Comparison of GEE-based methods for efficient estimation in cluster-randomized trial with missing data.

Mélanie Prague[(1)], Rui Wang[(2)], Eric Tchetgen Tchetgen[(1)] and Victor De Gruttola[(1)]

[(1)] Harvard School of Public Health, Boston MA, USA
[(2)] Brigham and Women's Hospital, Boston MA, USA

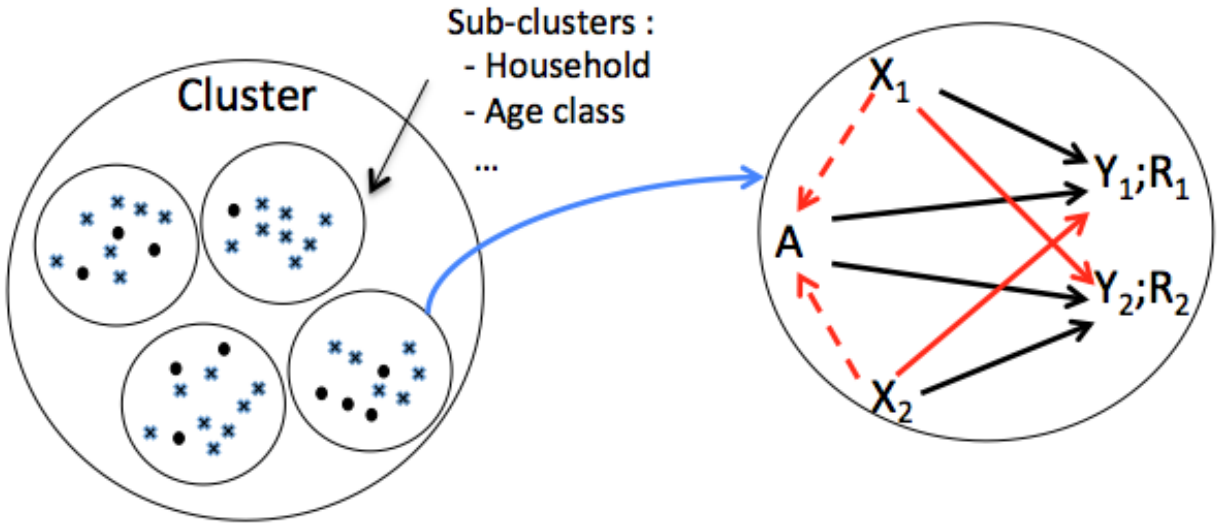
Boston - USA, JSM, August 6^{th} 2014

Background

→ Cluster randomized trials :
- Intervention study
  - Continuous outcome
  - Binary outcome
- Complex correlation structures (sub-clusters)
  - Individuals in household in communities
  - Longitudinal clustered data
  - Nested sub clusters according to observed or latent traits

→ Missing data :
- Multiple sources (**MAR** or MNAR)
  - Failure to locate village residents,
  - Denial of consent,
  - Dropout.
- Complex correlation structures
  - Sub-clusters

Outline

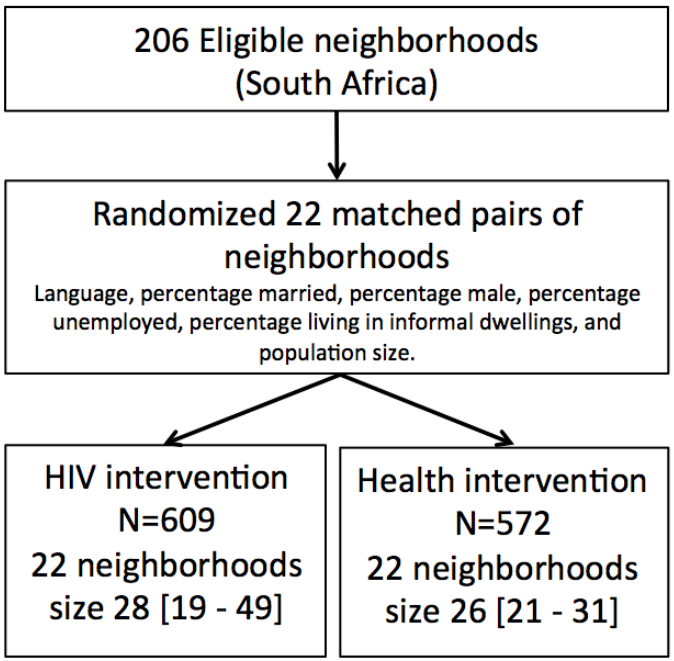


Outcomes and missingness may be affected by personal characteristics of others subjects in the same cluster.

