0.133-2.709) 0.5076

[a] Visiting family and relatives in the region of origin of self or family in a low/low-middle income country as defined by the World Health Organization's 2011 Human Development Index.
[b] Odds Ratios.
[c] Confidence Intervals.
[d] The variance of the random effects were significant (p<0.0001) in all the random effects models.

We applied 15 different models to our data to assess the degree to which clustering and the small number of clinics affect the results of our analysis in predicting the refusal of yellow fever vaccine. Results varied with the method chosen and the assumptions of the covariance structure. In our analysis most of the variables that were statistically significant or insignificant in the unadjusted analysis remained so even after adjusting for the clustering and the small number of clusters. However, we also noticed that failing to adjust for the clustering would have led to incorrect inferences, especially about age being an important predictor for the refusal of the vaccine. The p-value for age ranged from as small as 0.03 in the unadjusted analysis to a nonsignificant 0.24 (GEE with an *independent covariance structure*) to 0.73 (GEE with an *exchangeable covariance structure*) when adjusted for clustering, to greater than 0.76 when an additional adjustment was made for the small sample.

How does one choose a particular method as being the “correct” one when results vary by the chosen method? This difficult question can only be answered by conducting some sensitivity analyses. We chose one of the more conservative of approaches (GLIMMIX procedure in SAS with the MBN adjustment) for our analyses with this dataset, at least until we have a larger number of balanced clusters, since the variance of the random effect was significant and we wanted to adjust for the small number of clusters.

The clustering effect should never be ignored. Our results may convince readers that adjustments for small number of clusters are necessary. Although we only discuss procedures in SAS and SUDAAN, other statistical software (e.g., STATA, SPSS, R) have appropriate procedures to account for clustering, and some have additional adjustments for the small number of clusters. These procedures are all easy to apply, even for investigators not proficient in data analysis.

CONCLUSION

Inferences do vary by the method chosen, and the “wrong” choice could lead to incorrect inferences. Although our analysis does not indicate the correct method or the extent of the bias of incorrect methods, it does show the implications of the choice of different methods. We chose the random intercept model with the Morel, Bokossa and Neerchal (MBN) adjustment to be the most preferable method for the GTEN dataset since this was one of the more conservative models that accounted for clustering, small sample sizes and also the random effect due to site. Different methods might work better for other datasets. Investigators should not ignore clustering and should pay attention to the number and size of clusters, be aware of the assumptions of the different models, and consider the appropriate adjustments necessary for their studies.

ACKNOWLEDGEMENTS

Members of the Global TravEpiNet Consortium. Members of the Global TravEpiNet Consortium (in alphabetical order) are George M. Abraham, Saint Vincent Hospital (Worcester, MA); Salvador Alvarez, Mayo Clinic (Jacksonville, FL); Vernon Ansdell and Johnnie A. Yates, Travel Medicine Clinic, Kaiser Permanente (Honolulu, HI); Elisha H. Atkins, Chelsea HealthCare Center (Chelsea, MA); John Cahill, Travel and Immunization Center, St. Luke’s-Roosevelt (New York, NY); Holly K. Birich and Dagmar Vitek, Salt Lake Valley Health Department (Salt Lake, Utah); Bradley A. Connor, New York Center for Travel and Tropical Medicine, Cornell University (New York, NY); Roberta Dismukes and Phyllis Kozarsky, Emory TravelWell, Emory University (Atlanta, GA); Ronke Dosunmu, Journey Health (Maywood, NJ); Jeffrey A. Goad and Edith Mirzaian, International Travel Medicine Clinic, University of Southern California (Los Angeles, CA); DeVon Hale, International Travel Clinic, University of Utah (Salt Lake City, UT); Noreen A. Hynes, John Hopkins Travel and Tropical Medicine, Division of Infectious Diseases, John Hopkins School of Medicine (Baltimore, MD); Frederique Jacquerioz and Susan McLellan, Tulane University (New Orleans, LA); Mark Knouse, Keystone Travel Medicine, Lehigh Valley Health Network (Allentown, PA); Jennifer Lee, Northwestern Medical Group-Travel Medicine, Northwestern Memorial Hospital (Chicago, IL); Alawode Oladele and Hanna Demeke, DeKalb County Board of Health Travel Services-DeKalb North and Central-T.O. Vinson Centers (Decatur, GA); Roger Pasinski and Amy E. Wheeler, Revere HealthCare Center (Revere, MA); Jessica Rosen and Noreen N. Cabellon, Infectious Diseases and Travel Medicine, Georgetown University (Washington, DC); Brian S. Schwartz, Travel Medicine and Immunization Clinic, University of California (San Francisco, CA); William Stauffer and Patricia Walker, HealthPartners Travel Medicine Clinics (St. Paul, Minnesota); and Joseph Vinetz, Travel Clinic, Division of Infectious Diseases, Department of Medicine, University of California-San Diego School of Medicine (La Jolla, CA).

