Journal of Biomedical Analytics
Vol. 1, No. 1 (2018), pp. 37–50
doi:10.30577/jba.2018.v1n1.xxx

Tutorial Article

A Tutorial on GEE with Applications to Diabetes and Hypertension Data from a Complex Survey

Tahmina Akter[1], Elizabeth Bianca Sarker[1], and M. Shafiqur Rahman[1,*]

[1]Institute of Statistical Research and Training, University of Dhaka, Bangladesh
[*]Corrresponding author: shafiq@isrt.ac.bd

Received: January 31, 2018; revised: February 17, 2018; accepted: February 18, 2018.

Abstract: Correlated data frequently arise from cross-sectional studies with complex cluster design because individuals from the same cluster or region share some common characteristics. Analyzing correlated data using standard statistical methods, which are applicable for independent data, may produce misleading inference. This article reviews the GEE and its software implementations and provides some guidelines for using it in practice. To illustrate GEE, data from the 2011 Bangladesh Demographic and Health Survey, a two-stage complex cluster survey have been used to identify the risk factors for diabetes and hypertension. The results suggest that age, current working status, education, socioeconomic status, and body mass index are significantly associated with hypertension and diabetes. Further, we found significant positive correlation between the responses from the same cluster, justifying the use of GEE.

Keywords: Correlated data, cluster survey, risk factors, GEE, diabetes, hypertension

1 Introduction

Correlated data arise frequently in longitudinal studies conducted in many epidemiological and biomedical practices where subjects are followed over time with repeated monitoring of risk factors or health outcomes or both. Such correlated data also often arise from cross-sectional studies with complex study design such as multistage cluster design, where outcomes or responses from the same cluster tend to be correlated (Diggle, 2002). Analyzing

such data using standard statistical methods and models under independence assumption may provide misleading inference on the parameters of interest and hence it requires appropriate statistical methods and models for valid inference (Burton *et al.*, 1998). When such outcome is discrete, for example-binary or count, the complication of analyzing correlated data is partly due to the lack of rich class of models such as multivariate Gaussian for the joint distribution of the correlated outcomes (Zeger and Liang, 1986; Hanley *et al.*, 2003). Generalized estimating equations (GEE) was introduced by Zeger and Liang (1986) as an extension of generalized linear models (GLM) (McCullagh, 1984) for analyzing discrete correlated data, in which a working correlation matrix for the responses from the same subject or cluster is used. Later an exhaustive review of the development of the GEE approach was discussed by Ziegler *et al.* (1998). The GEE approach requires specification of only first two moments of the responses from the same subject or cluster rather than full specification of the joint distribution. The advantage of GEE is that it fits marginal mean models and hence requires only correct specification of marginal means, and provides asymptotically unbiased estimate of the regression coefficients even under mis-specification of correlation structure. In addition, the GEE estimates have population-average intuitive interpretation as compared with the other class of models for correlated data such as random effect models with conditional interpretation (Lee and Nelder, 2004, 1996).

As more advancements are made at biomedical studies, implementations of GEE methodology using built-in functions in common statistical software such as R, SAS, and Stata have become available. The GEE is increasingly being used by both applied statisticians and public health researchers for analyzing data from longitudinal studies(Fitzmaurice *et al.*, 2012; Feng *et al.*, 2001). Although data from nationally representative cross-sectional survey with complex cluster design, for example, Bangladesh Demographic and Health Survey (BDHS) (NIPORT *et al.*, 2011), are often correlated because individuals from the same cluster or region tend to be similar, use of GEE or similar methods for analyzing such survey data is very limited. The lack of application of GEE for survey data is perhaps due to unfamiliarity of the methods or other technical challenges that practitioners face such as availability of options in the commonly used software packages.

Thus, we present a brief review of GEE methodology and its implementation using standard statistical software packages. Further, we illustrate the GEE method with an application to data from Bangladesh demographic and health survey of 2011. In particular, we demonstrate the use of GEE to identify the associated risk factors for two common chronic diseases–diabetes and hypertension.

The paper is organized as follows. Section 2 reviews the GEE methodology and Section 3 discusses the implementation of GEE using standard statistical software. An illustration using health survey data is discussed in Section 4. Section 5 ends the paper with a general discussion.

