Cluster-based outcome-dependent sampling: inference and frameworks for efficient sampling designs

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Biostatistics

have examined a dissertation entitled

"Cluster-Based Outcome-Dependent Sampling: Inference and Frameworks for Efficient Sampling Designs"

presented by Sara M. Sauer

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature .....

Typed name: Prof. Sebastien Haneuse

Signature Michael D Hughes
Michael D Hughes (May 11, 2021 23:04 EDT)

Typed name: Prof. Michael Hughes

Signature bg
Bethany Hedt-Gauthier (May 12, 2021 12:15 EDT)

Typed name: Prof. Bethany Hedt-Gauthier

Signature Jukka-Pekka Onnela
Jukka-Pekka Onnela (May 12, 2021 12:34 EDT)

Typed name: Prof. Jukka-Pekka Onnela

Date: May 11, 2021

Cluster-based outcome-dependent sampling: inference and frameworks for efficient sampling designs

A DISSERTATION PRESENTED
BY
SARA M. SAUER
TO
THE DEPARTMENT OF BIOSTATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOSTATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2021

Dissertation Advisor: Professor Sebastien Haneuse Sara M. Sauer

Cluster-based outcome-dependent sampling: inference and frameworks for efficient sampling designs

Abstract

Efficient sampling designs are valuable in public health research when finite resources necessitate decisions regarding which individuals to sample for detailed data collection. In observational studies, when the outcome is rare, outcome-dependent sampling (ODS) is a cost-efficient strategy that leverages information on the subject outcomes at the design stage to inflate the outcome rate in the sample and thereby increase statistical efficiency. In many settings, the individuals in the target population are clustered, as are patients in health centers, and therefore exhibit cluster-correlation in their outcomes. Logistical, ethical, or resource constraints may require sampling clusters rather than individuals directly. In such settings, the question becomes which *clusters* should be sampled to yield the most 'informative' sample for the research question of interest.

This dissertation focuses on the design and analysis of cluster-based ODS designs, in which cluster-level summaries of the outcome, as well as possibly other pieces of cluster-level, readily-available information from sources such as a country's Health Management Information System (HMIS), is used to guide the decision regarding which clusters to sample. In particular, this dissertation proposes methods for i) valid estimation and inference given data collected through a cluster-based ODS design when the number of sampled clusters is small, and ii) a framework for designing efficient cluster-based ODS designs, when interest lies in estimating with precision one or multiple parameters in a marginal mean model.

In Chapter 1, I propose to carry out inference given data collected through a cluster-based ODS scheme using inverse-probability-weighted generalized estimating equations (IPW-GEE), where the cluster-specific weights are the inverse of a cluster's probability of selection into the sample. I provide a detailed treatment of the asymptotic properties of this estimator, together with an explicit expression for the asymptotic variance and a corresponding estimator. Furthermore, motivated by a study on risk factors for low birthweight in Rwanda, I propose a number of small-sample bias corrections to the point estimates and standard error estimates. In Chapter 2, I develop an approach for optimal allocation in single-stage stratified cluster-based ODS designs and investigate the potential for gains in statistical efficiency under such a design given one or multiple parameters of interest. As the optimal allocation formulae presented in Chapter 2 depend on quantities that are unknown in practice, Chapter 3 proposes and evaluates an adaptive sampling strategy for operationalizing the optimal allocation design in practice. Finally, in Chapter 4 I give concluding remarks and present some directions for future work.

Contents

Title page

Copyright

Abstract iii

Listing of figures viii

List of Tables xvii

Acknowledgements xix

0 Introduction 1

1 Small-sample inference for cluster-based outcome-dependent sampling schemes in resource-limited settings: investigating low birthweight in Rwanda 5

1.1 Introduction 7

1.2 Risk Factors for Low Birthweight in Rwanda 9
1.3 Analysis Based on Complete Data 11
1.4 Cluster-Based Outcome-Dependent Sampling 13
1.5 Analysis Based on Data from a Cluster-Based ODS Design 16
1.6 Simulation Study 20
1.7 Data Application 30
1.8 Discussion 32

2 Optimal allocation in stratified cluster-based outcome-dependent sampling designs **36**
2.1 Introduction 38
2.2 Increasing Facility-Based Births in Zanzibar 42
2.3 Setting 47
2.4 Optimal Allocation 49
2.5 Simulation Study 53
2.6 Results 60
2.7 Discussion 64

