Paper 869-2017

Using a Population Average Model to Investigate the Success of a Customer Retention Strategy

Colleen E McGahan, British Columbia Cancer Agency

ABSTRACT

In many healthcare settings, patients are like customers – they have a choice. An example is whether or not to receive a procedure. In population-based screening where the goal is to reduce deaths, the success of a screening program hinges on the patient's choice to accept the procedure. Like in many other industries, this not only relies on the program's ability to attract new eligible patients, it also relies on its ability to retain existing customers.

The success of a new customer retention strategy within the breast screening environment is examined. A population averaged (PA) model (aka marginal models) using Generalized Estimating Equations (GEE) in the GENMOD procedure is used to account for the lack of independence of the observations. This business case provides a great introductory session for people who are not familiar with mixed effects models or hierarchical modelling with a binary outcome. It provides guidance when to use a cluster-specific mixed effects model and when to apply a PA model. It then leads you through the decision process to understand why the PA model is the most appropriate method in this scenario, and implements a GEE within *SAS*®.

INTRODUCTION

This paper is aimed at analysts familiar with logistic regression models but who are seeking to learn more about mixed effects or hierarchical modelling with a binary outcome. It focuses on two methods; one is the 'population average' (PA) using Generalized Estimating Equations (GEE) by Liang and Zeger, and method two is the subject or cluster-specific (CS) method using random effects model with a binary outcome. Minimal methodology is specified as the focus is on the applied application of the methods.

Using a GEE, the success of a new customer retention strategy within the breast screening environment is examined (the business case). Correlation is inherent within the data, which both the GEE and CS methods can account for. Since many people can be unsure when it is appropriate to use each method, the motivation of this paper is to use this business case to provide guidance and understanding as to when each method is appropriate. The paper first presents background information and the design of the study implemented. It then gives a brief description of each method and leads you through the decision process so you can understand why the PA model is the most appropriate model in this scenario. It concludes with implementation of the GEE model using the GENMOD procedure in SAS/STAT® software Version 14.1.

Although this example is health related, the concept can be applied to all industries. It is critical to understand the objective of your study and the structure of your data to make the informed decision on which is the most appropriate method to apply. Once you understand and recognize the inherent nuances of your data, there is very little SAS® code required.

BUSINESS CASE

BACKGROUND

Regular breast cancer screening is an important part of women's health. British Columbia (BC) is among the best in the world when it comes to population breast cancer survival outcomes [2]. This success is largely due to the publicly funded health care system accessible by all residents, a centralized cancer treatment system and a population-based breast cancer screening program. According to randomized trials investigating the efficacy of breast cancer screening, the relative reduction in mortality due to breast

cancer ranges from 15 to 25% [3, 4, 5, 6], which emphasizes the value of patient participation in screening programs.

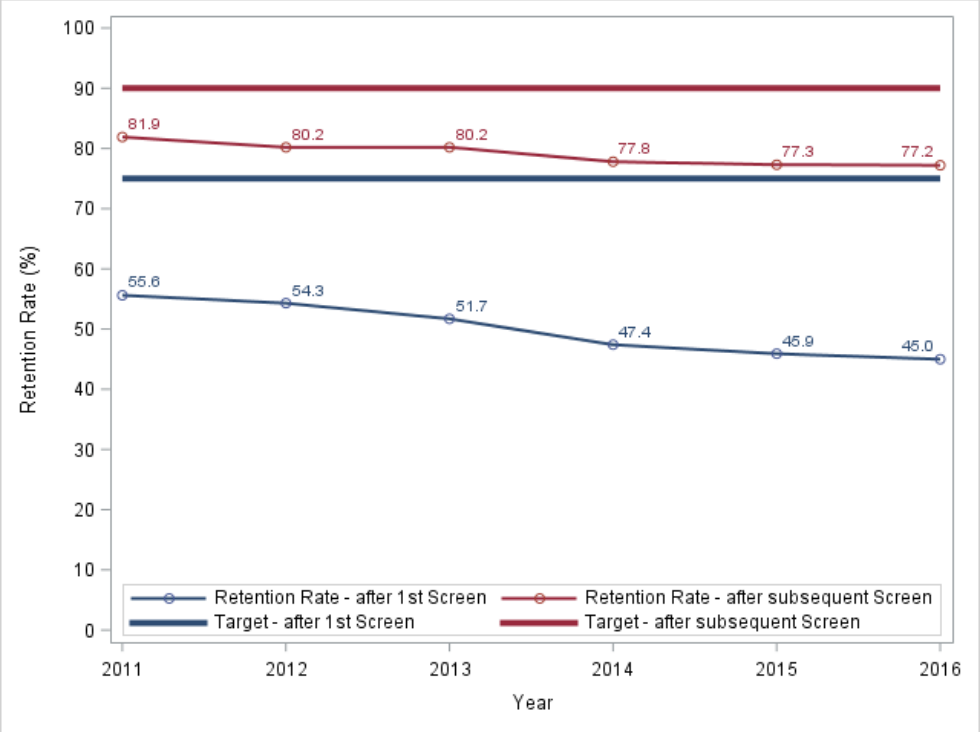
Within the BC Screening Mammography Program (SMP), the data collected is analyzed on an ongoing basis to monitor the program's efficacy and identify areas for improvement. For a screening program to be effective in reducing mortality and provide optimal benefit women not only need to participate in screening initially (first screens), they also need to return for ongoing screening at the appropriate interval (retention) [7].

In BC there are 5 major regional health authorities (HAs). Within each of the 5 HAs there are 3-4 health service delivery areas (HSDAs). Retention rates for the 16 HSDAs are monitored by the program.

THE PROBLEM – RETENTION RATES DECLINING

The BC Screening Mammography Program follows nationally set standards for retention rates. It has a target of ≥75% for women after their first screen and ≥90% for women who have had more than one screen. The BC SMP has experienced a decline in recent years and retention rates have fallen below targets (see Figure 1).

Figure 1: Retention Rates after 1st and Subsequent Mammography Screen for Women in Target Age Group 50-69, Years 2011-2016



MITIGATION EFFORTS – INVESTIGATIVE STUDY

One of the main elements of customer retention is building loyalty. A major component of this loyalty is emotional attachment [8]. Patients often feel the need to create – and often do create – a personal relationship with their GP [9]. The emotional attachment created by the patient-GP relationship lends itself to the GP playing a key role in shaping their patients' engagement in health-related behaviors. In fact, randomized control studies have been conducted where it has been shown that women's participation in breast screening increases if they received a letter from their GP rather than from the screening program [10, 11].