2 Brief Overview of GEE

2.1 Notations

Let $Y_{ij}, i = 1, 2, \ldots, N, j = 1, 2, \ldots, n_i$ be the j^{th} response from the i^{th} subject or cluster and $X_{ij} = \{x_{ij1}, x_{ij2}, \ldots, x_{ijp}\}$ be the vector of corresponding p covariates. The responses y_{ij}'s $(j = 1, \ldots, n_i)$ are assumed to be correlated for each subject but independent between the subjects (or clusters). The marginal expectation $E(Y_i|X_i) = \boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^T$ is

modelled by $g(\boldsymbol{\mu}_i) = X_i^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional vector of unknown regression coefficients and $g(\cdot)$ is a known link function. For binary data, logit or probit link and for count data log link are commonly used link functions. For example, for the logit link, the mean model can be expressed as

$$E(Y_{ij}|x_{ij}) = \mu_{ij} = \frac{\exp(X_i^T\boldsymbol{\beta})}{1 + \exp(X_i^T\boldsymbol{\beta})}.$$

As proposed by Liang and Zeger (1986), GEE uses a common correlation matrix for repeated measurements from the same subject or cluster. Using standard notations, let us consider $V_i = A_i^{1/2} M_i(\boldsymbol{\alpha}) A_i^{1/2}$ as an working covariance matrix, where A_i is a diagonal matrix with known variance function $\nu(\mu_{ij})$ and $M_i(\boldsymbol{\alpha})$ is the corresponding working correlation matrix, which may depend on some parameters $\boldsymbol{\alpha}$ which is generally unknown. Under an assumption of a common structure of $M_i(\boldsymbol{\alpha})$, the regression coefficient $\boldsymbol{\beta}$ can be estimated by solving the following equations (called GEE):

$$U(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} D_i^T V_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0,$$

where $D_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$.

2.2 Working Correlation Matrix

The working correlation matrix $M(\boldsymbol{\alpha})$ for a balanced design with same number of repeated measurements from each subject or cluster represents the within subject dependence and takes the following form:

$$M(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & Corr(Y_{i1}, Y_{i2}) & \ldots & Corr(Y_{i1}, Y_{in}) \\ Corr(Y_{i2}, Y_{i1}) & 1 & \ldots & Corr(Y_{i2}, Y_{in}) \\ \vdots & \vdots & & \vdots \\ Corr(Y_{in}, Y_{i1}) & Corr(Y_{in}, Y_{i2}) & \ldots & 1 \end{pmatrix}$$

The dimension of the matrix depends on the number of observations (n) for each subject or cluster. Different forms of working correlation matrix can be assumed. The most commonly used structures are independent, exchangeable (or compound symmetry), auto-regressive (AR), M-dependent and unstructured. The independent structure assumes that observations from the same subject or cluster are uncorrelated , i.e, $M_{i,j} = 0$ if $i \neq j$. Under this assumption, the GEE is equivalent to the GLM score equation. The exchangeable structure assumes equal correlation between the observations, i.e $M_{i,j} = \rho$ if $i \neq j$ with $-1 \leq \rho \leq +1$. In the AR structure, the observations from the same subject or cluster have auto-regression relationship, i.e, $M_{i,j} = \rho^{|i-j|}$ if $i \neq j$. The correlation between any two adjacent observations is ρ and ρ^2 for the observations that are separated by three consecutive measurements and so on. Similarly, in the M-dependent structure, the consecutive observations have common correlation (say ρ_1), pair of observations separated by three measurements common

correlation (say ρ_2) and so on. In general, $M_{i,j} = \rho_{|i-j|}$ if $i \neq j$. Observations with separation greater than m are assumed to be independent, where m is a arbitrary value that represents the order of separation. While specifying this structure, the choice of a value of m should be less than the dimension of matrix n. Finally, the unstructured correlation assumes no specific structure of the correlation, i.e, $M_{i,j} = \rho_{i,j}$ if $i \neq j$. However, the main disadvantage of this structure is that the number of parameters to be estimated increases with increasing dimension of the matrix. Although the estimate of $\boldsymbol{\beta}$ is not affected by the choices of the appropriate working correlation matrix, its mis-specification affects the efficacy of the estimated regression coefficients (Fitzmaurice, 1995; Wang and Carey, 2003; Mancl and Leroux, 1996; Sutradhar and Das, 2000). There are several methods that have been proposed in the literature for selection of appropriate working correlation structure, for example, the methods proposed by Cui and Qian (2007) and Jaman *et al.* (2016). In general, if the number of observations in each subject or cluster is small in a balanced and complete design, then unstructured correlation matrix is recommended. For a dataset from longitudinal study with mistimed observations, it is reasonable to choose a correlation model which is the function of time, for example, auto-regressive or M-dependent. For a data set from a complex clustered design where there is no chronological ordering of the observations from the same cluster, an exchangeable structure may be appropriate choice (Horton and Lipsitz, 1999).

2.3 Estimation

Under a given correlation structure, the $\boldsymbol{\beta}$ can be then estimated by the following standard iterative process proposed by Liang and Zeger (1986):

(i) Choose an initial estimate $\boldsymbol{\beta}^{(0)}$ of $\boldsymbol{\beta}$ as the estimate obtained by fitting GLM considering independent working correlation.