3 Practical strategies for operationalizing optimal allocation in stratified cluster-based outcome-dependent sampling designs **69**
3.1 Introduction 71
3.2 Review of Optimal Allocation in Stratified Cluster-Based ODS 73
3.3 Two-Wave Adaptive Sampling 76
3.4 Simulation Study 80
3.5 Results 83
3.6 Data Application 88

3.7 Discussion 98

4 Conclusion 101

4.1 Two-Stage Stratified Cluster-Based ODS Designs 103

4.2 Other Areas for Future Work 108

Appendix A Supplementary material to accompany Chapter 1 110

A.1 Asymptotic Theory for GEE 110

A.2 Asymptotic Theory for WGEE 117

A.3 Bias Correction for Point Estimates under Cluster-Based Outcome-Dependent Sampling Design 140

A.4 Mancl and DeRouen-type Variance Bias Correction 143

A.5 Kauermann and Carroll-type Variance Bias Correction 148

A.6 Fay and Graubard-type Variance Bias Correction 150

A.7 Smoothed Bootstrap 155

A.8 Simulation Results: Absolute Bias in Point Estimates 157

A.9 Simulation Results: Coverage Probabilities 162

Appendix B Supplementary material to accompany Chapter 2 194

B.1 Initial Simulation to Investigate Potential for Efficiency Gains 194

B.2 Complete Simulation Study Results 199

Appendix C Supplementary material to accompany Chapter 3 238

C.1 Algorithm for Handling Edge Cases 238

References 247

Listing of figures

1.1 Scatterplot of health center-specific number of live births and prevalence of low birthweight births between April and June 2017, at each of K=44 health centers in two districts in northern Rwanda. Relative sizes of the circles are proportional to the probability of selection into the outcome-dependent scheme, π_k. Black shading indicates which of the 44 health centers were ultimately selected. 10

1.2 The absolute bias in the mean of the unweighted point estimates $\widehat{\boldsymbol{\beta}}^{(s)}$, weighted uncorrected point estimates $\widehat{\boldsymbol{\beta}}_w$, and the bias-corrected weighted point estimates $\widehat{\boldsymbol{\beta}}_w^c$ in Simulation 1, as a function of $K_s \in \{12, 18, 24\}$, under designs #1, #2, and #3. The degree of correlation was determined by $\sigma_V = 0.5$. The true parameter values are given by $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ = (-1.6, 0.5, 0.3). Under design #1, $\widehat{\boldsymbol{\beta}}^{(s)}$ and $\widehat{\boldsymbol{\beta}}_w$ are equivalent. 25

1.3 The absolute bias in the mean of the unweighted point estimates $\widehat{\boldsymbol{\beta}}^{(s)}$, weighted uncorrected point estimates $\widehat{\boldsymbol{\beta}}_w$, and the bias-corrected weighted point estimates $\widehat{\boldsymbol{\beta}}_w^c$ in Simulation 2, as a function of $K_s \in \{15, 20, 30, 50, 100\}$, under designs #1, #2, and #3. The degree of correlation was determined by $\sigma_V = 0.5$. The true parameter values are given by $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ = (-3.1, 0.5, 0.5, 0.5, 1). Under design #1, $\widehat{\boldsymbol{\beta}}^{(s)}$ and $\widehat{\boldsymbol{\beta}}_w$ are equivalent. 26

1.4 The absolute bias in the mean of the unweighted point estimate $\widehat{\beta}_4^{(s)}$, weighted uncorrected point estimate $\widehat{\beta}_{4,w}$, and the bias-corrected weighted point estimate $\widehat{\beta}_{4,w}^c$ in Simulation 2 as a function of the prevalence of $X_4 \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$, with the degree of correlation determined by $\sigma_V = 0.5$. 27

2.1 Distribution of the standard errors of $\widehat{\beta}_{ANC}$ from the 500 simple random sampling designs across 1000 iterations. 44

2.2 Baseline scenario: K=280, $N_k = 40 \; \forall \; k = 1, \ldots, K$, $K_s = 80$, $\sigma_V = 0.5$. Shown is (i) the percent bias in the mean point estimates (top panel), (ii) the estimated coverage probabilities for confidence intervals (middle panel) using unadjusted estimated standard errors (UA), estimated standard errors with a degrees-of-freedom adjustment (DF), estimated standard errors with a Mancl and DeRouen-type correction (MD), and estimated standard errors with a Kauermann and Carroll-type correction (KC), and (iii) the uncertainty relative to simple random sampling (bottom panel) with the dark blue, red, and light blue bars corresponding to scenarios 1, 2, and 3 described in Section 2.2.5 respectively. 57