Currently, in the BC system, all women in the target age group receive invitation and reminder letters from the SMP. The SMP therefore performed an investigative study in one HSDA to examine the efficacy of GP invite letters on retention of women who had not returned to the screening program. It was initially planned to be a randomized control trial but due to cost and practical issues one could not be implemented. Instead the SMP invited all 344 GPs practicing as of 1st January 2016 within one HSDA to participate.

Participation included the GP sending an invitation letter to the female patients in their practice who were overdue for mammography screening (the intervention condition). Women were aged 50-69 and overdue by 24-96 months as of 31st December 2015. If a GP chose not to participate, the standard SMP invite letter was sent to female patients in their practice fitting the same criteria (the control condition). The SMP wanted to know if return rates were higher in women who received a letter from the GP compared to those who received the normal reminder letter from the SMP. A statistician was brought in after the completion of the study to analyze the data.

ANALYSIS DESIGN

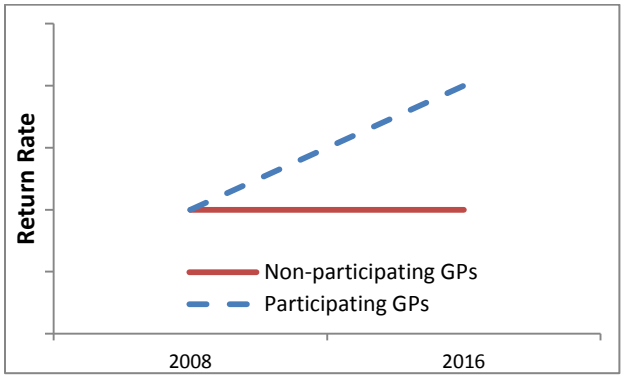
In order to try to estimate the success of the use of the GP invite letter, the least biased control group was chosen. There are inherent biases in using GPs who chose not to participate as the control. The fact they did not participate may indicate they are generally less engaged in breast screening so they could have a lower underlying retention rate compared to the GPs who agreed to participate. In this scenario, if GPs who participated were found to have a higher retention rate it would be difficult to know how much was due to the letter sent by the GP. To avoid this problem we look at the retention rate of the same GPs but from an earlier year and use the earlier year as a baseline.

Women included in the current study were all women who were overdue for a screen up to 8 years. In order to avoid overlap of women in the control condition, 2008 was used as the baseline year. Return rates of participating and non-participating GPs in 2016 were compared to their return rates in 2008. Four GPs were not practicing in 2008 and were excluded. GPs were classified as participated if they agreed to send invite letters in 2016 pilot study and non-participating if they chose not to.

HYPOTHESIS

If GPs who agreed to **participate** in 2016 send invite letters to women under their practice and the women under GPs who did not participate continue to receive the standard SMP invite letter then the retention rates of women under participating GPs will **improve** in 2016 compared to 2008 because patients in 2016 will be influenced by the GP's endorsement of breast screening, whereas the women in 2008 will have only received the SMP standard letter. In addition, the retention rates of women under the practice of GPs who did **not participate** will **not improve** in 2016 because neither group is influenced by the receipt of a GP invite letter, as per the pattern shown in Figure 2.

Figure 2: Hypothesis of Change in Retention Rate from 2008 to 2016



CHARACTERISTICS OF THE DATA STRUCTURE

- It is highly likely that women under the care of the same GP will be more similar to one another than they will to women who are seeing a different GP. For example, one GP, due to their own beliefs, may have different opinions on the effectiveness of breast screening women and this belief may have underlying influence on women that attend their practice but would not affect women under other GP practices. This leads to variability not only between the GPs but also between women who are nested within the GP. For analysis this means the assumptions of independent and identically distributed observed outcomes do not hold.
- There are 2 levels of information available – GPs and women within GP. Since women are nested within GP, each GP can be thought of as a 'cluster'.
- There is only one observation per woman.
- Each observation is unique to a specific cluster. As there is only one observation per woman, a woman can only be under one GP.
- The data is unbalanced. A different number of women are within a GP.
- The number of observations within a cluster varies quite considerably. The median number of women under one GP is 30 and ranges from 4 to 176.
- There are a large number of clusters. A total of 340 GPs; 205 participated, 135 did not.
- The data are nested, not longitudinal and so it is unlikely information from women observed would inform data on women not observed. It therefore seems reasonable to assume data are missing completely at random (MCAR) rather than missing at random (MAR).
- In each cluster the variable of interest takes only one value. All women under one GP received the same type of letter invite, whether that is the standard SMP letter or the letter from the GP.

VARIABLES

The outcome (returned) is binary, defined as return to screening within 6 months of the invite letter or not. It is measured at level 1, i.e. the outcome is for each woman. The intervention of interest (participate) is at level 2, i.e. it is a cluster-varying covariate as it informs whether the GP participated in the study or not. We also have level 1-invariant (or women-invariant) covariates as follows:

- Amount of time elapsed since a woman's last screen, categorized into 2 groups; time lapsed 24 to 36 months and 37-96 months (lapse_gp)
- Number of previous mammography screens a woman has had, categorized as 1, 2, 3, 4, 5, 6, 7 and 8+ (total_screens)
- Woman's age at her last screen grouped in 3 age groups; 49-53, 54-57, 58-70 (age_gp)
- Period at which point the woman had not returned for screening (2008 or 2016) (period)

CONSIDERATIONS FOR MODEL SELECTION

If a **standard logistic regression** is used and the cluster structure is ignored, the model is likely to **underestimate the standard errors of cluster-varying covariates** and hence inflate the Type I error. Not taking the cluster structure into account can also **overestimate the standard errors of level 1-invariant covariates** because the between-cluster variability has not been accounted for, causing Type II error inflation. In general, if there is failure to take the clustering into account, the stronger the intra-cluster variation is, the more profound the effect will be on the statistical analysis.

To account for this inherent clustering or correlation, there are several statistical modeling methods that can be used for a binary outcome. Typically these can be grouped into two classes; marginal or population averaged (PA) approach, and conditional or cluster specific (CS) approach. A GEE approach is often used for the PA approach. The GEE and CS models address the problem of correlated observations but incorporate it into the model differently. For binary outcomes, the regression coefficients or odds ratios obtained from the two approaches are numerically different, as are their interpretations, due to the non-linearity of the logistic regression function.