(ii) Given $\boldsymbol{\beta}^*$ ($\boldsymbol{\beta}^* = \boldsymbol{\beta}^{(0)}$ at the first iteration), calculate moment estimate $\boldsymbol{\alpha}^*$ of $\boldsymbol{\alpha}$ of the working correlation matrix $M(\boldsymbol{\alpha})$. For example, for exchangeable working correlation

$$\alpha^* = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i(n_i-1)}\sum_{j\neq k}^{n_i} e_{ij}^* e_{ik}^* \text{ where } e_{ij}^* = \frac{y_{ij}-\mu_{ij}^*}{\sqrt{\nu(\mu_{ij}^*)}}.$$

For AR(1),

$$\alpha^* = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i-1}\sum_{j\leq n_i-1}^{n_i} e_{ij}^* e_{i,j+1}^*.$$

(iii) Given the working correlation matrix $M(\boldsymbol{\alpha}^*)$ obtained in step (ii), the current estimate $\boldsymbol{\beta}^{*t}$ is updated using the Newton-Raphson method as

$$\boldsymbol{\beta}^{*(t+1)} = \boldsymbol{\beta}^{*t} + \left[I(\boldsymbol{\beta},\boldsymbol{\alpha})\right]^{-1}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{*t}} U^*(\boldsymbol{\beta},\boldsymbol{\alpha})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{*t}}.$$

(iv) Iterate steps (ii) to (iii) until a desired convergence achieved. At convergence, the estimate of $\boldsymbol{\beta}$ is denoted by $\boldsymbol{\beta}^*$ and the final estimates of $\boldsymbol{\alpha}$ are given by $\boldsymbol{\alpha}^*$ used in the final step of iteration. The estimates $\boldsymbol{\beta}^*$ of $\boldsymbol{\beta}$ will be referred as GEE estimates.

The variance of the GEE estimate (var($\hat{\beta}$)) can be estimated in two different approaches: sandwich-based robust estimate and model-based naive estimate. The sandwich-based robust estimator can be used to estimate (var($\hat{\beta}$)) empirically through the iterative process by substituting the estimate of $\hat{\boldsymbol{\beta}}$ into the following equation at each iteration and updated it for final estimate:

$$V_s = \left(\sum_{i=1}^{N} D_i' V_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{N} D_i' V_i^{-1} \text{Cov}(Y_i) V_i^{-1} D_i\right) \left(\sum_{i=1}^{N} D_i' V_i^{-1} D_i\right)^{-1},$$

where $\text{Cov}(Y_i) = E(Y_i - \boldsymbol{\mu}_i)(Y_i - \boldsymbol{\mu}_i)^T$. If the mean model is correct, this estimator is asymptotically unbiased and consistent even under mis-specification of working correlation structure. In contrast, the model-based estimator is consistent when both the mean model and covariance model is correctly specified. However, the correct form of working correlation matrix generally is usually unknown to the analyst and therefore if the number of subjects or cluster size is large sandwich-based estimator is a preferable choice (Horton and Lipsitz, 1999; Liang and Zeger, 1986).

3 Software Packages for Implementation of GEE

Most statistical software, such as R, SAS, and Stata, offer packages for fitting marginal models using GEE. We described here the available packages that are commonly used to implement the GEE methodology. There are two packages available in R: `gee` and `geepack`. The basic difference between these two packages is that `geepack` offers an option to identify the order of the observations within the groups (subjects or clusters) for fitting temporal correlation models to the datasets with missing values, which is not possible in `gee`. In addition to that, it provides an ANOVA method that allows to carry out multivariate Wald test. To implement GEE in Stata, package `xtgee` is widely used as part of `xt` cross-sectional time-series analysis tools. The SAS software uses `PROC GENMOD` to implement GEE. Table 1 describes how the correlation model can be implemented using each of the three software. All software have facilities for implementing mean models with different link functions. For example, for binary data, the commonly used link functions are `logit`, `probit` and `cloglog`, and for count data, only `log` link is widely used.

Table 1: Options for specifying correlation structure in different software.