2.3 Positive association X_1 and X_2: K=280, $N_k = 40 \; \forall \; k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. Shown is (i) the percent bias in the mean point estimates (top panel), (ii) the estimated coverage probabilities for confidence intervals (middle panel) using unadjusted estimated standard errors (UA), estimated standard errors with a degrees-of-freedom adjustment (DF), estimated standard errors with a Mancl and DeRouen-type correction (MD), and estimated standard errors with a Kauermann and Carroll-type correction (KC), and (iii) the uncertainty relative to simple random sampling (bottom panel) with the dark blue, red, and light blue bars corresponding to scenarios 1, 2, and 3 described in Section 2.2.5 respectively. 58

2.4 Negative association X_1 and X_2: K=280, $N_k = 40 \; \forall \; k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. Shown is (i) the percent bias in the mean point estimates (top panel), (ii) the estimated coverage probabilities for confidence intervals (middle panel) using unadjusted estimated standard errors (UA), estimated standard errors with a degrees-of-freedom adjustment (DF), estimated standard errors with a Mancl and DeRouen-type correction (MD), and estimated standard errors with a Kauermann and Carroll-type correction (KC), and (iii) the uncertainty relative to simple random sampling (bottom panel) with the dark blue, red, and light blue bars corresponding to scenarios 1, 2, and 3 described in Section 2.2.5 respectively. 59

3.1 Shown is (sd($\widehat{\beta}^{1:R}_{q,Adapt}$) - sd($\widehat{\beta}^{1:R}_{q,Opt}$))/sd($\widehat{\beta}^{1:R}_{q,Opt}$), $q = 1, \ldots, p$ under the adaptive sampling strategy using IPW estimation (reds) and multiple imputation (blues), for $K_s = 80$ and $K_{s,1} \in \{20, 40, 60\}$ under positive dependence X_1 and X_2 (top panel), and negative dependence X_1 and X_2 (bottom panel). 89

3.2 Shown is (sd($\widehat{\beta}^{1:R}_{q,Adapt}$) - sd($\widehat{\beta}^{1:R}_{q,Opt}$))/sd($\widehat{\beta}^{1:R}_{q,Opt}$), $q = 1, \ldots, p$ under the adaptive sampling strategy using IPW estimation (reds) and multiple imputation (blues), for $K_s = 80$ and $K_{s_1} \in \{20, 40, 60\}$ under positive dependence X_1 and X_3 (top panel), and negative dependence X_1 and X_3 (bottom panel). 90

3.3 The number of women delivering outside of a health facility, in Unguja and Pemba. Each point represents one shehia; those shaded blue indicate the $K_{s,1} = 40$ shehias selected in the first wave. 95

A.1 Absolute bias in point estimates, σ_V=0.25, working independence 158

A.2 Absolute bias in mean point estimates, σ_V=0.5, working independence 159

A.3 Absolute bias in mean point estimates, σ_V=0.75, working independence 160

A.4 Absolute bias in mean point estimates, σ_V=0.5, working exchangeable 161

A.5 Estimated coverage probabilities with $\widehat{\beta}_w$, using normal distribution for confidence interval construction, σ_V=0.5 162

A.6 Estimated coverage probabilities with $\widehat{\beta}_w$, using t distribution for confidence interval construction, σ_V=0.5 163

A.7 Estimated coverage probabilities with $\widehat{\beta}^c_w$, using normal distribution for confidence interval construction, σ_V=0.5 164

A.8 Estimated coverage probabilities with $\widehat{\beta}^c_w$, using t distribution for confidence interval construction, σ_V=0.5 165

A.9 Estimated coverage probabilities with $\widehat{\beta}_w$, using normal distribution for confidence interval construction, ignoring negative correlation, σ_V=0.5 166

A.10 Estimated coverage probabilities with $\widehat{\beta}_w$, using t distribution for confidence interval construction, ignoring negative correlation, σ_V=0.5 167