1. Definition of the problem & Augmented method for correction,
2. Simulations,
3. Implication for analysis.

Motivating example

The SAM study [Jemmott2014]



Primary outcome at 12 months :

- Score for HIV knowledge (continuous)
- Consistent use of condom (binary)

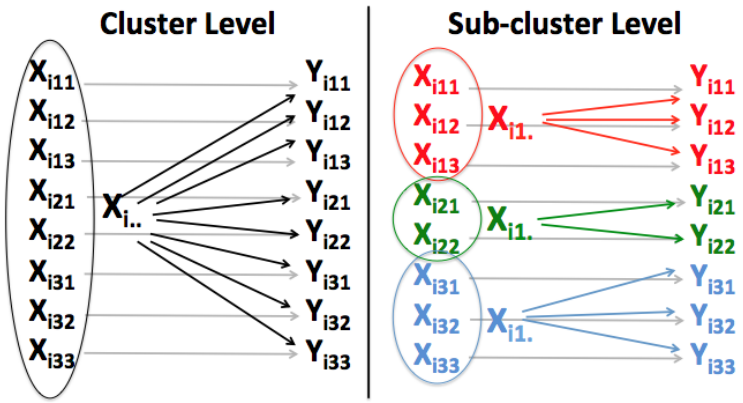
Missingness of the HIV outcome :

- 7%

[Jemmott2014] Jemmott et al. (2014) Cluster-Randomized Controlled Trial of an HIV/Sexually Transmitted Infection Risk-Reduction Intervention for South African Men *Research and practice 104(3) 467-474*

The SAM study [Jemmott2014]

p-value for association	with $A^{(a)}$	with $Y^{(b)}$	with $R^{(a)}$
At the cluster level :			
Mean education in the community	0.30	< 0.001	0.632
Mean alcool level in the community	0.68	0.24	0.024
At the sub-cluster level :			
Wealth class	0.025	0.010	0.900
Mean score for condom use in each wealth class	0.019	<0.001	0.682



(a) anova for continuous variables ; chi-squared test otherwise

(b) Pearson correlation test for continuous variables ; chi-squared test otherwise

Statistical framework

Notations

→ Notations :

Y_{ijk}	The response variable for patient k in sub-cluster j in cluster i and $Y_i = (Y_{i11}, \ldots, Y_{i1n_{i1}}, \ldots, Y_{iJ1}, \ldots, Y_{iJn_{iJ}})^T$ the observations for cluster i
A_i	the treatment in cluster i
X_{ijk}	a covariate for patient k in sub-cluster j in cluster i and $X_i = (X_{i11}, \ldots, X_{i1n_{i1}}, \ldots, X_{iJ1}, \ldots, X_{iJn_{iJ}})^T$
$\bar{X}_{ij.}$	a interference covariate from a mapping function $h((x_{ijk})\|k \in$ sub-cluster $j)$
R_{ijk}	the indicator of missingness (=1 if observed)

→ The regression model is :

$$\mu_{ijk} = \mathbb{E}[Y_{ijk}|X_i] = g(A_i; X_i; \beta)$$

Statistical framework

Interested β, the average effect on population : GEE [Hu98].

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

with C a Working Correlation Structure,

$$V_i^{-1} = U_i^{1/2} \mathbf{C}(\alpha) U_i^{1/2}$$

Missing data : Inverse probability weighting IPW [Robins94].

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

$$W_{ijk} = \frac{R_{ijk}}{\mathbb{P}(R_{ijk} = 1)}$$

[Hu98] Hu et al. (1998) Comparison of Population-Averaged and Subject-Specific Approaches... *AJE 147(7) 694-703*

[Robins94] Robins et al. (1994) Estimation of regression coefficients when some regressors are not always observed *JASA 89(427) 846-866*

Asymptotic properties

Classical GEE properties [LiangZeger86]

In MCAR patterns, consistent and asymptotically normal estimates for β does not depend on the working correlation structure (C) specification which only plays on efficiency.

Classical GEE-IPW properties [Robins94]

In MAR patterns, this result holds.