Correlation Structure	R (gee, geepack)	Stata	SAS
Independence	corstr=independence	corr(ind)	corr=inde
Exchangeable	corstr=exchangeable	corr (exc)	corr=exch
Unstructured	corstr=unstructured	corr(uns)	corr=un
Auto-regressive	corstr=AR-M, ar1	corr(ar 1)	corr=ar
M-dependent	corstr=stat_M_dep	corr(sta m)	corr=mdep(m)

4 Application of GEE to Health Survey Data

4.1 Data and Variables

An application of GEE is provided to diabetes and hypertension data extracted from the 2011 BDHS aiming at investigating the factors associated with these two common non-communicable diseases. Both diabetes and hypertension significantly increases the risk of cardiovascular diseases (CVDs) such as stroke, myocardial infarction, premature death. According to the WHO estimate in 2014 (WHO, 2014), about 17 million deaths occur worldwide due to CVDs of which almost 80% occurs in low and middle income countries like Bangladesh. The prevalence of diabetes and hypertension among the Bangladeshi adult population increased markedly in the last two decades (Joshi *et al.*, 2007; Saquib *et al.*, 2012). The prevalence of diabetes increased from 3.2% in 1992 to 13.5% in 2015 and hypertension from 11% in 1992 to 20.4% in 2015 (Zaman *et al.*, 2013). Several studies (Joshi *et al.*, 2007; Saquib *et al.*, 2012) have been conducted to identify the risk factors associated with hypertension and diabetes, most of which were small-scale and confined to a specific group or community or based on data from a clinic or hospital, which cannot provide sufficient and accurate results for Bangladesh at large. Therefore, findings of these studies make it difficult for policy makers to design appropriate policies for prevention or reduction of the burden of such diseases. Although few studies (Chowdhury *et al.*, 2016) focused on the nationally representative data from large scale survey such as BDHS, none of these applied appropriate statistical methods and models for analyzing such data allowing for dependence between the responses from the same cluster and may fail to provide valid conclusion regarding the parameter of interest (Hanley *et al.*, 2003; Moen *et al.*, 2016). This paper analyzes diabetes and hypertension data of the 2011 BDHS using GEE to identify the associated risk factors and make some recommendations.

The BDHS is a national representative cross-sectional survey, which has been conducted in every three years since 1993 as a part of worldwide demographic and health survey conducted in developing countries under the authority of the National Institute for Population Research and Training (NIPORT) and Measure DHS (NIPORT *et al.*, 2011). The BDHS employs two-stage stratified cluster sampling method, where in first stage, enumeration areas (EA) were randomly selected (75% from rural area and 25% from urban area) from each of the seven administrative divisions, and in the second stage, households were selected using systematic sampling method from the each selected EA. For hypertension and diabetes data, one in three of the eligible households (subsample) were taken in consideration for gathering associated biomarker information of all members of age 35 and older. Details on sampling methods, sample size, survey procedure and questionnaires can be found elsewhere (NIPORT *et al.*, 2011).

For each eligible adult in the selected household, three measurements of both systolic and diastolic blood pressure were taken almost 10 minute intervals, where average of last two measurements was considered as final blood pressure values. An individual was considered hypertensive if systolic blood pressure (SBP) $\geq$ 140 mmHg (millimeters of mercury) and/or, diastolic blood pressure (DBP) $\geq$ 90 mmHg and/or if he/she is taking antihypertensive medication during the survey, otherwise not. For identifying the individuals with diabetes, the individuals who are told by doctor or nurse to have diabetes and/or taking medication for it are considered as diabetic patients, otherwise not. Information on both the diabetes and hypertension are considered as response variables.

Based on the literature (Chowdhury *et al.*, 2016; Joshi *et al.*, 2007) and exploratory analysis, the associated risk factors of both diabetes and hypertension considered here are age, sex, education (no education, primary, secondary, higher), wealth index (poorest, poorer, middle, richer, richest), place of residence (urban, rural), administrative division (Barisal, Chittagong, Dhaka, Khulna, Rajshahi, Rangpur, Sylhet), body mass index (BMI: underweight if BMI $<$18.5, normal if BMI 18.5-24.9, and overweight/obese if BMI $\geq$25), and current working status (yes, no). The calculated wealth index in BDHS was re-categorized as 'poor' for the poorest and poorer, 'middle', and 'rich' for the richest and richer, respectively. For exploratory analysis, age was categorized as 35-45 yrs, 46-55 yrs and 56 and above. However, age was treated as continuous in the multivariable regression model.

4.2 Statistical Analysis

Firstly, exploratory analysis based on contingency table was performed to estimate the prevalence of diabetes and hypertension with 95% binomial confidence interval for true prevalence, by the background characteristics (risk factors) of the respondents. The estimate was obtained by incorporating survey weights. Secondly, as the responses (binary outcome) on diabetes and hypertension from the same EA (cluster and then household) were expected to be correlated as they share the same cluster-level information such as food, environment, treatment, awareness about disease prevention etc., we preferred GEE based marginal logistic regression models to analyse the data. Separate models with logit link were fitted for diabetes and hypertension to identify the associated risk factors. Thirdly, we fitted the models under three different correlation structures namely, independence, exchangeable, and autoregressive of order 1 (AR 1) to examine if the results of the mean models (e.g. estimates of the regression coefficients of the model and the corresponding standard error and p-value) vary across the correlation models. The robust sandwich based estimate of the standard error of the regression coefficient was reported. The following commands in Stata and R were used to implement GEE for the hypertension data. Similar commands can be used for diabetes data by replacing the response variable.