A.11 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using normal distribution for confidence interval construction, ignoring negative correlation, σ_V=0.5 168
A.12 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using t distribution for confidence interval construction, ignoring negative correlation, σ_V=0.5 169
A.13 Estimated coverage probabilities with $\widehat{\beta}_{w}$, using normal distribution for confidence interval construction, σ_V=0.25 170
A.14 Estimated coverage probabilities with $\widehat{\beta}_{w}$, using t distribution for confidence interval construction, σ_V=0.25 171
A.15 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using normal distribution for confidence interval construction, σ_V=0.25 172
A.16 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using t distribution for confidence interval construction, σ_V=0.25 173
A.17 Estimated coverage probabilities with $\widehat{\beta}_{w}$, using normal distribution for confidence interval construction, σ_V=0.25, ignoring negative correlation 174
A.18 Estimated coverage probabilities with $\widehat{\beta}_{w}$, using t distribution for confidence interval construction, σ_V=0.25, ignoring negative correlation 175
A.19 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using normal distribution for confidence interval construction, σ_V=0.25, ignoring negative correlation 176
A.20 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using t distribution for confidence interval construction, σ_V=0.25, ignoring negative correlation 177
A.21 Estimated coverage probabilities with $\widehat{\beta}_{w}$, using normal distribution for confidence interval construction, σ_V=0.75 178
A.22 Estimated coverage probabilities with $\widehat{\beta}_{w}$, using t distribution for confidence interval construction, σ_V=0.75 179

A.23 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using normal distribution for confidence interval construction, σ_V=0.75 180
A.24 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using t distribution for confidence interval construction, σ_V=0.75 181
A.25 Estimated coverage probabilities with $\widehat{\beta}_{w}$, using normal distribution for confidence interval construction, σ_V=0.75, ignoring negative correlation 182
A.26 Estimated coverage probabilities with $\widehat{\beta}_{w}$, using t distribution for confidence interval construction, σ_V=0.75, ignoring negative correlation 183
A.27 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using normal distribution for confidence interval construction, σ_V=0.75, ignoring negative correlation 184
A.28 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using t distribution for confidence interval construction, σ_V=0.75, ignoring negative correlation 185
A.29 Estimated coverage probabilities with $\widehat{\beta}_{w}$, using normal distribution for confidence interval construction, σ_V=0.5, working exchangeable 186
A.30 Estimated coverage probabilities with $\widehat{\beta}_{w}$, using t distribution for confidence interval construction, σ_V=0.5, working exchangeable 187
A.31 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using normal distribution for confidence interval construction, σ_V=0.5, working exchangeable 188
A.32 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using t distribution for confidence interval construction, σ_V=0.5, working exchangeable 189
A.33 Estimated coverage probabilities with $\widehat{\beta}_{w}$, using normal distribution for confidence interval construction, σ_V=0.5, working exchangeable, ignoring negative correlation 190
A.34 Estimated coverage probabilities with $\widehat{\beta}_{w}$, using t distribution for confidence interval construction, σ_V=0.5, working exchangeable, ignoring negative correlation 191

A.35 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using normal distribution for confidence interval construction, σ_V=0.5, working exchangeable, ignoring negative correlation 192

A.36 Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using t distribution for confidence interval construction, σ_V=0.5, working exchangeable, ignoring negative correlation 193

B.1 Distribution of the standard errors of β_{ANC} under the 500 simple random sampling designs across 1000 iterations. 198

B.2 The top panel shows the 500 designs plotted as a function of the number of X_{ANC} cases and the mean number of outcome cases across the samples. The dark blue points represent the best (lowest standard deviation of $\widehat{\beta}_{ANC}$) designs, and the red points represent the worst (highest standard deviation of $\widehat{\beta}_{ANC}$) designs. In this setting, the number of outcome cases is proportional to the cluster size. 200

B.3 The top panel shows the 500 designs plotted as a function of the number of X_{ANC} cases and the mean number of outcome cases across the samples. The dark blue points represent the best (lowest standard deviation of $\widehat{\beta}_{ANC}$) designs, and the red points represent the worst (highest standard deviation of $\widehat{\beta}_{ANC}$) designs. In this setting, the number of outcome cases is inversely proportional to the cluster size. 201

B.4 Baseline scenario: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. 202

B.5 Positive association X_1 and X_2: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. 203

B.6 Negative association X_1 and X_2: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. 204

B.7 Positive association X_1 and X_3: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. 205

B.8 Negative association X_1 and X_3: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. 206