[LiangZeger86] Liang et Zeger (1986) Longitudinal data analysis using generalized linear models *Biometrika (73) 13-22*

Important underlying assumption

Underlying assumption for outcome [Pepe94]

For the outcome model :

$$\mathbb{E}(Y_{ijk}|X_i) = \mathbb{E}(Y_{ijk}|X_{ijk})$$

OR

$$C^{(O)} = \text{Independence}$$

For the IPW model (to compute W) :

$$\mathbb{E}(R_{ijk}|X_i) = \mathbb{E}(R_{ijk}|X_{ijk})$$

OR

$$C^{(M)} = \text{Independence}$$

Underlying assumption for missingness [TchetgenTchetgen12]

For the outcome model :

$$\mathbb{E}(R_{ijk}|X_i) = \mathbb{E}(R_{ijk}|X_{ijk})$$

OR

$$C^{(O)} = \text{Independence (in the outcome model)}$$

[Pepe94] Pepe et Anderson (1994) A cautionary note on Inference for Marginal regression models with longitudinal data and general correlated response data *Comm. Statistics-Simulations (23) 939-951*

[TchetgenTchetgen12] Tchetgen Tchetgen et al. (2012) A Cautionary Note on Specification of the Correlation Structure in Inverse-Probability-Weighted Estimation for Repeated Measures *Harvard Tech. Report.*

Goal for our investigation

1. **Extend these results to multilevel correlated data**
   - Difference between longitudinal and clustered :
     - No natural ordering,
     - patterns of correlation complex in different ways.
   - Combine [Pepe94] and [TchetgenTchetgen12] result.

2. **Improve efficiency by accounting for other subjects' covariate**
   - Use an Augmented approach extended for missing data [Stephens12]

$$\sum_{i=1}^{n} \frac{\partial \mu_i(\beta)}{\partial \beta} V_i^{-1} W_i (Y_i - \mu_i) - \sum_{l} (I(A_i = l) - \mathbb{P}(A_i = l)) \gamma_l^T = 0$$

   - Set $C^{(O)} = C^{(M)} =$ Independence
   - γ_l^T depends on other patient covariates ($\bar{x}_{ij.}$) whose selection may be guided by the correlation structure.

[Stephens12] Stephens et al. (2012) Augmented generalized estimating equations for improving efficiency and validity of estimation in cluster randomized trials by leveraging cluster-level and individual-level covariates *Stat. Med. (31) 915-930*

Preliminary results : Simulations

Simulations of RCT [$\mathbb{P}(A_i = 1) = 1/2$; $n_i = [20; 50]$; $n = 100$]

$$
\begin{aligned}
(\mathsf{S};\mathsf{X}) &\sim MVN\left((30, b_i), \begin{pmatrix} 10 & \rho_C/\sqrt{(10*5)} \\ \rho_C/\sqrt{(10*5)} & 5 \end{pmatrix}\right) \\
\mathsf{S} &: \text{sub-clustering variable } \{< Q_1; [Q_1; Q_3]; > Q_3\} \\
X_{ijk} &= \mathsf{X}_{ijk} + \mathcal{N}_i(0, 1.0) \\
\bar{X}_{ij.} &= \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} X_{ijk}
\end{aligned}
$$

Outcome model :

$$
\begin{aligned}
Y_{ijk} &= \beta_0^{*(o)} + \beta_1^{*(o)} A_i + \beta_2^{*(o)} X_{ijk} + \beta_3^{*(o)} \bar{X}_{ij.} + b_i^o \\
b^o &\sim \mathcal{N}(0, \rho^{(o)})
\end{aligned}
$$

Missingness model :

$$
\begin{aligned}
\text{logit}(\mathbb{P}(R_{ijk}=1)) &= \beta_0^{*(m)} + \beta_1^{*(m)} A_i + \beta_2^{*(m)} X_{ijk} + \beta_3^{*(m)} \bar{X}_{ij.} + b_i^m \\
b^{(m)} &\sim \mathcal{B}_l(\rho^m)
\end{aligned}
$$

Models for analysis

Outcome model :

$$
\begin{aligned}
Y_{ijk} &= \beta_0^{(o)} + \beta_1^{(o)} A_i && (o.trt) \\
Y_{ijk} &= \beta_0^{(o)} + \beta_1^{(o)} A_i + \beta_2^{(o)} X_{ijk} && (o.x) \\
Y_{ijk} &= \beta_0^{(o)} + \beta_1^{(o)} A_i + \beta_2^{(o)} X_{ijk} + \beta_3^{(o)} \bar{X}_{ij.} && (o.true)
\end{aligned}
$$