With respect to interpretation, the principle distinction between the two methods is the **GEE method** models the average response effect **across all clusters** i.e. it models the marginal expectations of the outcome. In a **CS model**, the response effect is specific **for a given cluster**. The decision of which one to use should therefore depend on the objectives of the study.

GEE

To obtain a marginal probability, GEE uses a quasi-likelihood approach [12] which separately models the mean response across all clusters and the within-cluster association, assuming the primary interest is in the mean response and the within-cluster correlation is just a nuisance that must be taken into account for valid inference. A cluster effect is not explicitly included in the model. Instead the within-cluster correlation is specified through a working correlation whose parameters are estimated by methods of moments. Rather than having a single constant variance and zero covariance for all the residuals (as in logistic regression), observations within the same cluster are allowed to have different variances and nonzero covariances. It also assumes the observations in different clusters are independent. The pattern of variances and covariances is specified by a working correlation structure. There are different specifications of the working correlation that can be used. These are described later in the paper.

Since the focus of GEE is on the 'mean model', the GEE should only be used if we want to estimate the average effect over the entire sample rather than estimate the effect for a particular cluster. It cannot make inferences about the reasons for the variation at the cluster level nor estimate cluster-specific effects. If you are interested in the cluster effect another method needs to be used. Relating this to our business case, the question a GEE could answer is:

How does the probability of a woman returning to screening mammography change if she received a GP invite letter compared to receiving the SMP standard letter?

The behavior of GEEs is asymptotic to the number of clusters so the more clusters you have, the better. If you do not have a large number of clusters, confidence intervals are too narrow and false positive rates are increased [13]. There is no formal cutoff for what quantifies as 'too few' clusters but some research puts the threshold of concern around 40 clusters [14]. With that said, MacKinnon and Webb found that if the clusters have different numbers of observations then as many as 100 clusters can still be problematic. For GEE, the computational complexity increases with the size of the largest cluster rather than the number of clusters. This can be an advantage if there are many clusters of smaller size. The smaller the number of clusters there are, the more important it is that the working correlation is accurately specified.

Currently, one drawback of the GEE is standard statistical software does not easily implement more than 2 clustering levels. For example, it could be feasible in addition to the correlation between women within the same GP, that GPs within the same clinic or within the same city would be correlated. If the marginal approach is the relevant analysis to perform and there are multiple clustering levels then modelling at the highest cluster level could be considered, especially if there is little evidence of a strong correlation within the dataset. The general idea would be to use bigger and more aggregate clusters when possible, up to the point at which there is concern about having too few clusters [15].

Both GEE and CS can handle unbalanced data. GEE works well if you have data missing and it is missing completely at random (MCAR). Under this assumption the GEE approach provides consistent estimators of the regression coefficients and of their robust variances even if the assumed working correlation is misspecified. However, if the missing data is related to some of the observed data i.e. missing at random (MAR) then it requires estimation of weights and a more complicated correlation structure. This can be performed in PROC GEE but is beyond the scope of this paper.

One advantage of the GEE model approach is that it involves fewer assumptions than the CS model and does not require knowledge of the ways in which individuals are correlated within a GP nor how GPs vary in order to provide consistent regression parameter estimates.

CLUSTER-SPECIFIC MIXED EFFECTS

The CS method estimates the variance among cluster means and among observations within a cluster. It partitions the variance into cluster-level and observation-level parts. It assumes the within-cluster correlation has a specific distribution and attempts to model it. For example, since the focus of using a CS model is the response pertaining to clusters, in our study, we could model the GP as a fixed effect, so each GP has its own intercept within the model. However, if we have many GPs, this leads to a high number of additional parameters (basically, the total number of GPs – 1) in the model with relatively few degrees of freedom left with which to try to accurately estimate the parameters of interest. A superior approach is to treat the intercepts relating to each GP as a random variable with a specified probability distribution. This is what the CS approach does, it treats the cluster effect as a random variable and thus, each cluster has a different intercept estimated from this random effect and the change in absolute risk for different levels of covariates from one cluster to another will depend on the baseline rate for the cluster. This leads to what is known as the random intercept model. It can also be extended to include a random slope effect. The concept of random intercepts and random slopes is shown in Figure 3 and Figure 4.

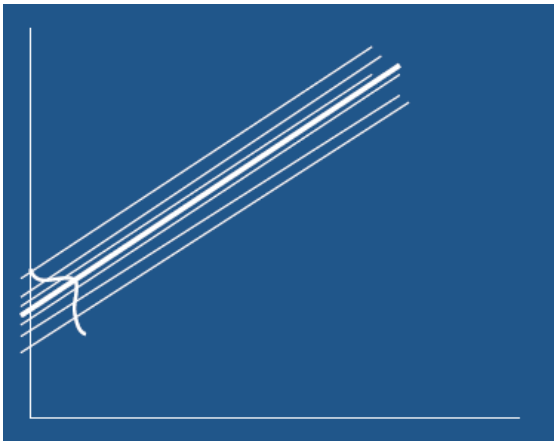


Figure 3: Random Intercepts

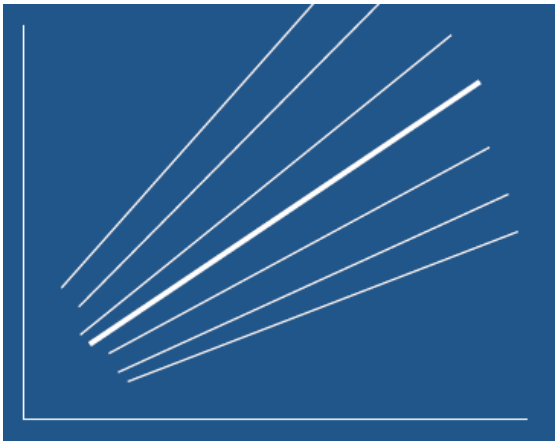


Figure 4: Random Intercepts and Slopes

Since the change in absolute risk for different levels of covariates from one cluster to another will depend on the baseline rate for the cluster there is no concept of 'average' across a covariate level in this scenario. The CS approach is therefore of greatest use when the main interest is to make inference about cluster differences or if the focus is to understand the role of a specific cluster on the outcome. It can estimate the cluster effect adjusting for covariates of interest, it can evaluate and compare the performance of the clusters with respect to the outcome and it can compare the difference in cluster effects between two clusters. In our case, the CS model could answer the following questions:

> *How does the probability of a woman returning to screening mammography change, if she received an invite letter from a specific GP compared to receipt of the standard SMP letter instead?*
>
> *How does the change in return rates differ if a woman received an invite letter from one GP compared to another GP?*

It is of interest to note, in our study, a CS model would describe the difference in retention when a particular GP sends invite letters compared to when they do not send invite letters. Since a single GP only either sends the GP letters or does not, the CS model estimates an unobserved effect. This can be concerning for some people and for this reason, it can often be recommended to use PA models for analyzing cluster-specific covariate.