B.9 Positive association X_1 and X_4: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. 207
B.10 Negative association X_1 and X_4: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. 208
B.11 Positive association X_1 and X_5: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. 209
B.12 Negative association X_1 and X_5: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. 210
B.13 Baseline scenario: K=280, varying $N_k, K_s = 80, \sigma_V = 0.5$. 211
B.14 Positive association X_1 and X_2: K=280, varying $N_k, K_s = 80, \sigma_V = 0.5$. 212
B.15 Negative association X_1 and X_2: K=280, varying $N_k, K_s = 80, \sigma_V = 0.5$. 213
B.16 Positive association X_1 and X_3: K=280, varying $N_k, K_s = 80, \sigma_V = 0.5$. 214
B.17 Negative association X_1 and X_3: K=280, varying $N_k, K_s = 80, \sigma_V = 0.5$. 215
B.18 Positive association X_1 and X_4: K=280, varying $N_k, K_s = 80, \sigma_V = 0.5$. 216
B.19 Negative association X_1 and X_4: K=280, varying $N_k, K_s = 80, \sigma_V = 0.5$. 217
B.20 Positive association X_1 and X_5: K=280, varying $N_k, K_s = 80, \sigma_V = 0.5$. 218
B.21 Negative association X_1 and X_5: K=280, varying $N_k, K_s = 80, \sigma_V = 0.5$. 219
B.22 Baseline scenario: K=280, equal N_k=40, $K_s = 80, \sigma_V = 1$. 220
B.23 Positive association X_1 and X_2: K=280, equal N_k=40, $K_s = 80, \sigma_V = 1$. 221
B.24 Negative association X_1 and X_2: K=280, equal N_k=40, $K_s = 80, \sigma_V = 1$. 222
B.25 Positive association X_1 and X_3: K=280, equal N_k=40, $K_s = 80, \sigma_V = 1$. 223
B.26 Negative associaiton X_1 and X_3: K=280, equal N_k=40, $K_s = 80, \sigma_V = 1$. 224
B.27 Positive association X_1 and X_4: K=280, equal N_k=40, $K_s = 80, \sigma_V = 1$. 225
B.28 Positive association X_1 and X_4: K=280, equal N_k=40, $K_s = 80, \sigma_V = 1$. 226
B.29 Positive association X_1 and X_5: K=280, equal N_k=40, $K_s = 80, \sigma_V = 1$. 227
B.30 Negative association X_1 and X_5: K=280, equal N_k=40, $K_s = 80, \sigma_V = 1$. 228
B.31 Baseline scenario: K=280, equal N_k=40, $K_s = 40, \sigma_V = 0.5$. 229

B.32 Positive association X_1 and X_2: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$. 230

B.33 Negative association X_1 and X_2: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$. 231

B.34 Positive association X_1 and X_3: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$. 232

B.35 Negative association X_1 and X_3: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$. 233

B.36 Positive association X_1 and X_4: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$. 234

B.37 Negative association X_1 and X_4: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$. 235

B.38 Positive association X_1 and X_5: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$. 236

B.39 Negative association X_1 and X_5: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$. 237

List of Tables

1.1 Maternal and Newborn Characteristics 12

1.2 Estimated coverage of Wald-based 95% confidence intervals in Simulation 1 using methods proposed in Sections 1.5.2 and 1.5.3, using the bias-corrected weighted estimator, $\widehat{\beta}^{c}_{w}$, as the point estimate. See Section 1.6 for details. 28

1.3 Estimated coverage of Wald-based 95% confidence intervals in Simulation 2 using methods proposed in Sections 1.5.2 and 1.5.3, using the bias-corrected weighted estimator, $\widehat{\beta}^{c}_{w}$, as the point estimate. Shown are results for estimators that acknowledge negative correlation in the selection indicators. See Section 1.6 for details. 29

1.4 Results from the analysis of the Rwandan birth dataset. See Section 1.7.1 for details. 32

2.1 Characteristics of the 'best' (lowest standard deviation of $\widehat{\beta}_{ANC}$) and 'worst' (highest standard deviation of $\widehat{\beta}_{ANC}$) designs among the set of 'unbiased' designs for the estimation of β_{ANC}. In Scenario 1, the number of outcome cases is proportional to the cluster size, while in Scenario 2, the number of outcome cases is inversely proportional to the cluster size. See Section 2.2.4 for details. 45

2.2 Covariate distributions for nine simulation scenarios considered in Section 2.5. . . 54

2.3 $K = 280$, varying N_k, K_s=80, σ_V=0.5. Shown are the average stratum-specific cluster-level sample sizes, and the average overall individual-level sample size n, under the eight data scenarios in which there is a dependence between X_1 and one of the other covariates in the model, across 10000 iterations. The average number of clusters sampled from stratum $(Y^* \geq Y^*_{0.80}, X_1 = 1)$ is k_{11}, the average number of clusters sampled from stratum $(Y^* < Y^*_{0.80}, X_1 = 1)$ is k_{01}, the average number of clusters sampled from stratum $(Y^* \geq Y^*_{0.80}, X_1 = 0)$ is k_{10}, and the average number of clusters sampled from stratum $(Y^* < Y^*_{0.80}, X_1 = 0)$ is k_{00}. See Sections 2.5.2 and 2.5.3 for details. 65