To estimate the mean model with robust sandwich standard error, Stata command for GEE:

```
xi: xtgee hypertension age i.sex i.education i.wealth i.residence\\
 i.division i.bmi i.diabetes i.cworking, link(logit) \\
 corr(exchangeable) family(binomial) robust
 xtcorr  // to estimate correlation matrix
```

To estimate GEE under different correlation structure, for example AR(1), replace `corr(exchangeable)` by `corr(ar 1)`. To estimate the mean model with robust sandwich standard error R codes for GEE:

```
modelgee<-gee(hypertension~age+factor(sex)+factor(education)\\
+factor(wealth)+factor(residence)+factor(division)\\
+factor(bmi)+factor(diabetes)+factor(cworking), \\
id=clusterid, family=binomial, corstr="exchangeable")
summary(modelgee)
```

4.3 Results

The overall prevalence of hypertension and diabetes were reported as 24.9% and 5.7%, respectively (Table 2). The prevalence of both diseases increased with the increasing level of age, education, socioeconomic status, and BMI. Female and urban people are at higher risk of developing both hypertension and diabetes. There was also regional variation in the prevalence of the diseases. From the GEE model for hypertension, the results suggest that under three correlation models, the estimates of the regression coefficients are similar, except for the estimates of the associated standard error. The results revealed that age, sex, education, socio-economic status, diabetes and BMI have significantly positive association with the odds of having hypertension (Table 3). For example, the odds of having hypertension among the individuals with normal BMI was 85% percent higher (OR=$e^{0.6161} = 1.85$) compared to those individuals with BMI less than 18.5. Similarly, the odds among the overweight and obese was reported to 3.69 (OR=$e^{1.3075} = 3.69$) times of those individuals with BMI less than 18.5. Further, place of residence and current working status were found to be negatively associated with the odds of having hypertension. For example, people living in rural area and the people working outside were found to have significantly lower odds of having hypertension than their counterparts. Under both AR(1) and exchangeable correlation, the estimated within cluster correlation was positive and statistically significant, suggesting the necessity of using GEE model rather than standard logistic model. Similar results can be observed for the diabetes (Table 4). For both data, the number of clusters was reported as 457 with an average of 11.5 individuals per cluster.

5 Discussion and Conclusion

The GEE based marginal models have been widely used by the applied statisticians for analyzing longitudinal data, however, the public health or biomedical researchers who frequently use complex survey data rarely apply GEE or similar methods, despite that data from such survey are often correlated. The lack of popularity of these methods among the practitioners is probably due to the unavailability of user friendly manuals describing the methods and their implementations using common statistical software packages. This paper reviewed GEE with an aim to provide some guidelines for the public health and biomedical researchers so that they can understand the method and its application to analyzing correlated data. The paper also showed an application of the GEE for analyzing diabetes and hypertension data from a cross-sectional survey (BDHS 2011) with complex design. Findings suggest that the odds of having hypertension and diabetes increase with the increasing level of age, education, socioeconomic status, and BMI. For BMI, although the individuals with low BMI (say less than 18.5) tend to have smaller odds of having such cardiovascular disease compared to those with high BMI, they may have other health hazards due to nutrition deficiency (Lavie *et al.*, 2011). Further, males, people from rural area, people working outside have lower odds of having both hypertension and diabetes. The results (particularly the SE of the estimates of the regression coefficients and the associated p-values) are slightly different from those found in other studies which analyzed similar data using the standard logistic regression approach (Chowdhury *et al.*, 2016; Saquib *et al.*, 2012). This is because there was significantly positive correlation between the responses of hypertension and diabetes from the same cluster, and the GEE provided estimates allowing for such correlation, which was not possible by using standard logistic regression.

Further findings of the study suggests that the GEE estimates of the regression coefficient, particularly the SE and p-value, under both AR(1) and exchangeable correlation structure is slightly different from the estimates under independence correlation structure. Because, under independence correlation structure, the GEE provides equivalent estimates to those of the standard logistic regression model that does not account for the effect of correlation between the responses from the same cluster. The difference in the results between these two approaches increases with increasing degree of intra-cluster correlation. However, the advantage of using GEE for analyzing cluster data is that, even if intra-cluster correlation does not exist in practice, the GEE provides the same results as the standard logistic models. We therefore recommend to use GEE for analyzing complex survey data. Alternatively, mixed-effect models, a random intercept mixed model which is equivalent to GEE with exchangeable correlation (Lee and Nelder, 2004, 1996), can be applied to analyze such data if the number of clusters and their size is large (Moen *et al.*, 2016). However, the GEE estimate has intuitive population average interpretation compared to those of the mixed effect models that have cluster-specific conditional interpretations.

