Informative cluster size in observational studies

Aya Mitani

November 1, 2019

Research Fellow, Department of Biostatistics, Harvard Chan School of Public Health

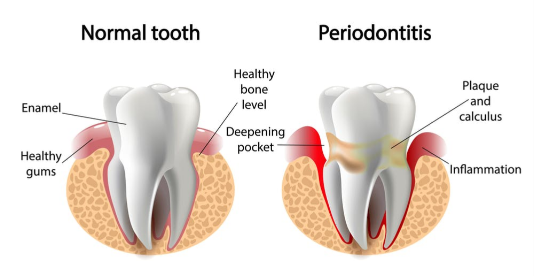
Outline

- Background and Motivation
- Marginal analysis of multiple correlated outcomes with ICS
- ICS in HIV/STD research
- ICS in other settings

ICS: Informative cluster size

Periodontal Disease

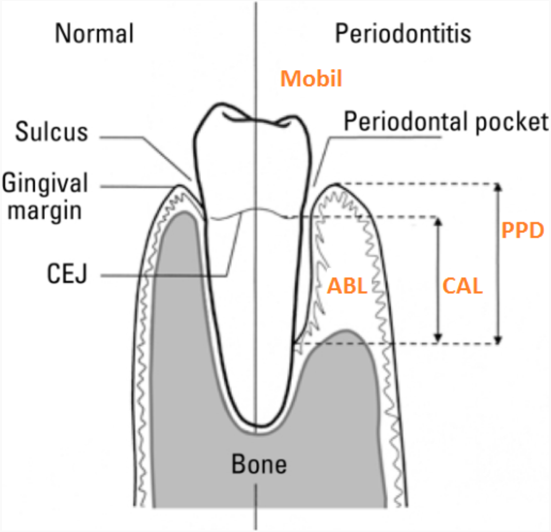
- Inflammatory disease affecting gums and bones surrounding teeth
- Progress is measured by many factors including clinical attachment loss (CAL)
- Mild periodontal disease – swollen and bleeding gums
- Severe periodontal disease – loosening teeth and teeth loss



Periodontal Disease

- Affects 30-50% of adult population in US
- Associated with
  - Age
  - Smoking
  - Low SES
  - Cardiovascular disease
  - Diabetes
  - HIV
  - ? Metabolic syndrome or MetS (Presence of $\geq$3 of the 5 following metabolic risk factors)
    1. Large waistline ($\geq$102 cm)
    2. High triglyceride level ($\geq$150 mg/dl)
    3. Low HDL cholesterol level ($<$40 mg/dl)
    4. High blood pressure (SBP$\geq$130 or DBP$\geq$85 mmHg)
    5. High fasting blood sugar ($\geq$100 mg/dL or antidiabetic drug use)

Periodontal Disease



ABL: Alveolar bone loss; CAL: Clinical attachment loss;
Mobil: Mobility; PPD: Probing pocket depth

Periodontal Disease

No universal definition of advanced periodontal disease

	Periodontal disease outcomes			
	ABL	CAL	Mobil	PPD
Ordinal score	0: None			
	1: $<$20%	0: $<$2mm	0: None	0: $<$2mm
	2: 20-39%	1: 2-2.9mm	1: $<$0.5mm	1: 2-2.9mm
	3: 40-59%	2: 3-4.9mm	2: 0.5-0.9mm	2: 3-4.9mm
	4: 60-79%	3: $\geq$5mm	3: $\geq$1mm	3: $\geq$5mm
	5: $\geq$80%			

ABL: Alveolar bone loss

CAL: Clinical attachment loss

Mobil: Mobility

PPD: Probing pocket depth

Motivating data set: VA Dental Longitudinal Study

Table 1: Description of Veterans Affairs Dental Longitudinal Study (1981-2011)

Number of subjects	760
Percentage of Men	100%
Number of visits per subject	1-11
Subject-level baseline variables	Age, Education, etc.
Subject-level time-varying variables	MetS, Smoking, etc.
Tooth-level variables	PPD, CAL, ABL, Mobil
Baseline number of teeth per subject	1-28

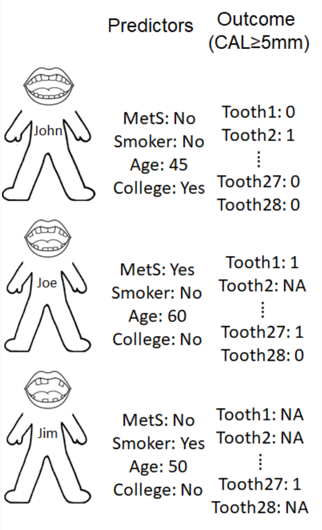
Kaye et al, 2016

Overall research question

What is the relationship between periodontal disease and MetS?

Overall research question

What is the relationship between periodontal disease and MetS?