DISCLAIMER

The opinions expressed in this manuscript do not necessarily represent the official views of the Department of Veterans Affairs. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention (CDC).

FUNDING

This work was supported by U19CI000514 and U01CK000175 from the U.S. Centers for Disease Control & Prevention.

REFERENCES

[1] Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73: 13-22. http://dx.doi.org/10.1093/biomet/73.1.13

[2] Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986; 121-30. http://dx.doi.org/10.2307/2531248

[3] Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. Hoboken, New Jersey: John Wiley & Sons, Inc. 2004.

[4] Kleinbaum DG, Klein M. Logistic Regression: A Self-Learning Text (Chapters 14-16). Third Edition. New York Dordrecht Heidelberg London: Springer Publishers 2010. http://dx.doi.org/10.1007/978-1-4419-1742-3

[5] Horton NJ, Lipsitz SR. Review of software to fit generalized estimating equation regression models. Am Stat 1999; 53: 160-9.

[6] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc 1993; 88: 9-25.

[7] Binder DA. On the variances of asymptotically normal estimators from complex surveys. Int Stat Rev Int Stat 1983; 279-92.

[8] Morel JG, Neerchal N, Neerchal NK. Overdispersion models in SAS. Sas Inst 2012.

[9] Mancl LA, DeRouen TA. A Covariance Estimator for GEE with Improved Small-Sample Properties. Biometrics 2001; 57: 126-34. http://dx.doi.org/10.1111/j.0006-341X.2001.00126.x

[10] Fay MP, Graubard BI. Small-Sample Adjustments for Wald-Type Tests Using Sandwich Estimators. Biometrics 2001; 57: 1198-206. http://dx.doi.org/10.1111/j.0006-341X.2001.01198.x

[11] Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. J Am Stat Assoc 2001; 96: 1387-96. http://dx.doi.org/10.1198/016214501753382309

[12] Morel JG. Logistic regression under complex survey designs. Surv Methodol 1989; 15: 203-23.

[13] Morel JG, Bokossa MC, Neerchal NK. Small sample correction for the variance of GEE estimators. Biom J 2003; 45: 395-409. http://dx.doi.org/10.1002/bimj.200390021

[14] LaRocque RC, Rao SR, Lee J, Ansdell V, Yates JA, Schwartz BS, *et al*. Global TravEpiNet: a national consortium of clinics providing care to international travelers—analysis of demographic characteristics, travel destinations, and pretravel healthcare of high-risk US international travelers, 2009–2011. Clin Infect Dis 2012; 54: 455-62. http://dx.doi.org/10.1093/cid/cir839

[15] WHO | WHO regional offices [Internet]. WHO. [cited 2013 Jun 15]. Available from: http://www.who.int/about/regions/en/index.html. Accessed 15 January 2014.

[16] Indices & Data | Human Development Reports (HDR) | United Nations Development Programme (UNDP) [Internet]. [cited 2013 Jun 15]. Available from: http://hdr.undp.org/en/statistics/. Accessed 15 January 2014.

[17] Keystone JS. Immigrants returning home to visit friends and relatives (VFRs). Health Inf Int Travel 2012; 547-51.

[18] SAS, Guide SU. Version 9.2. Cary, NC, USA: SAS Institute Inc. 2008.

[19] SUDAAN Language Manual, Volumes 1 and 2, Release 11. Research Triangle Park, NC, USA: Research Triangle Institute 2012.

Received on 23-04-2014 Accepted on 27-06-2014 Published on 05-08-2014

http://dx.doi.org/10.6000/1929-6029.2014.03.03.1