There is one limitation of the study that needs to be mentioned. The application of the study dealt with a secondary data from retrospective cross-sectional survey, where data on diabetes were collected by asking question that whether he/she was told by doctor or nurse to have diabetes or taking medication, rather than collecting data from their blood sample. Therefore, there is very high chance of under reporting of diabetes data.

Declarations

Acknowledgments

We acknowledge the authority of Bangladesh Demographic and Health Survey 2011 (National Institute of Population Research and Training and Measures DHS) for making the dataset available in public domain and providing approval for using it. In addition, the authors are grateful to editor and the reviewer for their valuable suggestions and comments, which have improved the presentation of the paper.

Funding

The authors did not receive any fund for this research.

Conflict of Interest

None declared.

Ethical Approval

As the study used data from a secondary source(NIPORT and DHS Program), the ethics approval and consents have been approved by the authority who made the data available for public use.

References

Burton P, Gurrin L, Sly P (1998). "Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multilevel mixed modeling." *Statistics in Medicine*, **17**, 1261–1291. doi:10.1002/(SICI)1097-0258(19980615)17:11<1261::AID-SIM846>3.0.CO;2-Z.

Chowdhury MAB, Uddin MJ, Haque MR, Ibrahimou B (2016). "Hypertension among adults in Bangladesh: evidence from a national cross-sectional survey." *BMC Cardiovascular Disorders*, **16**(1), 22. doi:10.1186/s12872-016-0197-3.

Cui J, Qian G (2007). "Selection of Working Correlation Structure and Best Model in GEE Analyses of Longitudinal Data." *Communications in Statistics - Simulation and Computation*, **36**(5), 987–996. doi:10.1080/03610910701539617.

Diggle P (2002). *Analysis of longitudinal data*. Oxford University Press.

Feng Z, Diehr P, Peterson A, McLerran D (2001). "Selected statistical issues in group randomized trials." *Annual Review of Public Health*, **22**(1), 167–187. doi:10.1146/annurev.publhealth.22.1.167.

Fitzmaurice GM (1995). "A caveat concerning independence estimating equations with multivariate binary data." *Biometrics*, pp. 309–317. doi:10.2307/2533336.

Fitzmaurice GM, Laird NM, Ware JH (2012). *Applied longitudinal analysis*, volume 998. John Wiley & Sons.

Hanley JA, Abdissa N, Michael DdE, Janet EF (2003). "Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation." *American Journal of Epidemiology*, **157**(4), 364–375. doi:10.1093/aje/kwf215.

Horton NJ, Lipsitz SR (1999). "Review of Software to Fit Generalized Estimating Equation Regression Models." *The American Statistician*, **53**(2), 160–169. doi:10.2307/2685737.

Jaman A, Latif MAHM, Bari W, Wahed AS (2016). "A determinant-based criterion for working correlation structure selection in generalized estimating equations." *Statistics in Medicine*, **35**(11), 1819–1833. doi:10.1002/sim.6821.

Joshi P, Islam S, Pais P, Reddy S, Dorairaj P, Kazmi K ea (2007). "Risk factors for early myocardial infarction in South Asians compared with individuals in other countries." *JAMA*, **297**(3), 286–294. doi:10.1001/jama.297.3.286.

Lavie CJ, De Schutter A, Patel D, Artham SM, Milani RV (2011). "Body Composition and Coronary Heart Disease Mortality–An Obesity or a Lean Paradox?" *Mayo Clinic Proceedings*, **86**(9), 857–864. doi:10.4065/mcp.2011.0092.

Lee Y, Nelder JA (1996). "Hierarchical generalized linear models." *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(4), 619–678.

Lee Y, Nelder JA (2004). "Conditional and marginal models: another view." *Statistical Science*, **19**(2), 219–238. doi:10.1214/088342304000000305.

Liang KY, Zeger SL (1986). "Longitudinal data analysis using generalized linear models." *Biometrika*, **73**(1), 13–22. doi:10.1093/biomet/73.1.13.

Mancl LA, Leroux BG (1996). "Efficiency of regression estimates for clustered data." *Biometrics*, pp. 500–511. doi:10.2307/2532890.

McCullagh P (1984). "Generalized linear models." *European Journal of Operational Research*, **16**(3), 285–292. doi:10.1016/0377-2217(84)90282-0.