3.1 Covariate distributions for nine simulation scenarios considered in Section 3.4. . . 81

3.2 Maternal Characteristics 92

3.3 Estimated odds ratios (OR) and 95% confidence intervals (CI) from IPW-GEE analysis, based on five samples drawn under five different sampling designs. 97

3.4 Stratum specific sample sizes under optimal allocation, adaptive sampling with IPW estimation, and adaptive sampling with MI estimation. 98

B.1 Characteristics of the best (lowest standard deviation of $\widehat{\beta}_{ANC}$) and worst (highest standard deviation of $\widehat{\beta}_{ANC}$) designs among the set of 'unbiased' designs for the estimation of β_{ANC}. In Scenario 1, the number of outcome cases is proportional to the cluster size, while in Scenario 2, the number of outcome cases is inversely proportional to the cluster size. 198

Acknowledgments

I AM DEEPLY GRATEFUL to my advisor Sebastien Haneuse for his support, guidance, and mentorship throughout my time as a PhD student in this department. From him I have learned a tremendous amount about what it means to be a strong statistician, effective researcher, and reliable collaborator. I'd also like to thank Bethany Hedt-Gauthier for her mentorship, for helping me carve a path towards a career at the intersection of global health and statistics, and for her thoughtful questions always aimed at ensuring my work is useful in practical settings. In addition, I'd like to thank my committee members Michael Hughes and Jukka-Pekka Onnela for their engagement, insightful questions, and useful advice that has helped shape the direction and focus of my dissertation.

Finally, I am thankful for the constant support of my family (mom, dad, and brother) and friends throughout my doctoral training, without whom this endeavor would not have been possible.

0

Introduction

Conducting research in low-and-middle-income country (LMIC) settings or community-based settings in the US is often challenging for a host of reasons, including that resources for data collection are limited. When resources are finite, decisions must be made at the study design stage regarding which individuals to sample for detailed data collection. Efficient sampling designs that aim to select the most informative individuals to answer the research question(s) of interest are therefore indispensable in such settings.

Sampling designs that leverage information available at the design stage of a study can increase the statistical efficiency of the final analysis, which is particularly important when the outcome and/or exposure of interest is rare. Designs that use information on the outcome of the individuals in the target population belong to a class of designs known as outcome-dependent sampling (ODS) designs. A classic example of an ODS design is the case-control design, which is known to yield substantial efficiency gains over simple random sampling when the outcome of interest is rare [53]. Other examples of ODS schemes include the nested case-control, case-cohort, and two-phase designs [10,31,74,75]. A rich literature exists on the analysis and design of outcome-dependent sampling designs tailored to different contexts. For the most part, the literature on ODS has focused on the setting where individuals are treated as independent, although methods have been recently proposed for longitudinal and cluster-correlated data settings [23,44,45,46,57,61,65]. However, the majority of previous research on ODS designs has focused on settings in which sampling occurs at the level of the individual.

In this dissertation, we instead focus on ODS designs that involve sampling clusters as opposed to individuals, even while the analysis remains at the level of the individual. Such cluster-based ODS schemes may be useful in settings where i) individuals are cluster-correlated, and ii) logistical constraints permit researchers to visit only a subset of the clusters in the target population for data collection. Rather than using individual-level information on the outcome to guide sampling decisions, cluster-level summaries of the outcome and possible other relevant pieces of information may be used. In the context of public health research, such cluster-level summary measures of health indicators are increasingly available through centralized databases such as a country's Health Management Information System (HMIS) [3,48].

This dissertation aims to answer several questions related to the design and analysis of cluster-based ODS designs. In Chapter 1, we propose to carry out inference given data collected through a cluster-based ODS design using IPW-GEE. Furthermore, we propose a number of small-sample bias

corrections to the point and variance estimates to be used when the number of sampled clusters is 'small'. Through a comprehensive simulation study, we show that i) analysis via IPW-GEE is valid when the number of sampled clusters is large enough, and ii) that the proposed small-sample bias corrections reduce the bias in both the point and variance estimates when the number of sampled clusters is small.