Marginal models for clustered data

Notation

- $i = 1, ..., N$ Subjects
- $j = 1, ..., n_i$ teeth for ith subject at baseline
- $\mu_i = E(Y_i|X_i)$ where $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$

Generalized estimating equations (GEE)

$$\sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}' V_i^{-1}(Y_i - \mu_i) = 0$$

where $V_i = A_i^{1/2} R_i A_i^{1/2}$ and A_i is the diagonal matrix of variance $\mu_i(1 - \mu_i)$ and R_i is the working correlation matrix

Marginal models for clustered data

Notation

- $i = 1, ..., N$ Subjects
- $j = 1, ..., n_i$ teeth for ith subject at baseline
- $\mu_i = E(Y_i|X_i)$ where $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$

Generalized estimating equations (GEE)

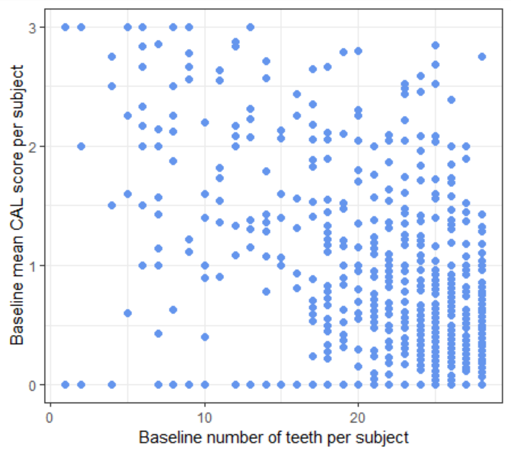
$$\sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta}' V_i^{-1}(Y_i - \mu_i) = 0$$

where $V_i = A_i^{1/2} R_i A_i^{1/2}$ and A_i is the diagonal matrix of variance $\mu_i(1 - \mu_i)$ and R_i is the working correlation matrix

Assumption of GEE
Independence between cluster size (number of teeth per subject, n_i) and outcome

Informative cluster size

Figure 1: Baseline number of teeth vs. mean CAL score
Pearson correlation coefficient = -0.470 (-0.553, -0.378)



Informative cluster size

What is informative cluster size (ICS)?

- Cluster size (number of teeth per subject, n_i) varies
- Outcome (CAL) is not independent of cluster size (number of teeth) given the exposure (MetS)

$$E(Y_i|X_i = x_i, n_i) \neq E(Y_i|X_i = x_i)$$

Issues with informative cluster size (ICS)

- Standard methods for clustered data analysis assume independence between outcome and cluster size
- When assumption is violated, analysis may result in biased estimates

Hoffman et al, 2001

Cluster weighted generalized estimating equations (CWGEE)

CWGEE for cross-sectional data

$$\sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}}' V_{ij}^{-1}(Y_{ij} - \mu_{ij}) = 0$$

- $E(\hat{\beta}_{CWGEE}) = \beta$
- $\sqrt{N}(\hat{\beta}_{CWGEE} - \beta) \xrightarrow{d} MN(\mathbf{0}, \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1})$ where
  - $\boldsymbol{B} = \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}}' V_{ij}^{-1} \frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}}$
  - $\boldsymbol{M} = \sum_{i=1}^{N} \left[\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}}' V_{ij}^{-1}(Y_{ij} - \mu_{ij}) \right] \left[\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}}' V_{ij}^{-1}(Y_{ij} - \mu_{ij}) \right]'$

Williamson et al, 2003

Difference in inference

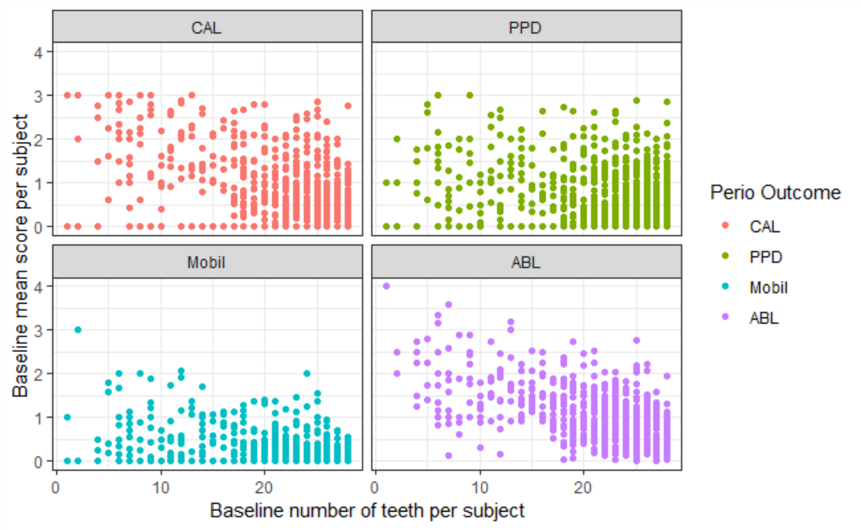
GEE with independence working correlation

- Inference for population of all units
- Larger clusters contribute more to inference than smaller ones
- May be preferred in economic assessment of how many, and which, teeth among patients seen at dental clinic require costly intervention

CWGEE

- Inference for typical unit of typical cluster
- All clusters contribute to inference equally
- May be preferred in study of patient factors linked to disease status of teeth