For a CS model, the number of groups, number of level-one units, and the amount of dependency between the observations affect the accuracy of the estimates. Because the distribution of the cluster-specific intercepts is modeled, as with GEE it is better to have more clusters than more observations within a cluster. It has been shown that as few as 10 clusters, with 5 observations within each of the 10 clusters is on the verge of unacceptable, leading to biased (smaller) estimates of the level 2 (cluster-specific) standard errors. However, more than 50 clusters is ideal [16]. More cases may be needed for convergence if the model is more complex or data is unbalanced [17].

Even for a standard logistic regression, if only 10 observations were collected, one would have limited confidence in the inference. The same is true for GEE and CS modeling; if information on only 10 clusters is collected it is unlikely accurate estimates of the variability are obtained.

CS modeling assumes that the missing data are dependent on the covariates and observed responses (MAR) which can be preferable and the advantage of using CS modeling with longitudinal data but is more difficult to justify in the case of only clustered data (e.g. in our study where women are nested within GPs) [20].

In a CS model, interpretation of regression coefficients and their estimates can be susceptible to biases from sensitive and difficult-to-verify assumptions about the random effects distribution, particularly the dependence of the latent variable distribution on covariates [19]. They can also be computationally intensive, especially as the number of random effects to be estimated increases beyond one or two.

MODEL SELECTION

For this study, the SMP wanted to evaluate if customer retention would improve using the GP invite letter compared to the standard SMP invite letter. It was not of interest to compare retention within specific GPs or between GPs, but rather examine the overall effect. This is the most important aspect that leads to the decision a PA model is more appropriate than a CS model.

WORKING CORRELATION STRUCTURE SELECTION

The next step is deciding how to specify the working correlation structure which describes the pattern of association amongst the observations that are within each cluster. Although in GEE, the working correlation is thought of as being a 'nuisance' parameter, and estimates from the model are meant to be robust even if the incorrect correlation structure is specified, it is good scientific practice to carefully select the correlation structure of your data. If you have small sample sizes or strong correlations, you can have substantive loss of efficiency in the estimation of the regression parameters if the working correlation is misspecified [21].

SAS® provides a choice of a few correlation structures that can be used:

- **Independent** (IND): This structure assumes zero correlation between measurements within the same cluster and is recommended when there are few clusters [22]. Our data has many clusters (340 GPs).
- **Autoregressive** (AR): The AR working correlation forces the correlation between consecutive measurements on a subject (cluster) to decrease with increased separation in time or space. This is often relevant when modelling longitudinal data where measurements in time are approximately equally spaced but this is not applicable for our data. The different women under the same GP are not linked by measurement of time or distance.
- **Unstructured** (UN)**:** The unstructured correlation assumes that correlation between pair of residuals is unique. This is straight forward only if you have balanced data [23]. For studies involving few measurements, the flexibility of the unstructured correlation can make it the optimal choice. However, when there are many measurements within a cluster, so much computational power is required it is often inestimable. For our study, as mentioned previously, data is not balanced and we can have many women under one GP.
- **Exchangeable** (EXCH): This structure assumes all pairs of observations in the same cluster have the same correlation i.e. there is no logical ordering. It is often the standard choice if the correlation structure is unknown and can be useful for clustered observations such as students in the same classroom or members of the same household. This is the scenario in our business case.

There is no statistical test to assess the correctness of the working correlation structure but understanding the study design and data can guide which is the most applicable correlation structure to use. If there is uncertainty, the quasi-likelihood information criteria (QIC) statistic can be used to compare correlation structures. The one resulting in the smallest QIC value is preferred. The QIC statistic is analogous to the likelihood-based Akaike information criteria (AIC).

The estimation of parameter standard errors can be model-based or empirical. The model-based standard error is based on the estimated correlation (i.e. the specified working correlation matrix). The empirical method, also known as the robust standard error method, uses the actual variation in the cluster-level statistics. Unless data are sparse, the empirical standard error is more trustworthy than the model-based one [24] and is preferred when using GEE. If the correlation matrix is totally accurate, model-based and empirical standard errors would be the same. Hence, agreement between the model-based and empirical standard errors will suggest that the assumed correlation structure is reasonable. Even if the working correlation structure is not the true correlation, the empirical standard error provides a consistent estimator [24].

In summary, a GEE model using an exchangeable correlation structure is used for this study and the agreement between the model-based and empirical standard errors are investigated.

IMPLEMENTING THE MODEL USING PROC GENMOD

The following code was used to implement the model:

```
ODS GRAPHICS ON;
PROC GENMOD DATA=study DESCEND;
  CLASS gp_id total_screens age_group lapse_group period (REF="2008") participate (REF="0")
        / PARAM=GLM;
  MODEL returned = total_screens age_gp lapse_gp period participate period*participate
       / DIST=BIN LINK=LOGIT TYPE3;
  REPEATED SUBJECT=gp_id / TYPE=EXCH MODELSE;
RUN;
```

The DATA= specifies the dataset to use as in all the modelling procedures.

The default for ordering the response variable is the order of the format for the variable. For a binomial response, the DESCEND option specifies that the levels of the response variable be sorted in the reverse of the default order. In our data it is ordered 0 (=woman did not return) then 1 (=woman returned) and so using the DESCEND option *SAS*® models the probability of returned=1.

The first part of the output generated by PROC GENMOD is the model information, shown in Output 1. It shows there are 11,660 observations with 2,173 women who returned to screening.

Output 1: Model Information

The GENMOD Procedure

Model Information	
Data Set	WORK.STUDY5
Distribution	Binomial
Link Function	Logit
Dependent Variable	returned

Number of Observations Read	11660
Number of Observations Used	11660
Number of Events	2173
Number of Trials	11660

The CLASS statement and the MODEL statement specify the model for the mean of the response variable, returned, as a logistic regression with total_screens, age_gp, lapse_gp, period and participate as independent variables, just as for a typical logistic regression in PROC LOGISTIC. Including the interaction of period and participate enables the investigation of whether the return rate of participating GPs compared to non-participating GPs differed in 2008 compared to 2016.