Chapter 1 provides a way to analyze data that has been collected through a cluster-based ODS sampling scheme; it does not, however, address the question of how to choose an efficient sampling strategy among the class of cluster-based ODS designs. In Chapter 2, we propose a framework for optimal allocation in the class of *single-stage stratified cluster-based ODS* designs. In such a design, the readily available cluster-level information on the outcome, as well as possibly other covariates, is used to stratify the clusters in the population of interest. Then, given resources to sample a fixed number of clusters, the optimal allocation of the cluster-level sample size across the defined strata is determined according to a predefined optimality criterion. The primary goal of Chapter 2 is to develop a comprehensive understanding of the potential value (in terms of efficiency) of pursuing an optimal allocation strategy. That is, we seek to provide insight into how the potential for efficiency gain is impacted by different factors such as the optimality criterion, the type of the covariate of interest (cluster-level or individual-level, binary or continuous) and the relationship between the covariate of interest and the stratification variable(s).

One major obstacle to implementing the optimal allocation design in practice, however, is the fact that the formulae for the optimal stratum-specific sample sizes depend on quantities that are unknown, such as the true parameter values, β_0. In Chapter 3, we therefore propose a two-wave adaptive sampling scheme, in which the data collected in the first wave serves as an internal pilot study that is used to estimate the optimal allocation of the remaining resources. We develop and evaluate two approaches to estimating the optimal stratum-specific sample sizes given the first wave data: an inverse-probability weighting (IPW) approach and a multi-level multiple imputation (MI)

approach that employs the imputation approach proposed by Jolani (2015).[29]

Chapters 1-3 in this dissertation are each presented as a self-contained paper. Chapter 4 provides some concluding remarks, extensions, and directions for future work. In particular, we extend the methods presented in the first three chapters for single-stage cluster-based ODS schemes to settings where the sampling design is *two-stage* cluster-based ODS, which arises when researchers are interested not only in sampling clusters, but would also like to sample individuals within the selected clusters. Towards this, we describe how to carry out estimation and inference for data that has been collected through a two-stage stratified cluster-based ODS design. Furthermore, we extend the idea of optimal allocation to this setting, in which decisions must be made not only regarding which clusters to sample, but also which individuals to sample within the selected clusters. We conclude with a brief discussion regarding challenges that arise when implementing such a design, and provide some additional areas for future work.

1

Small-sample inference for cluster-based outcome-dependent sampling schemes in resource-limited settings: investigating low birthweight in Rwanda

Sara Sauer[1], Bethany Hedt-Gauthier[1,2], Claudia Rivera-Rodriguez[3], Sebastien Haneuse[1]

1 *Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA USA*

2 *Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA USA*

3 *Department of Statistics, University of Auckland, Auckland, NZ*

Abstract

The neonatal mortality rate in Rwanda remains above the United Nations Sustainable Development Goal 3 target of 12 deaths per 1,000 live births. As part of a larger effort to reduce preventable neonatal deaths in the country, we conducted a study to examine risk factors for low birthweight. The data was collected via a cost-efficient cluster-based outcome-dependent sampling scheme wherein clusters of individuals (health centers) were selected on the basis of, in part, the outcome rate of the individuals. For a given dataset collected via a cluster-based outcome-dependent sampling scheme, estimation for a marginal model may proceed via inverse-probability-weighted generalized estimating equations, where the cluster-specific weights are the inverse probability of the health center's inclusion in the sample. In this paper, we provide a detailed treatment of the asymptotic properties of this estimator, together with an explicit expression for the asymptotic variance and a corresponding estimator. Furthermore, motivated by the study we conducted in Rwanda, we propose a number of small-sample bias corrections to both the point estimates and the standard error estimates. Through simulation, we show that applying these corrections when the number of clusters is small generally reduces the bias in the point estimates, and results in closer to nominal coverage. The proposed methods are applied to data from 18 health centers and 1 district hospital in Rwanda.

1.1 Introduction

Neonatal mortality is defined as the probability of dying within the first 28 days of life. While the majority of neonatal deaths are preventable, due to slower progress in neonatal mortality reduction since 1990 compared to the advances made in child mortality reduction, neonatal deaths represented 47% of all under-five deaths in 2018[76]. The United Nations Sustainable Development Goal 3 aims to reduce neonatal mortality to 12 deaths per 1,000 live births by 2030[26]. In Rwanda, however, the rate of neonatal deaths remains high at 20 deaths per 1,000 live deaths[58].

Low birthweight is associated with a higher risk of neonatal death[35], strategies that target reduction of low birthweight births may therefore help to reduce neonatal mortality. This motivated a study we recently conducted in Rwanda, with the goal to investigate risk factors for low birthweight in two northern districts of the country. Due to time and resource constraints, it was feasible to visit only 18 of the 44 health centers for collection of the individual-level data. The 18 health centers were sampled based on an outcome-dependent sampling (ODS) design, and in the summer of 2017, the paper's first author traveled to the sampled health centers and collected data on all live births recorded in the maternity registers between April-June 2017.