Informative cluster size



Marginal analysis of multiple correlated outcomes

Solutions for analysis of multiple correlated binary outcomes with ICS

- Define composite binary outcome and use one model

$$\text{Perio} = \begin{cases} 1 & \text{if ABL} \geq 40\% \text{ and CAL/PPD} \geq 5\text{mm and Mobil} \geq 0.5\text{mm} \\ 0 & \text{otherwise} \end{cases}$$

Marginal analysis of multiple correlated outcomes

Solutions for analysis of multiple correlated binary outcomes with ICS

- Define composite binary outcome and use one model

$$\text{Perio} = \begin{cases} 1 & \text{if ABL} \geq 40\% \text{ and CAL/PPD} \geq 5\text{mm and Mobil} \geq 0.5\text{mm} \\ 0 & \text{otherwise} \end{cases}$$

  - How to define single outcome?
  - Can obscure true effect

Marginal analysis of multiple correlated outcomes

Solutions for analysis of multiple correlated binary outcomes with ICS

- Define composite binary outcome and use one model

$$\text{Perio} = \begin{cases} 1 & \text{if ABL} \geq 40\% \text{ and CAL/PPD} \geq 5\text{mm and Mobil} \geq 0.5\text{mm} \\ 0 & \text{otherwise} \end{cases}$$

  - How to define single outcome?
  - Can obscure true effect

- Use four separate models, one for each outcome

Marginal analysis of multiple correlated outcomes

Solutions for analysis of multiple correlated binary outcomes with ICS

- Define composite binary outcome and use one model

$$\text{Perio} = \begin{cases} 1 & \text{if ABL} \geq 40\% \text{ and CAL/PPD} \geq 5\text{mm and Mobil} \geq 0.5\text{mm} \\ 0 & \text{otherwise} \end{cases}$$

  - How to define single outcome?
  - Can obscure true effect

- Use four separate models, one for each outcome
  - Ignores correlation between outcomes
  - Need to correct for multiple comparison

Marginal analysis of multiple correlated outcomes

Solutions for analysis of multiple correlated binary outcomes with ICS

- Define composite binary outcome and use one model

$$\text{Perio} = \begin{cases} 1 & \text{if ABL} \geq 40\% \text{ and CAL/PPD} \geq 5\text{mm and Mobil} \geq 0.5\text{mm} \\ 0 & \text{otherwise} \end{cases}$$

  - How to define single outcome?
  - Can obscure true effect

- Use four separate models, one for each outcome
  - Ignores correlation between outcomes
  - Need to correct for multiple comparison

- Multivariate approach to jointly analyze all outcomes in one model

Marginal analysis of multiple correlated outcomes with ICS

Three dichotomized periodontal disease outcomes

	Periodontal disease outcomes		
	ABL	CAL	Mobil
Dichotomized score	0: $<$40%	0: $<$5mm	0: $<$0.5mm
	1: $\geq$40%	1: $\geq$5mm	1: $\geq$0.5mm

ABL: Alveolar bone loss

CAL: Clinical attachment loss

Mobil: Mobility

Method

- $i = 1, ..., N$ Subjects
- $j = 1, ..., n_i$ teeth for ith subject at baseline
- $k = 1, 2, 3$ outcome variables
- Y_{ijk} is kth binary outcome for jth tooth of ith subject, $Y_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3})$
- X_i is subject-level predictor
- $\mu_{ijk} = \Pr(Y_{ijk} = 1)$

General model

$$\begin{aligned} \text{logit}(\mu_{ij1}) &= a_1 + X_i\beta, \\ \text{logit}(\mu_{ij2}) &= a_2 + X_i(\beta + \beta_{12}), \\ \text{logit}(\mu_{ij3}) &= a_3 + X_i(\beta + \beta_{13}). \end{aligned} \tag{1}$$

Hypothesis test

$$H_0 : \beta_{12} = \beta_{13} = 0$$

Method

- $i = 1, ..., N$ Subjects
- $j = 1, ..., n_i$ teeth for ith subject at baseline
- $k = 1, 2, 3$ outcome variables
- Y_{ijk} is kth outcome for jth tooth of ith subject, $Y_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3})$
- X_i is subject-level predictor
- $\mu_{ijk} = \Pr(Y_{ijk} = 1)$

Parsimonious model

$$\begin{aligned} \text{logit}(\mu_{ij1}) &= a_1 + X_i\beta, \\ \text{logit}(\mu_{ij2}) &= a_2 + X_i\beta, \\ \text{logit}(\mu_{ij3}) &= a_3 + X_i\beta. \end{aligned} \tag{2}$$

Method

How to model correlation between outcomes?

Generalized sum of squares for error

$$Q_{GEE} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} Z_{ij} R_{ij}(\alpha)^{-1} Z_{ij}^T$$

where $Z_{ij} = (Y_{ij} - \mu_{ij})/\sqrt{\mu_{ij}(1-\mu_{ij})}$ and $R_{ij}(\alpha)$ is correlation matrix between outcomes (Chaganty & Shults, 1999)

Method

How to model correlation between outcomes?