PARAM=GLM specifies the GLM parameterization global-option be used, although, since this is the default it is not necessary to specify it. The GLM parameterization has been used as it is required for the LSMEANS statement which is used later. It is also required if the TEST, LSMESTIMATE, and SLICE statements are to be used. Using the global-option PARAM=GLM and the individual variable option REF= for period and participate variables, requests that the specified levels of the two classification variables correspond to the reference level. Since a reference level was not specified for the other

variables, the last formatted value in alphabetic order is chosen as the reference level (the default). The class level information for the model can be seen below in Output 2.

Output 2: Class Level Information

Class Level Information		
Class	Levels	Values
gp_id	340	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 ...
total_screens	8	1 2 3 4 5 6 7 8
age_group	3	49-53 54-57 58-70
lapse_group	2	25-36 mo 37-96 mo
period	2	2008 2016
participate	2	0 1

In the MODEL statement the DIST=BIN and LINK=LOGIT options are used to specify a logistic regression. If only the DIST=BIN was specified then the default canonical link function for the distribution specified is used. Since GEE is not a likelihood-based method, inference is not based on likelihoods and the default analysis is to compute score statistics. Since the REPEATED statement has been used and TYPE3 specified in the MODEL statement, generalized score tests for Type III contrasts are produced (Output 3). Wald tests for Type III GEE analysis can also be obtained by additionally specifying WALD in the MODEL statement. The Wald test can be used when the number of clusters is large but when the number of clusters is small it is recommended to use the score test as it performs better for hypothesis testing for correlated data [25, 26].

From Output 3 we can see that all effects included in the model are statistically significant, apart from period. However, since we require the interaction of period and participate to investigate whether the return rate of participating GPs compared to non-participating GPs differed in 2008 compared to 2016, period is kept in the model.

Output 3: Score Statistics for Type 3 GEE Analysis

Score Statistics For Type 3 GEE Analysis			
Source	DF	Chi-Square	Pr > ChiSq
total_screens	7	153.32	<.0001
age_group	2	25.81	<.0001
lapse_group	1	105.63	<.0001
period	1	1.49	0.2216
participate	1	38.22	<.0001
period*participate	1	10.95	0.0009

The REPEATED statement informs PROC GENMOD to fit a GEE with an exchangeable correlation structure (TYPE=EXCH). The option SUBJECT= identifies the 'subject-effect' or cluster effect. Each distinct value of the variable specified in SUBJECT= identifies a different cluster whereby responses from different clusters are assumed to be statistically independent, and responses within clusters are assumed to be correlated. In our case, the gp_id variable identifies the GP cluster. Each woman in our dataset has a unique id (as per Table 1). However, if women had a unique id within the cluster (as per Table 2) then SUBJECT=woman_id(gp_id) or the equivalent, SUBJECT=woman_id*gp_id, would need to be specified instead.

You will notice gp_id is also listed in the CLASS statement since the effect specified as the SUBJECT= effect must also be listed in the CLASS statement.

Table 1: Unique ID

woman_id	gp_id
1	1
2	1
3	1
4	2
5	2
6	2

Table 2: Unique ID within GP

woman_id	gp_id
1	1
2	1
3	1
1	2
2	2
3	2

GEE model information in Output 4 shows we have 340 GPs, with the minimum and maximum number of woman within a GP of 4 and 176, respectively.

Output 4: GEE Model Information

GEE Model Information	
Correlation Structure	Exchangeable
Subject Effect	gp_id (340 levels)
Number of Clusters	340
Correlation Matrix Dimension	176
Maximum Cluster Size	176
Minimum Cluster Size	4

Our code requests an exchangeable working correlation matrix. This matrix assumes all pairs of observations in the same cluster have the same correlation. Hence, there is only one value of interest. *SAS*®, by default, prints a table that contains the single estimated correlation, as shown in Output 5. However, had a different correlation matrix been specified, the CORRW option in the REPEATED statement can be used to output the matrix. In Output 5, we see the correlation between women within the same GP in our study is actually relatively small.

Output 5: Exchangeable Working Correlation

Exchangeable Working Correlation	
Correlation	0.0071330977

The QIC value for our model is shown in Output 6. The QICu statistic is also automatically output. The QIC can be used to compare different correlation structures. The QICu statistic adds a penalty for the number of parameters and so can be used to compare different models. The parameters in one model do not need to be a subset of the parameters in the other model. For both statistics, the smaller the value

the better. In our scenario it is known that the number of previous screens, the length of time since the last screen and age all influence the probability of a woman returning. Thus we want to adjust for these parameters rather than perform model selection so for the purposes of our study we do not need to use QICu.

Output 6: GEE Fit Criteria

GEE Fit Criteria	
QIC	10016.4047
QICu	10011.6860

By default, the Analysis of GEE Parameter Estimates table provides standard errors, confidence intervals, Z scores and p-values based on empirical standard errors. Using the MODELSE option in the REPEATED statement requests an *additional* parameter estimates table which uses the model-based standard errors.

For ease of comparison, Table 3 provides only the estimates, standard errors and p-values for analyses based on both empirical and model-based standard errors. The last column provides the percent difference in model-based standard error estimates compared to the empirical standard error estimates. The model-based parameter standard errors are generally smaller than empirical standard error estimates and can be as much as 29% smaller.

Since there are some differences in the standard errors between empirical and model-based estimates, an unstructured working correlation structure could be considered instead (the AR is not plausible in this scenario). However, the unstructured matrix introduces many more computations that may become far too complex for the amount of data available. This is certainly the case in our scenario. In fact, performing the analysis but instead specifying an unstructured working correlation (TYPE=UN) produces the warning and error messages within the *SAS*® log, shown in Output 7, indicating this specific issue.