Missingness model :

$$
\begin{aligned}
\text{logit}(\mathbb{P}(R_{ijk}{=}1)) &= \beta_0^{(m)} + \beta_1^{(m)} A_i && (m.trt) \\
\text{logit}(\mathbb{P}(R_{ijk}{=}1)) &= \beta_0^{(m)} + \beta_1^{(m)} A_i + \beta_2^{(m)} X_{ijk} && (m.x) \\
\text{logit}(\mathbb{P}(R_{ijk}{=}1)) &= \beta_0^{(m)} + \beta_1^{(m)} A_i + \beta_2^{(m)} X_{ijk} + \beta_3^{(m)} \bar{X}_{ij.} && (m.true)
\end{aligned}
$$

Analysis : R software package geepack [Halekoh06]

[Halekoh06] Halekoh et al. (2006) The R Package geepack for Generalized Estimating Equations *J. Stat. Soft. 15(2) 1-11*

Biais for treatment parameter (no missingness)

Interest for A_i	$\frac{1}{n}\sum_k \left(\beta_{1k}^{(o)} - \beta_1^{*(o)}\right)$		Mean SE		Coverage	
	Ind.	Exch.	Ind.	Exch	Ind.	Exch
o.trt	0.005	0.004	1.25	1.23	94.5	93.8
o.x	0.013	0.013	0.79	0.8	93.8	93.7
o.true	0.021	0.023	0.61	0.59	94.3	94.5

- No bias for treatment effect because of randomization ($A \perp X$)

n=100 communities of size [25 ;50] - n=1000 replications for simulations
ρ_C=0.6 ; ρ^o=3.0 ; ρ^m=0.1 ; $\beta^{*(o)}$=(0.3,1.0,1.0,3.0) ; $\beta^{*(m)}$=(-3.5,0.2,-0.25,-1.0)

Biais for treatment parameter (no missingness)

Interest	$\frac{1}{n}\sum_k \left(\beta_{1k}^{(o)} - \beta_1^{*(o)}\right)$		Mean SE		Coverage	
for A_i	Ind.	Exch.	Ind.	Exch	Ind.	Exch
o.trt	0.005	0.004	1.25	1.23	94.5	93.8
o.x	0.013	0.013	0.79	0.8	93.8	93.7
o.true	0.021	0.023	0.61	0.59	94.3	94.5

- No bias for treatment effect because of randomization ($A \perp X$)

Interest	$\frac{1}{n}\sum_k \left(\beta_{2k}^{(o)} - \beta_2^{*(o)}\right)$		Mean SE		Coverage	
for X_{ijk}	Ind.	Exch.	Ind.	Exch	Ind.	Exch
o.x	0.023	0.132	0.78	0.79	94.5	92.7
o.true	0.018	0.022	0.68	0.64	94.3	94.1

- Bias for covariate effect, use $C^{(O)}$ =Independence.

n=100 communities of size [25 ;50] - n=1000 replications for simulations
ρ_C=0.6 ; ρ^o=3.0 ; ρ^m=0.1 ; $\beta^{*(o)}$=(0.3,1.0,1.0,3.0) ; $\beta^{*(m)}$=(-3.5,0.2,-0.25,-1.0)

Analysis of the Complete Case

No missingness	Bias		Mean SE		Coverage	
	Ind.	Exch.	Ind.	Exch	Ind.	Exch
o.trt	0.005	0.004	1.25	1.23	94.5	93.8
o.x	0.013	0.013	0.79	0.8	93.8	93.7
o.true	0.021	0.023	0.61	0.59	94.3	94.5

Complete case (25%)	Bias		Mean SE		Coverage	
	Ind.	Exch.	Ind.	Exch	Ind.	Exch
o.trt	0.247	0.254	1.09	1.08	93.1	93.4
o.x	0.125	0.136	0.81	0.81	93.6	93.6
o.true	0.039	0.040	0.63	0.59	94.9	94.5

- Missing data increase bias because $R \not\perp A$

n=100 communities of size [25 ;50] - n=1000 replications for simulations
ρ_C=0.6 ; ρ^o=3.0 ; ρ^m=0.1 ; $\beta^{*(o)}$=(0.3,1.0,1.0,3.0) ; $\beta^{*(m)}$=(-3.5,0.2,-0.25,-1.0)

Missing data analysis : IPW

Complete case IPW (25%)	Bias		Mean SE		Coverage	
	Ind.	Exch.	Ind.	Exch	Ind.	Exch
o.trt.m.trt	0.247	0.254	1.09	1.08	93.1	93.4
o.trt.m.x	0.168	1.402	1.22	2.75	93.2	81.1
o.trt.m.true	0.031	0.958	2.18	2.17	94.3	77.2
o.x.m.trt	0.125	0.136	0.81	0.81	93.6	93.6
o.x.m.x	0.157	0.245	0.81	1.03	93.9	92.4
o.x.m.true	0.036	0.139	1.02	2.41	93.3	83
o.true.m.trt	0.039	0.040	0.63	0.59	94.9	94.5
o.true.m.x	0.022	0.024	0.63	0.9	95.1	90
o.true.m.true	0.027	0.077	0.64	1.13	94	83.4