Generalized sum of squares for error

$$Q_{GEE} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} Z_{ij} R_{ij}(\alpha)^{-1} Z_{ij}^T$$

where $Z_{ij} = (Y_{ij} - \mu_{ij})/\sqrt{\mu_{ij}(1-\mu_{ij})}$ and $R_{ij}(\alpha)$ is correlation matrix between outcomes (Chaganty & Shults, 1999)

Cluster weighted generalized sum of squares for error

$$Q_{CWGEE} = \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{ij} R_{ij}(\alpha)^{-1} Z_{ij}^T$$

Method

How to model correlation between outcomes?

Estimation of β

$\frac{\partial Q(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = 0 \Rightarrow$

$$U_{CWGEE}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}}' V_{ij}(\alpha)^{-1}(Y_{ij} - \mu_{ij}) = 0 \tag{3}$$

Estimation of α

$\frac{\partial Q(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = 0 \Rightarrow$

$$\sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{ij} \frac{\partial R_{ij}(\alpha)^{-1}}{\partial \alpha} Z_{ij}^{T} = 0 \tag{4}$$

Iterate between Equations (3) and (4) until convergence.

Method

Working correlation structures for $R_{ij}(\alpha)$

1. Unstructured:

$$R_{ij}(\alpha) = \begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} \\ \alpha_{12} & 1 & \alpha_{23} \\ \alpha_{13} & \alpha_{23} & 1 \end{pmatrix}$$

2. Exchangeable:

$$R_{ij}(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix}$$

3. Independence:

$$R_{ij}(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Analysis

VA Dental Longitudinal Study (Baseline)

- $N = 760$ subjects
- $1 - 28$ teeth per subject
- $K = 3$ binary outcomes: CAL≥5mm, ABL≥40%, Mobil≥0.5mm
- Subject-level predictors: $\boldsymbol{X}_i = (X'_{\text{Age}}, X'_{\text{Smoking}}, X'_{\text{Education}}, X'_{\text{MetS}})$

General model

$$\text{logit}(\mu_{ij\text{CAL}}) = a^{\text{CAL}} + \boldsymbol{X}_i\boldsymbol{\beta}$$
$$\text{logit}(\mu_{ij\text{ABL}}) = a^{\text{ABL}} + \boldsymbol{X}_i(\boldsymbol{\beta} + \boldsymbol{\beta}^{\text{ABL}})$$
$$\text{logit}(\mu_{ij\text{Mobil}}) = a^{\text{Mobil}} + \boldsymbol{X}_i(\boldsymbol{\beta} + \boldsymbol{\beta}^{\text{Mobil}})$$

Table 2: Results from general model assuming unstructured corr structure. P-values are for $H_0 : \beta^{ABL} = \beta^{Mobil} = 0$

	GEE		CWGEE	
	Estimate (SE)	P-value	Estimate (SE)	P-value
Int (CAL)	-4.500 (0.879)		-4.810 (0.888)	
Int (ABL)	-4.042 (0.843)		-3.750 (0.887)	
Int (Mobil)	-4.821 (0.888)		-4.174 (0.958)	
Age	0.041 (0.105)		0.051 (0.106)	
Age (ABL)	-0.017 (0.100)	0.231	-0.024 (0.096)	0.018
Age (Mobil)	-0.010 (0.102)		-0.023 (0.104)	
Smoking	0.710 (0.445)		0.657 (0.470)	
Smoking (ABL)	0.253 (0.421)	0.360	0.132 (0.421)	0.726
Smoking (Mobil)	0.078 (0.426)		-0.018 (0.455)	
Edu	-0.401 (0.334)		-0.424 (0.350)	
Edu (ABL)	0.002 (0.316)	0.683	-0.041 (0.320)	0.454
Edu (Mobil)	-0.083 (0.323)		-0.157 (0.353)	
MetS	0.403 (0.420)		0.336 (0.430)	
MetS (ABL)	-0.197 (0.401)	0.288	-0.267 (0.406)	0.197
MetS (Mobil)	0.096 (0.422)		0.067 (0.434)	

Analysis

Table 3: Results from parsimonious models assuming unstructured corr structure

	GEE		CWGEE	
	Estimate (SE)	P-value	Estimate (SE)	P-value
Int (CAL)	-4.086 (0.671)	$<$0.001	-4.774 (0.887)	$<$0.001
Int (ABL)	-4.659 (0.676)	$<$0.001	-3.751 (0.873)	$<$0.001
Int (Mobil)	-5.133 (0.675)	$<$0.001	-4.295 (0.926)	$<$0.001
Age	0.035 (0.010)	$<$0.001	0.052 (0.106)	$<$0.001
Age (ABL)			-0.025 (0.096)	0.007
Age (Mobil)			-0.024 (0.106)	0.028
Smoking	0.794 (0.171)	$<$0.001	0.695 (0.441)	$<$0.001
Edu	-0.413 (0.010)	$<$0.001	-0.458 (0.329)	$<$0.001
MetS	0.360 (0.154)	0.019	0.277 (0.404)	0.089