Output 7: Log Warning and Error Messages when Unstructured Correlation Working Matrix is specified

```
NOTE: Class levels for some variables were not printed due to excessive size.
NOTE: PROC GENMOD is modeling the probability that returned='1'.
NOTE: Algorithm converged.
WARNING: The number of response pairs for estimating correlation is less than or equal to the
         number of regression parameters. A simpler correlation model might be more
         appropriate.
NOTE: The working correlation has been ridged with a maximum value of 18.447931549 to avoid
      singularity.
NOTE: The working correlation has been ridged with a maximum value of 13.396272134 to avoid
      singularity.
NOTE: The working correlation has been ridged with a maximum value of 4.0911454974 to avoid
      singularity.
ERROR: Error in computing the variance function.
ERROR: Error in parameter estimate covariance computation.
ERROR: Error in estimation routine.
NOTE: The scale parameter was held fixed.
NOTE: The SAS System stopped processing this step because of errors.
NOTE: PROCEDURE GENMOD used (Total process time):
      real time           1.95 seconds
      cpu time            1.70 seconds
```

Table 3: Empirical and Model-Based Standard Error Estimates

			Empirical Standard Error Estimates			Model-Based Standard Error Estimates			
Parameter			Estimate	Std Error	p-value	Estimate	Std Error	p-value	% Diff
Intercept			-1.3774	0.1278	<.0001	-1.3774	0.1081	<.0001	-15
total_screens	1		-2.0872	0.1183	<.0001	-2.0872	0.1128	<.0001	-5
total_screens	2		-1.0223	0.0901	<.0001	-1.0223	0.0918	<.0001	2
total_screens	3		-0.7709	0.0876	<.0001	-0.7709	0.0913	<.0001	4
total_screens	4		-0.6059	0.0946	<.0001	-0.6059	0.0939	<.0001	-1
total_screens	5		-0.3762	0.1029	0.0003	-0.3762	0.0981	0.0001	-5
total_screens	6		-0.2877	0.1008	0.0043	-0.2877	0.1041	0.0057	3
total_screens	7		-0.2892	0.1177	0.0140	-0.2892	0.1154	0.0122	-2
total_screens	8		0	0	.	0	0	.	
age_group	49-53		0.3395	0.0684	<.0001	0.3395	0.0686	<.0001	0
age_group	54-57		0.1934	0.0605	0.0014	0.1934	0.0623	0.0019	3
age_group	58-70		0	0	.	0	0	.	
lapse_group	25-36 mo		0.8136	0.0538	<.0001	0.8136	0.0516	<.0001	-4
lapse_group	37-96 mo		0	0	.	0	0	.	
period	2016		-0.1839	0.1371	0.1798	-0.1839	0.0972	0.0585	-29
period	2008		0	0	.	0	0	.	
participate	1		0.1942	0.1195	0.1041	0.1942	0.0994	0.0508	-17
participate	0		0	0	.	0	0	.	
period*participate	2016	1	0.5528	0.1543	0.0003	0.5528	0.1166	<.0001	-24
period*participate	2016	0	0	0	.	0	0	.	
period*participate	2008	1	0	0	.	0	0	.	
period*participate	2008	0	0	0	.	0	0	.	

Odds ratios for comparisons of interest can be obtained by adding the LSMEANS statement to the code. Our main interest is in the interaction effect of period*participate as specified below:

```
ODS GRAPHICS ON;
PROC GENMOD DATA=study DESCEND;
  CLASS gp_id total_screens age_group lapse_group period (REF="2008") participate (REF="0")
        / PARAM=GLM;
  MODEL returned = total_screens age_gp lapse_gp period participate period*participate
       / DIST=BIN LINK=LOGIT TYPE3;
  REPEATED SUBJECT=gp_id / TYPE=EXCH MODELSE;
  LSMEANS period*participate / CL DIFF=ALL ODDSRATIO;
RUN;
```

The CL option requests t type 95% confidence limits (CIs) (default) for each of the LS-means. This can be changed using ALPHA= option. The DIFF=ALL option requests that LS-means differences be provided on all pairwise comparison. This is the default so just DIFF could be specified to obtain the same output.

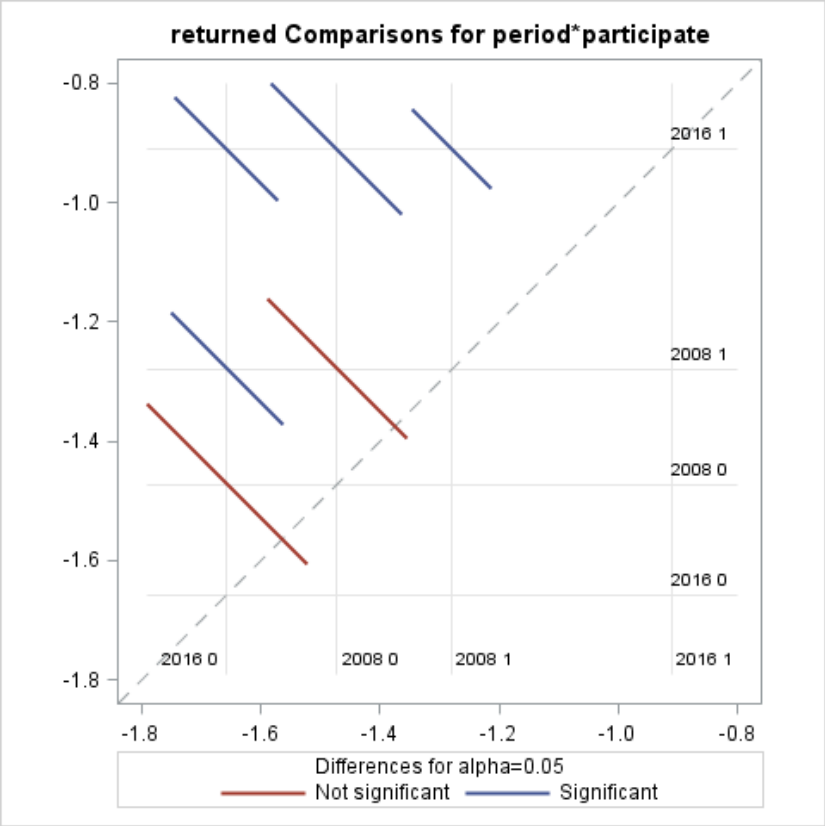
By adding ODDSRATIO to the statement the odds ratios for the LS-means are also reported. As CL is specified the 95% CIs for the odds ratios are output (Output 8). Since ODS GRAPHICS ON is specified the default plot is also provided (Figure 5).