- No correction with m.trt compared to Complete Case analysis
- Bias and low coverage, use $C^{(O)}$ =Independence.

n=100 communities of size [25 ;50] - n=1000 replications for simulations
ρ_C=0.6 ; ρ^o=3.0 ; ρ^m=0.1 ; $\beta^{*(o)}$=(0.3,1.0,1.0,3.0) ; $\beta^{*(m)}$=(-3.5,0.2,-0.25,-1.0)

Application of the AIPW

Most common model :

- o.trt for the outcome model,
- m.true for the IPW model.

Complete case	Bias		Mean SE		Coverage	
IPW (25%)	Ind.	Exch.	Ind.	Exch.	Ind.	Exch.
o.trt.m.trt	0.247	0.254	1.09	1.08	93.1	93.4
o.trt.m.x	0.168	1.402	1.22	2.75	93.2	81.1
o.trt.m.true	0.031	0.958	2.18	2.17	94.3	77.2
o.x.m.trt	0.125	0.136	0.81	0.81	93.6	93.6
o.x.m.x	0.157	0.245	0.81	1.03	93.9	92.4
o.x.m.true	0.036	0.139	1.02	2.41	93.3	83
o.true.m.trt	0.039	0.040	0.63	0.59	94.9	94.5
o.true.m.x	0.022	0.024	0.63	0.9	95.1	90
o.true.m.true	0.027	0.077	0.64	1.13	94	83.4

Application of the AIPW

We focus on the case o.trt and m.true and $C = I$:

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \frac{\partial \mu_i(\beta)}{\partial \beta} V_i^{-1} W_i (Y_i - \mu_i) \\ &\quad - \sum_{k=\{0,1\}} (A_i - \pi) \left\{ D_i(k)^T V_i(k)^{-1} W_i(k) \left[E(Y_i | A_i = k, X_i) - \mu_i(k) \right] \right\} \end{aligned}$$

with,

$$E(Y_i | A_i = k, X_i) = \beta_0^{(o)} + \beta_2^{(o)} X_{ijk} + \beta_3^{(o)} \bar{X}_{ij}.$$

	Bias		Mean SE		Coverage	
	Ind.	Exch.	Ind.	Exch	Ind.	Exch
o.trt (no missing)	0.001	0.001	1.25	1.25	93.5	93.5
o.trt (complete case)	0.246	0.253	1.24	1.25	92.3	78.9
o.trt.m.true (ipw)	0.069	0.902	1.06	1.90	94.3	77.2
Augmented IPW	0.001	-	0.67	-	94.5	-

Conclusion and prospects

Conclusion and prospects

Interference in clustered randomized trials with missing data have an implication for data analysis

→ Need of a prior analysis to assess if outcome and missingness depends on other patients covariates.

→ If so, need to handle it :

- [Pan02] used a resampling bootstrap-based approach to select the best working correlation structure but it is computationally intensive,
- Use an independence working correlation matrix and gain efficiency by augmentation.

[Pan02] Pan et Connett (2002) Selecting the working correlation structure in generalized estimating equations with application to the lung health study *Stat. Sin. (12) 475-490*

Conclusion and prospects

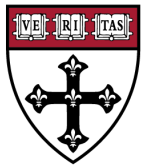
- **Step 1 : Interference covariate definition**
  - → **Select variables for sub-clustering (potentially arm-specific)**
    - GEE2 with significative association parameters
    - Network clustering
  - → **Select variables for interference**
    - Compute the subcluster-specific value through a mapping function.
    - Test for association with the outcome

- **Step 2 : Augmented estimation**
  - → **Run the AIPW estimator**

$$\sum_{i=1}^{n} \frac{\partial \mu_i(\beta)}{\partial \beta} V_i^{-1} W_i (Y_i - \mu_i) - \sum_{l} (I(A_i = l) - \mathbb{P}(A_i = l)) \gamma_l^T = 0$$

Acknowledgement

Thanks for your listening !



HARVARD
SCHOOL OF PUBLIC HEALTH
Powerful ideas for a healthier world

- My colleagues : in particular Victor, Rui and Eric.
- The trial participants and trial project teams,
- You again !