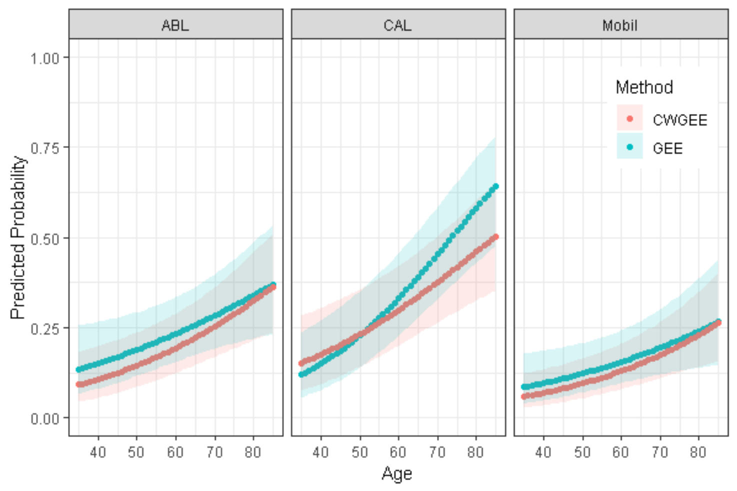
Analysis

Table 4: Estimates of the working correlation matrices (unstructured and exchangeable): GEE estimates are shown in the upper half of the matrices and CWGEE estimates are shown in the lower half of the matrices.

	Unstructured				Exchangeable		
	CAL	ABL	Mobil		CAL	ABL	Mobil
CAL	-	0.40	0.33	CAL	-	0.31	0.31
ABL	0.40	-	0.29	ABL	0.34	-	0.31
Mobil	0.32	0.29	-	Mobil	0.34	0.34	-

Analysis

Figure 2: Predicted probability of each outcome by age of a smoker with MetS and no college education



Simulation study

Simulation study to assess performance between multivariate CWGEE and GEE

- N=750 subjects, K=3 outcomes
- Induced ICS
- Varied correlation between teeth and correlation between outcomes

Result

- GEE
    - Performs well when applied to data with no ICS
    - Type I error inflated in scenarios with higher levels of correlation
    - Relative bias increase with increasing levels of correlation
- CWGEE
    - Type I error close to 5% across varying levels of correlation
    - Low relative biases and excellent coverage probabilities across varying levels of correlation
    - Performs well when applied to data with no ICS

Conclusions

Research question
What is the relationship between periodontal disease and MetS?

Answer
MetS is not an important predictor

ICS in HIV/STD research, Williamson et al

ELSEVIER

Weighting Condom Use Data to Account for Nonignorable Cluster Size

JOHN M. WILLIAMSON, MSc, ScD, HAE-YOUNG KIM, MSc, AND LEE WARNER, MPH, PhD

PURPOSE: We examined the impact of weighting the generalized estimating equation (GEE) by the inverse of the number of sex acts on the magnitude of association for factors predictive of recent condom use.
METHODS: Data were analyzed from a cross-sectional survey on condom use reported during vaginal intercourse during the past year among male students attending two Georgia universities. The usual GEE model was fit to the data predicting the binary act-specific response indicating whether a condom was used. A second cluster-weighted GEE model (i.e., weighting the GEE score equation by the inverse of the number of sex acts) was also fit to predict condom use.
RESULTS: Study participants who engaged in a greater frequency of sex acts were less likely to report condom use, resulting in nonignorable cluster-size data. The GEE analysis weighted by sex act (usual GEE) and the GEE analysis weighted by study subject (cluster-weighted GEE) produced different estimates of the association between the covariates and condom use in last year. For example, the cluster-weighted GEE analysis resulted in a marginally significant relationship between age and condom use (odds ratio of 0.49 with 95% confidence interval (0.23–1.03) for older versus younger participants) versus a nonsignificant relationship with the usual GEE model (odds ratio of 0.67 with a 95% confidence interval of 0.28–1.60).
CONCLUSIONS: The two ways of weighting the GEE score equation, by the sex act or by the respondent, may produce different results and a different interpretation of the parameters in the presence of nonignorable cluster size.
Ann Epidemiol 2007;17:603–607.

KEY WORDS: Condom use, Generalized Estimating Equations, HIV Infections, Informative Cluster Size, Sex Behavior, Sexually Transmitted Diseases.

ICS in HIV/STD research, Williamson et al

- Male condom use has been associated with reduced risk of HIV and many other STDs
- Identify demographic and behavioral characteristics of persons who report using condoms for STD prevention
- A cross-sectional study on condom use was conducted on a sample of male students attending two Georgia universities during 1993–1994
- Eligibility
  - Age 18-29 years
  - Full-time student
  - Lifetime use of ≥ 5 condoms during vaginal intercourse
- Confidential standardized interview to ascertain information about their use of condoms during vaginal intercourse, including condom use during the past year

ICS in HIV/STD research, Williamson et al

- $i = 1, ..., 85$ students
- $j = 1, ..., n_i$ sex acts
- $Y_{ij} = 1$ if condom used

TABLE 1. Percentage of condom use in last year by number of sex acts

Number of sex acts	Number of respondents	Percent condom use (no.)
0	5	
1–15	18	77.9 (109/140)
16–50	23	68.9 (519/753)
51–85	19	57.0 (743/1304)
86–280	20	25.2 (750/2980)
Total	85	41.0 (2121/5177)