Output 8: Differences of LS-Means Requested Along with Odds Ratios and 95% Confidence Intervals

Differences of period*participate Least Squares Means

period	ITT_participate	_period	ITT_participate	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper	Odds Ratio	Lower Confidence Limit for Odds Ratio	Upper Confidence Limit for Odds Ratio
2016	1	2016	0	0.7470	0.08849	8.44	<.0001	0.05	0.5735	0.9204	2.111	1.775	2.510
2016	1	2008	1	0.3689	0.06809	5.42	<.0001	0.05	0.2355	0.5024	1.446	1.266	1.653
2016	1	2008	0	0.5631	0.1118	5.04	<.0001	0.05	0.3441	0.7822	1.756	1.411	2.186
2016	0	2008	1	-0.3780	0.09607	-3.94	<.0001	0.05	-0.5663	-0.1898	0.685	0.568	0.827
2016	0	2008	0	-0.1839	0.1371	-1.34	0.1798	0.05	-0.4525	0.08480	0.832	0.636	1.088
2008	1	2008	0	0.1942	0.1195	1.63	0.1041	0.05	-0.03998	0.4284	1.214	0.961	1.535

Figure 5: Default Plot (Diffogram) for Differences Requested in LSMEANS Statement



It can also be useful to obtain the predicted probabilities of a woman returning to breast screening based on whether their GP participated and the period they had not returned in.

This can be obtained specifying the ILINK option in a LSMEANS statement as shown below. Estimates of predicted probabilities and their standard errors are provided along with 95% CIs because the CL option is also specified as shown in Output 9.

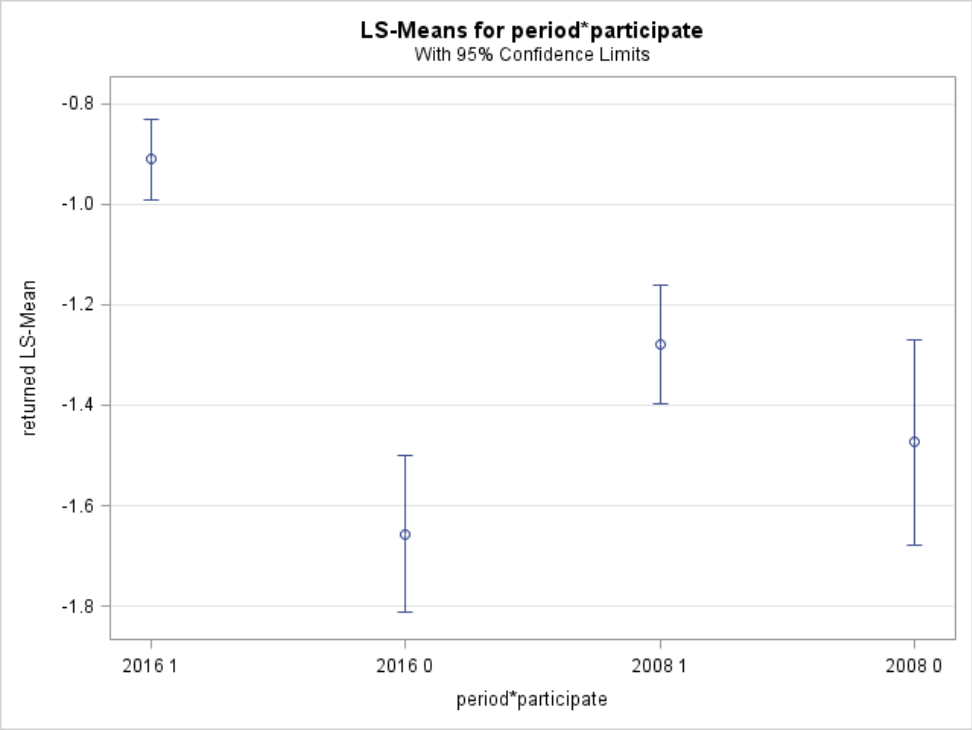
Once again, because ODS GRAPHICS ON is specified, the default plot for predicted probabilities is also provided (Figure 6).

```
ODS GRAPHICS ON;
PROC GENMOD DATA=study DESCEND;
  CLASS gp_id total_screens age_group lapse_group period (REF="2008") participate (REF="0")
          / PARAM=GLM;
  MODEL returned = total_screens age_gp lapse_gp period participate period*participate
        / DIST=BIN LINK=LOGIT TYPE3;
  REPEATED SUBJECT=gp_id / TYPE=EXCH MODELSE;
  LSMEANS period*participate / CL DIFF ODDSRATIO;
  LSMEANS period*participate / CL ILINK;
RUN;
```

Output 9: Predicted Probabilities and 95% Confidence Intervals Requested using ILINK and CL option

period*participate Least Squares Means												
period	ITT_participate	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper	Mean	Standard Error of Mean	Lower Mean	Upper Mean
2016	1	-0.9098	0.04112	-22.12	<.0001	0.05	-0.9904	-0.8292	0.2870	0.008415	0.2708	0.3038
2016	0	-1.6568	0.07937	-20.87	<.0001	0.05	-1.8123	-1.5012	0.1602	0.01068	0.1404	0.1822
2008	1	-1.2787	0.05977	-21.39	<.0001	0.05	-1.3959	-1.1616	0.2178	0.01018	0.1985	0.2384
2008	0	-1.4729	0.1042	-14.13	<.0001	0.05	-1.6771	-1.2686	0.1865	0.01581	0.1575	0.2195

Figure 6: Default Plot (LS-Means Plot) when Requesting Predicted Probabilities



Multiple LSMEANS statements can be specified within one PROC GENMOD, as in the code above. Alternatively, both the odds ratios and predicted probabilities could be combined into just one LSMEANS statement by adding the ILINK option to the first LSMEANS statement, as below:

```
ODS GRAPHICS ON;
PROC GENMOD DATA=study DESCEND;
  CLASS gp_id total_screens age_group lapse_group period (REF="2008") participate (REF="0")
        / PARAM=GLM;
  MODEL returned = total_screens age_gp lapse_gp period participate period*participate
       / DIST=BIN LINK=LOGIT TYPE3;
  REPEATED SUBJECT=gp_id / TYPE=EXCH MODELSE;
  LSMEANS period*participate / CL DIFF ODDSRATIO ILINK;
RUN;
```

However, only the 'returned comparisons' plot would be provided (Figure 5), the LSMEANS plot (Figure 6) would not be.

INTERPRETING THE RESULTS

If the GP invite letter is a successful customer retention strategy then we can expect to see the probability of women returning in 2016 increase more for the GP participating group compared to the non-participating GP group accounting for baseline levels in 2008.