Moen EL, Fricano-Kugler C, Luikart B, OŠMalley A (2016). "Analyzing Clustered Data: Why and How to Account for Multiple Observations Nested within a Study Participant?" *PLoS ONE*, **11**(1), e0146721. doi:10.1371/journal.pone.0146721.

NIPORT, Mitra Associates, ICF International (2011). "Bangladesh Demographic and Health Survey." *Technical report*, National Institute of Population Research and Training (NIPORT), Dhaka, Bangladesh.

Saquib N, Saquib J, Ahmed T, Khanam MA, Cullen MR (2012). "Cardiovascular diseases and Type 2 Diabetes in Bangladesh: A systematic review and meta-analysis of studies between 1995 and 2010." *BMC Public Health*, **12**(1). doi:10.1186/1471-2458-12-434.

Sutradhar BC, Das K (2000). "On the accuracy of efficiency of estimating equation approach." *Biometrics*, **56**(2), 622–625. doi:10.1111/j.0006-341X.2000.00622.x.

Wang YG, Carey V (2003). "Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance." *Biometrika*, **90**(1), 29–41. doi:10.1093/biomet/90.1.29.

WHO (2014). "Global status report on noncommunicable diseases 2014." *Technical report*, World Health Organization, Geneva, Switzerland.

Zaman M, Taleb A, Rahmani S, Acharyyai A, Islami F, Ahmedi M, Zamanii M (2013). "Prevalence of hypertension among the Bangladeshi adult population: a meta-analysis." *Journal of Epidemiology and Community Health*, **65**(1). doi:10.1136/jech.2011.142976n.89.

Zeger SL, Liang KY (1986). "Longitudinal data analysis for discrete and continuous outcomes." *Biometrics*, pp. 121–130. doi:10.2307/2531248.

Ziegler A, Kastner C, Blettner M (1998). "The generalised estimating equations: an annotated bibliography." *Biometrical Journal*, **40**(2), 115–139. doi:10.1002/(SICI)1521-4036(199806)40:2<115::AID-BIMJ115>3.0.CO;2-6.

Table 2: Prevalence of diabetes and hypertension by the background characteristics.

	Hypertension		Diabetes	
	n at risk	Prevalence (95% CI)	n at risk	Prevalence (95% CI)
Age				
35-45	3196	16.4 (15.1-17.8)	3060	4.1 (3.3-4.9)
46-55	2146	25.2 (23.1-27.3)	2026	6.3 (5.2-7.7)
56+	2529	35.3 (33.2-37.5)	2259	7.5 (6.3-8.9)
Sex				
Male	3896	18.7 (17.3-20.1)	3751	5.1 (4.3-5.9)
Female	3975	31.1 (29.5-32.7)	3594	6.4 (5.5-7.4)
Education				
No education	4798	25.3 (23.9-26.7)	4331	4.1 (3.5-4.9)
Primary	1502	21.7 (19.4-24.1)	1449	5.9 (4.6-7.4)
Secondary	1033	22.9 (20.2-25.9)	1019	8.4 (6.7-10.6)
Higher	538	35.7 (31.1-40.6)	546	15.1 (11.8-19.0)
Wealth index				
Poor	2827	19.6 (18.0-21.2)	2482	2.0 (1.4-2.8)
Middle	1526	21.5 (19.3-23.9)	1415	3.4 (2.5-4.6)
Rich	3518	31.6 (29.9-33.4)	3448	10.0 (8.8-11.2)
Place of residence				
Urban	2581	31.6 (29.4-33.9)	2518	9.7 (8.4-11.3)
Rural	5290	22.9 (21.7-24.2)	4827	4.4 (3.8-5.1)
Division				
Barisal	936	24.3 (21.5-27.3)	867	4.4 (3.2-6.1)
Chittagong	1181	21.1 (18.8-23.5)	1102	7.1 (5.7-8.8)
Dhaka	1343	26.2 (23.9-28.6)	1295	6.9 (5.7-8.4)
Khulna	1238	29.4 (26.8-32.1)	1195	4.1 (3.1-5.4)
Rajshahi	1114	23.5 (20.9-26.2)	1039	5.7 (4.4-7.4)
Rangpur	1095	26.8 (24.1-29.6)	955	2.7 (1.9-3.9)
Sylhet	964	20.2 (17.6-23.0)	892	5.7 (4.4-7.5)
BMI				
Underweight	2145	16.6 (14.9-18.4)	1896	2.7 (2.0-3.8)
Normal	4304	24.3 (22.9-25.7)	4068	5.2 (4.5-6.0)
Overweight/Obese	1175	42.6 (39.3-45.9)	1154	13.6 (11.4-16.1)
Diabetes				
No	6863	23.7 (22.6-24.8)		
Yes	439	48.3 (42.9-53.8)		
Hypertension				
No			5409	4.0 (3.4-4.6)
Yes			1893	11.1 (9.5-12.9)
Currently working				
No	4051	32.5 (30.9-34.1)	3675	7.0 (6.1-8.0)
Yes	3818	16.8 (15.5-18.2)	3668	4.4 (3.7-5.3)
Overall	7871	24.9 (23.8-26.0)	7345	5.7 (5.2-6.4)

Table 3: GEE estimates of the logistic regression models for hypertension.