ICS in HIV/STD research, Williamson et al

TABLE 2. Results of GEE analyses of condom use data from a cross-sectional survey of males attending two Georgia universities

		Unweighted GEE[a,b]		Weighted GEE[a,c]	
Predictor	No. of persons	Adjusted OR[a]	95% CI[a]	Adjusted OR[a]	95% CI[a]
Intercept					
Age	80				
≥23 years	36	0.67	[0.28–1.60]	0.49	[0.23–1.03]
18-22 years	44	1.0		1.0	
Race	80				
Black	26	2.69	[1.04–6.97]	1.90	[0.83–4.36]
Other	54	1.0		1.0	
Number of sex partners	78				
≥10	39	0.96	[0.32–2.91]	0.77	[0.32–1.90]
<10	39	1.0		1.0	
Condom use at first sex	77				
Yes	41	1.31	[0.54–3.19]	1.36	[0.63–2.93]
No	36	1.0		1.0	

[a]GEE = generalized estimating equations, OR = odds ratio, CI = confidence interval.
[b]Usual unweighted GEE analysis with independence working correlation matrix based on 75 subjects, after deleting five observations with missing values.
[c]Cluster-weighted GEE analysis with independence working correlation matrix based on 75 subjects, after deleting five observations with missing values.

ICS in HIV/STD research, Williamson et al

- Cluster size (number of sex acts) was informative on the outcome (condom use)
    - Cluster size varied
    - Strong association between cluster size and outcome
- Some differences observed in results from unweighted GEE vs. CWGEE
- Differences may be due to relationships between
    - cluster size and outcome
    - covariate and outcome
    - covariate and cluster size

More recent studies that address ICS

Outcome (Y)	Unit w/n cluster	Cluster	Study
Neonatal complication	Infant	Birth	Yelland, 2015
Fetal malformation	Live fetus	Litter	Zhang, 2015
Alcohol consumption	Student	School	Innocenti, 2018
Surgical outcome	Patient	Hospital	Panageas, 2007

Other methods for ICS

Marginal inference

- Longitudinal data (Wang et al, 2011 & Bible et al, 2016 & Mitani et al, 2019)
- With informative empty clusters (McGee et al, 2019)

Cluster-specific inference

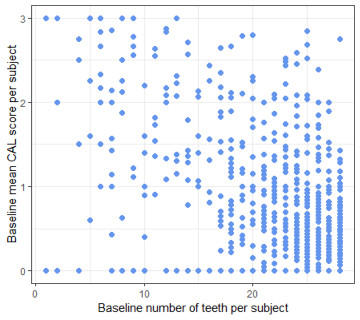
- Joint modelling of cluster size and outcomes (Dunson et al, 2003 & Gueorguieva, 2005)
- GLMM (Neuhaus and McCulloch, 2011)

Time-to-event analysis

- Williamson et al, 2008 & Zhang et al, 2013

How to check for ICS?

- Plot outcome and cluster size
  - Compute correlation
- Formal tests
  - Wald test (Benhin et al, 2005)
  - Bootstrap (Nevalainen et al, 2017)
- Sensitivity analysis

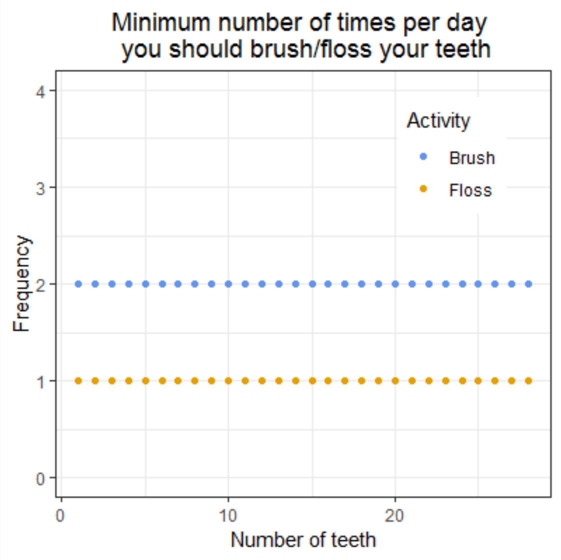


Software for CWGEE

- For cross-sectional data with single outcome
  - Use weights argument in R package geepack
  - Use WEIGHTS statement in SAS PROC GEE or PROC GENMOD
- R package CWGEE (https://github.com/AyaMitani/CWGEE)
  - Use mvoCWGEE function for cross-sectional data with multiple outcomes
  - Use ordCWGEE function for longitudinal data with ordinal outcomes (Mitani et al, 2019)

Final Message

Brush your teeth ≥2 and floss ≥1 times every day for all $n_i = 1, ..., 28$!!



References

Kaye, E. K. *et al.* Metabolic Syndrome and Periodontal Disease Progression in Men. *Journal of Dental Research* **95,** 822–828 (2016).

Hoffman, E. B. *et al.* Within-cluster resampling. *Biometrika* **88,** 1121–1134 (2001).