The interaction effect for period and participate, of 0.5528, based on inference using an exchangeable working correlation structure and the empirical standard errors shown in Table 3, is significant (p=0.0003). This indicates the probability of a women returning increased in 2016 compared to 2008 and more so for the GP intervention group than for the standard invite letter group. This pattern can also be seen quite clearly in Figure 6. From Output 8, we can see the odds of a woman, returning in 2016 when under a participating GP is 1.45 times (95% CI: 1.27 to 1.65) that of women under a participating GP back in 2008, when the standard SMP invite letter was used, not the GP letter. However, in the non-participating GPs, where in both 2008 and 2016 only the standard SMP invite letter was used, there was no significant change (OR: 0.83; 95% CI: 0.64 to 1.09). The probability of a woman returning to breast screening increased by 12.7% (from 16% in 2008 to 29% in 2016) if they were a client of a GP willing to participate and send women in their practice an invite letter.

It is clear that the use of a GP invite letter increases customer retention within the BC SMP.

CONCLUSION

Due to the success of this customer retention strategy, the BC SMP has started to implement GP invite letters across a few other areas where customer retention was low. However, as you can imagine, the amount of increase in the number of women returning to breast screening can have significant impacts on costs and resources for the SMP across their multiple sites in British Columbia and further strategic planning is required before the decision to implement this province wide is made.

Although this example is health related, the concept applies to all industries. Improper analysis of correlated data can lead to erroneous statistical inference. If the structure of the data informs that correlation between the observations is present, appropriate analyses accounting for the correlation, should be performed. This paper examined the appropriateness of two models, the CS model and the GEE, for investigating the success of a new customer retention strategy within the breast screening environment. With a binary outcome, results from a CS model and a GEE model are not equivalent. The CS model parameter interpretation is cluster specific where as in GEE it is population average. After consideration of the objectives of the study and the nuances and structure of the data were understood, the choice of which of the two methods to use was clear and a GEE model was implemented successfully.

Analysis is not always straightforward. Understanding the assumptions and limitations of the methods and using a common sense approach is required.

REFERENCES

[1] KY Liang, SL Zeger (1986). Longitudinal data analysis using generalized linear models. Biometrika; 73: 13-22.

[2] MP Coleman et al. 2010. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995–2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. The Lancet; 377(9760):127-138.

[3] P Gøtzsche, KJ Jørgensen (2013). Screening for breast cancer with mammography. Cochrane Database Syst Rev.6, CD001877.

[4] Independent UK Panel on Breast Cancer Screening (2012). The benefits and harms of breast cancer screening: an independent review. Lancet; 380:1778-86.

[5] H Vainio, F Bianchini (2002). IARC Handbook of Cancer Prevention. Volume 7. Breast Cancer Screening. Lyon, France: IARC Press.

[6] LL Humphrey, M Helfand, BK Chan, SH Woolf (2002). Breast cancer screening: a summary of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med; 137:347-60.

[7] Public Health Agency of Canada (2011). Organized Breast Cancer Screening Programs in Canada - Report on Program Performance in 2005 and 2006. Canadian Breast Cancer Screening Initiative. Cat.: HP32-1/2006E-PDF ISBN: 978-1-100-17774-8.

[8] D Baack (2012). Customer Retention and the Health Care Industry. Economics and Business Journal: Inquiries and Perspectives; 4(1):35-43.

[9] HB Gelund Frederiksen, J Kragstrup, B Dehlholm-Lambertsen (2010). Attachment in the doctor-patient relationship in general practice: A qualitative study. Scandinavian Journal of Primary Health Care; 28: 185–190.

[10] KM Turner, BJ Wilson, FJ Gilbert (1994). Improving breast screening uptake: persuading initial non-attenders to attend. J Med Screen; 1: 199-202.

[11] D Giorgi, L Giordano, C Senore, G Merlino, R Negri, M Cancian, M Lerda, N Segnan, MR Del Turco (2000). General practitioners and mammographic screening uptake: influence of different modalities of general practitioner participation. Working Group. Tumori; 86(2): 124–129.

[12] RWM Wedderburn (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. Biometrika; 61 (3): 439-447.

[13] J Esarey, A Mengerz (2017). Practical and E_ective Approaches to Dealing with Clustered Data. http://jee3.web.rice.edu/cluster-paper.pdf

[14] S Teerenstra, B Lu, JS Preisser, T van Achterberg, and GF Borm (2010). Sample size considerations for GEE analyses of three-level cluster randomized trials. Biometrics; 66(4): 1230–1237.

[15] AC Cameron, DL Miller (2015). A Practitioner's Guide to Cluster-Robust Inference. The Journal of Human Resources; 50(2):317-372.

[16] Maas and JJ Hox (2005). Sufficient Sample Sizes for Multilevel Modeling Methodology. CJM; 1(3):86-92.

[17] SW Raudenbush, X Liu (2000). Statistical power and optimal design for multisite randomized trials. Psychological Methods; 5:199-213.

[18] M Parzen, S Ghosh, Stuart Lipsitz, D Sinha, GM Fitzmaurice, BK Mallick, JG Ibrahi (2011). A generalized linear mixed model for longitudinal binary data with a marginal logit link function. Ann Appl Stat. 2011 ; 5(1): 449–467.

[19] SL Zeger, K-Y Liang, PS Albert (1988). Methods for longitudinal data: a generalized estimating equation approach. Biometrics; 44(4):1049-60.

[20] RD Gibbons, D Hedeker, S DuToit (2010). Advances in Analysis of Longitudinal Data. Annual Review of Clinical Psychology; 6:79-107.

[21] Y-G Wang, V Carey (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. Biometrika; 90(1): 29-41.

[22] NJ Horton, SR Lipsitz (1999). Review of software to fit Generalized Estimating Equation regression models. The American Statistician; 53:160–169

[23] J Shults, W Sun, X Tu, H Kim, J A, JM Hilbe, T Ten-Have (2009). A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data. Statistics in Medicine; 28:2338–2355

[24] JA Hanley, A Negassa, MD deB Edwardes, JE Forrester (2003). Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation. American Journal of Epidemiology; 157(4): 364-375

[25] X Guo, W Pan, JE Connett, PJ Hannan, SA French (2005). Small-sample performance of the robust score test and its modifications in generalized estimating equations. Statistics in Medicine; 24(22):3479-3495

[26] X Guo, W Pan. Small-sample Performance of the Score Test in GEE (2002). http://www.sph.umn.edu/faculty1/wp-content/uploads/2012/11/rr2002-013.pdf

ACKNOWLEDGMENTS

I would like to thank my team in Cancer Surveillance & Outcomes, BC Cancer Agency, for their support and discussions during my writing of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Colleen McGahan
Cancer Surveillance & Outcomes, British Columbia Cancer Agency
cmcgahan@bccancer.bc.ca