	Independence		Exchangeable		AR(1)	
	Coef.	S.E	Coef.	S.E	Coef.	S.E
Age	0.0443 **	0.0026	0.0444**	0.0026	0.046**	0.0038
Sex						
Male						
Female	0.6797**	0.1043	0.6729**	0.1038	0.5831**	0.1545
Education						
No education						
Primary	0.0336	0.0863	0.023	0.0859	0.0521	0.1273
Secondary	0.2837**	0.1077	0.274*	0.1075	0.3475*	0.1477
Higher	0.7259**	0.1267	0.7057**	0.1271	0.6748**	0.1682
Wealth index						
Poor						
Middle	-0.0396	0.0911	-0.0379	0.0903	-0.0617	0.1422
Rich	0.1957*	0.0791	0.1739*	0.0787	0.1558	0.1248
Place of residence						
Urban						
Rural	-0.1713*	0.0744	-0.1853*	0.0743	-0.2457*	0.0992
Division						
Barisal						
Chittagong	-0.2032	0.1282	-0.1989	0.1298	0.1974	0.1735
Dhaka	0.0846	0.1193	0.0756	0.1212	0.3616*	0.1693
Khulna	0.2397	0.1288	0.2343	0.1293	0.3376*	0.1703
Rajshahi	0.1163	0.1247	0.1163	0.1269	0.1672	0.1957
Rangpur	0.4167**	0.1258	0.4121**	0.127	0.8068**	0.1908
Sylhet	-0.0863	0.1429	-0.0869	0.1458	0.021	0.215
BMI						
Underweight						
Normal	0.6162**	0.0793	0.5934**	0.0785	0.6625**	0.1196
Overweight/Obese	1.3075**	0.1028	1.2917**	0.1025	1.3644**	0.1521
Diabetes						
No						
Yes	0.5573**	0.1216	0.5526**	0.1199	0.4553**	0.1734
Currently working						
No						
Yes	-0.3221**	0.0938	-0.3217**	0.0936	-0.42**	0.1362

$*p < 0.05, **p < 0.01$

Table 4: GEE estimates of the logistic regression models for diabetes.

	Independence		Exchangeable		AR(1)	
	Coef.	S.E	Coef.	S.E	Coef.	S.E
Age	0.0156**	0.0045	0.0156**	0.0045	0.0143*	0.0057
Sex						
Male						
Female	0.1773	0.162	0.1721	0.162	0.1429	0.2115
Education						
No education/ Preschool						
Primary	0.6102**	0.1528	0.6052**	0.153	0.4184*	0.1844
Secondary	0.8578**	0.1608	0.8523**	0.161	0.8803**	0.1894
Higher	1.3532**	0.1798	1.3458**	0.1801	1.2393**	0.2223
Wealth index						
Poor						
Middle	0.3479	0.2289	0.3489	0.2284	0.2126	0.2664
Rich	1.036**	0.1711	1.0329**	0.171	0.9379**	0.1942
Place of residence						
Urban						
Rural	-0.1216	0.115	-0.127	0.1151	-0.1871	0.1408
Division						
Barisal						
Chittagong	0.3211	0.1894	0.3229	0.1897	0.1786	0.2495
Dhaka	0.3593	0.1917	0.3578	0.1918	0.048	0.2482
Khulna	-0.0926	0.2112	-0.097	0.2109	-0.0769	0.2599
Rajshahi	0.3125	0.202	0.3142	0.2026	0.2413	0.2561
Rangpur	-0.25	0.2496	-0.2476	0.25	-0.4546	0.3056
Sylhet	0.3265	0.2188	0.3269	0.22	0.3074	0.2874
BMI						
Underweight						
Normal	0.5378**	0.1685	0.5339**	0.168	0.6436**	0.2169
Overweight/Obese	1.0403**	0.1993	1.0346**	0.1992	1.1086**	0.2526
Hypertension						
No						
Yes	0.5465**	0.1219	0.5464**	0.1218	0.5264**	0.1555
Currently working						
No						
Yes	-0.6203**	0.175	-0.6217**	0.1751	-0.7034**	0.2274

$*p < 0.05, **p < 0.01$