Williamson, J. M. *et al.* Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59,** 36–42 (2003).

Chaganty, N. R. *et al.* On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. *Journal of Statistical Planning and Inference* **76,** 145–161 (1999).

Mitani, A. A. *et al.* Marginal analysis of multiple outcomes with informative cluster size. Under revision.

Mitani, A. A. *et al.* Marginal analysis of ordinal clustered longitudinal data with informative cluster size. *Biometrics* **75,** 938–949 (2019).

Williamson, J. M. *et al.* Weighting Condom Use Data to Account for Nonignorable Cluster Size. *Annals of Epidemiology* **17,** 603–607 (2007).

Benhin, E. *et al.* Mean Estimating Equation Approach to Analysing Cluster-Correlated Data with Nonignorable Cluster Sizes. *Biometrika* **92,** 435–450 (2005).

Nevalainen, J. *et al.* Tests for informative cluster size using a novel balanced bootstrap scheme. *Statistics in Medicine* **36,** 2630–2640 (Mar. 2017).

Yelland, L. N. *et al.* Analysis of Randomised Trials Including Multiple Births When Birth Size Is Informative. *Paediatric and Perinatal Epidemiology* **29,** 567–575 (2015).

Innocenti, F. *et al.* Relative efficiencies of two-stage sampling schemes for mean estimation in multilevel populations when cluster size is informative. *Statistics in Medicine* **38,** 1817–1834 (Dec. 2018).

Panageas, K. S. *et al.* Properties of analysis methods that account for clustering in volume–outcome studies when the primary predictor is cluster size. *Statistics in Medicine* **26,** 2017–2035 (2007).

Iosif, A.-M. *et al.* A model for repeated clustered data with informative cluster sizes. *Statistics in Medicine* **33,** 738–759 (Sept. 2013).

Wang, M. *et al.* Inference for marginal linear models for clustered longitudinal data with potentially informative cluster sizes. *Statistical Methods in Medical Research* **20,** 347–367 (2011).

Bible, J. *et al.* Cluster adjusted regression for displaced subject data (CARDS): Marginal inference under potentially informative temporal cluster size profiles. *Biometrics* **72,** 441–451 (2016).

McGee, G. *et al.* Informatively empty clusters with application to multigenerational studies. *Biostatistics* (Apr. 2019).

Dunson, D. B. *et al.* A Bayesian Approach for Joint Modeling of Cluster Size and Subunit-Specific Outcomes. *Biometrics* **59,** 521–530 (2003).

Gueorguieval, R. V. *et al.* Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine* **25,** 1307–1322 (2006).

Neuhaus, J. M. *et al.* Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. *Biometrika* **98,** 147–162 (2011).

Williamson, J. M. *et al.* Modeling survival data with informative cluster size. *Statistics in Medicine* **27,** 543–555 (2008).

Zhang, X. Y. *et al.* Semiparametric Regression Analysis of Clustered Interval-Censored Failure Time Data with Informative Cluster Size. *International Journal of Biostatistics* **9,** 205–214 (2013).

Thank you!

Questions?

Co-authors
Kerrie Nelson, Biostatistics, BU School of Public Health
Elizabeth Kaye, BU Henry M. Goldman School of Dental Medicine

Funding
F31DE027589 (PI: Mitani)
R01CA226805 (PI: Nelson)

Simulation study

Design of simulation study

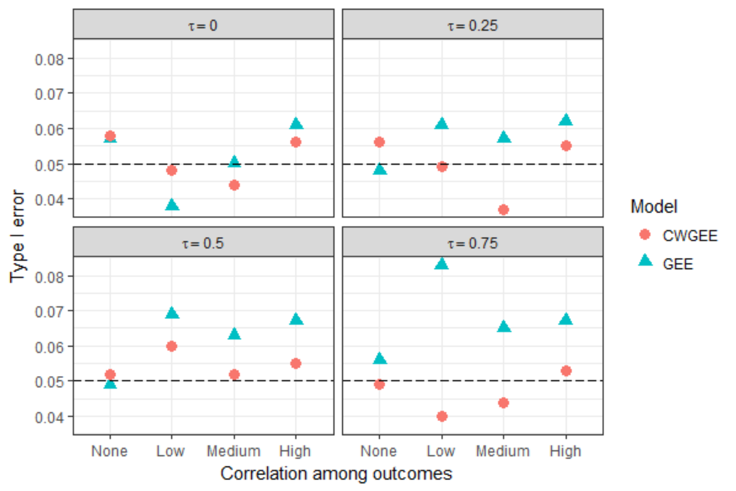
- N=750 subjects, K=3 outcomes
- $n_i \sim \text{Bin}(size = 28, prob = \lambda_i)$
- $\Pr(Y_{ijk} = 1) \sim f(\lambda_i, a_k, \boldsymbol{X}_i)$
- True model: $\text{logit}\{\Pr(Y_{ijk} = 1)\} = a_k + \boldsymbol{X}_i\beta$
- Compare performance of GEE and CWGEE while varying
  1. Correlation between teeth, $\tau : (0, 0.25, 0.5, 0.75)$
  2. Correlations between outcomes $(\alpha_{12}, \alpha_{13}, \alpha_{23})$:

None	Low	Medium	High
$(0, 0, 0)$	$(0.4, 0.35, 0.3)$	$(0.6, 0.55, 0.5)$	$(0.8, 0.75, 0.7)$

- Number of simulations: 1,000

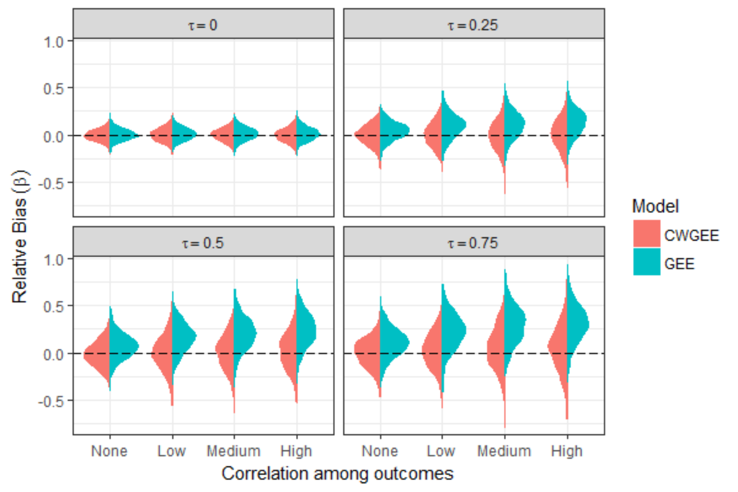
Simulation results

Figure 3: Simulation results of type I error rate ($H_0 : \beta_1 = \beta_2 = 0$) when fitting general model



Simulation results

Figure 4: Simulation results of relative bias ($\hat{\beta}$)



- interpretations here refer to a random member of a random cluster
- Intuitively, by balancing the contributions of all clusters, this approach downweights the largest clusters and upweights the smallest.
- There are two types of sampling inherent when analyzing clustered data with marginal modeling. The first is unitbased sampling that is implicit in the usual marginal models such as GEE, and the second is cluster-based sampling where one selects a random observation from a randomly selected cluster. For the former, larger clusters are weighted more than smaller ones. For the latter, all clusters are given equal weight regardless of size and accordingly the marginal parameter will have a cluster-based interpretation. Asymptotically the two marginal analyses will reach the same conclusion if cluster

Notes ii

size is unrelated to the outcome of interest. However, the two marginal analyses are different for informative cluster size data.

- When the total number of members in the cluster is informative, then inference may be for a typical member of a typical cluster or the population of all cluster members. Applying the GEE with independence working correlation provides inference for the population of all members, and with additional weighting by the inverse cluster size gives inference for the population of typical members.

- there are two marginal analyses of interest: one for the population of all cluster members (population M), where larger clusters contribute more to inference than smaller ones; and one for a typical member of a typical cluster, where all clusters contribute equally. We view the latter as inference for the population of typical cluster members (population C), which is a subpopulation of population M, formed by selecting one member at random from each cluster.
- In an economic assessment of how many, and which, teeth among patients seen at a dental clinic require a costly intervention, the population of all members (teeth) might be preferred, as clustering by patient may not be of direct relevance. Conversely, in a study of patient factors linked to the disease status of teeth, the population of typical cluster members (typical teeth for patients) might be of more interest.

Notes iv

- Inference for population M can be obtained by applying the standard GEE with independence working correlation. For population C two inference methods were initially proposed: the computationally-intensive within-cluster resampling method (WCR - Hoffman, Sen, and Weinberg (2001)) and the simpler inversely-weighted-by-cluster-size GEE with independence working correlation
- Unless cluster size is a predictor of primary scientific interest, such as in volume-outcome studies (see, for example, French et al. (2012)), there are two reasons why we do not wish to formulate a regression model involving N. First, N might lie in the causal pathway between Y and X. In the toxicology application, adjusting for the cluster size may cause misleading inferences for the effect of the exposure if unobserved factors that contribute to the foetal loss induced by the toxin are also associated with the foetal weight. Second, if the effect

of X on Y is different in clusters of different sizes then the effect of X conditional on N is a quantity which might not be scientifically useful.

- CWGEE weights each cluster by choosing the working correlation matrix Ri as the identity matrix (i.e., assuming independence) and weighting the GEE equation by 1/ni . This approach weights each cluster equally because the independence working correlation matrix represents ni independent observations and is canceled out by the factor 1/ni . Choosing a different working correlation matrix other than independence will require a different weight ([10 Ri 1 1]1 where 1ni is a ni 1 vector of 1s) to weight each cluster equally and achieve unbiased parameter estimation. In contrast, with the usual GEE model, specifying the working correlation matrix closer to the true correlation matrix will result in increased efficiency (asymptotically). However, with CWGEE there is little difference in efficiency based on

Notes vi

the choice of working correlation matrix as the cluster weight ([10 Ri 1 1]1) is chosen to weight each cluster equally and cancels out the choice of working correlation matrix Ri. (Williamson et al 2007)