ODS is an indispensable tool for conducting research in resource-limited settings. Examples of outcome-dependent sampling designs include the case-control, nested case-control, case-cohort, and two-phase designs[10,31,53,74,75]. For the most part, the literature on ODS has focused on the setting where individuals are treated as independent, although methods have been recently proposed for longitudinal and cluster-correlated data settings[23,44,45,46,57,61,65]. However, in both the independent and correlated data settings, the majority of designs proposed involve sampling at the level of the individual (i.e. the study units within a cluster). In some settings researchers may opt to use readily-available outcome information (either aggregated or at the level of the individual) to perform

cluster-based sampling; that is, to sample clusters rather than individuals [12]. Such sampling may be preferred when, for example, the costs associated with travel to a clinic are high compared to the cost of collecting information on individuals once at the clinic.

For data collected through a cluster-based ODS design, Cai et. al (2001) [12] proposed to carry out estimation for a marginally-specified regression model via inverse-probability-weighted generalized estimating equations (IPW-GEE), with the weights taken to be the inverse of the cluster-specific probabilities of being selected by the scheme. Furthermore, they proposed that inference be based on a sandwich estimator for the asymptotic variance. They did not, however, formally establish the asymptotic properties of the estimator for the regression coefficients nor did they provide explicit expressions for the variance that acknowledge the inherent negative correlation among the cluster-specific sampling indicators. In this paper we resolve these gaps by using results by Xie and Yang (2003) [81] for GEE in the complete data setting to establish the asymptotic properties of the IPW-GEE estimator for cluster-based ODS designs. Through this we derive an expression for the asymptotic variance and propose a corresponding consistent estimator.

In contrast to Cai et. al (2001) [12] where the focus was on settings with a large number of small clusters (specifically pairs of eyes within the Baltimore Eye Study), our study of low birthweight risk factors in Rwanda consists of a small number of relatively large clusters. That the number of clusters is small may be of concern, specifically in regard to small-sample bias in point estimates [51] and undercoverage of confidence intervals based on the usual sandwich estimator [16,30,38,43,50]. To the best of our knowledge, however, no attempts have been made to investigate the extent to which these issues manifest in the cluster-based ODS designs that are the focus of this paper. Furthermore, while small-sample corrections have been proposed in the complete data setting, these have not been adapted to the ODS setting. Therefore, a final contribution of this paper is that we provide expressions for a bias-correction to the point estimates, as well as several bias-corrections for the variance estimator.

1.2 Risk Factors for Low Birthweight in Rwanda

The motivating study for this paper is one that we recently conducted on risk factors for low birthweight ($<$ 2,500g) among facility-based births in two districts in northern Rwanda (Gakenke and Rulindo) between April and June in 2017. Within these districts, pregnant women may receive care at one of K=44 health centers. While patient records, that include information on the mother, the pregnancy and the infant, are maintained locally at the health centers, aggregated information on a range of health indicators are tallied on a monthly basis by each of the health centers and entered into the Rwanda Health Management Information System (HMIS), a centralized database maintained by the Rwandan Ministry of Health. Individual-level data, however, is not readily-available through the HMIS and must, therefore, be obtained by traveling to a given health center and abstracting the relevant information.

1.2.1 Sampling Design and Data Collection

At the design phase of this study, for a number of logistical and financial reasons, an early decision was made that individual-level patient data would only be collected from 18 health centers. Towards selecting which health centers to include, health center-specific information on the total number of births (i.e. $N_k, k = 1, \ldots, K$) as well as the (unadjusted) prevalence of low birthweight (i.e. $\overline{Y}_k = N_k^{-1} \sum_i Y_{ki}$) during the three-month study period was made available by HMIS. Figure 1.1 provides a scatterplot; across the 44 health centers N_k ranged from 24 to 194 while $\overline{Y}_k$ ranged from 0 to 0.13.

The information available on N_k and $\overline{Y}_k$ was then used to form the following selection strategy: (i) the six health centers with the highest prevalence were sampled with probability π_k=1.0; (ii) the remaining 12 centers were sampled via Poisson sampling on the basis of selection probabilities determined by the model: logit $\pi_k = \theta_0 + \theta_1 * maxRank_k$, where $maxRank_k$ is the maximum of

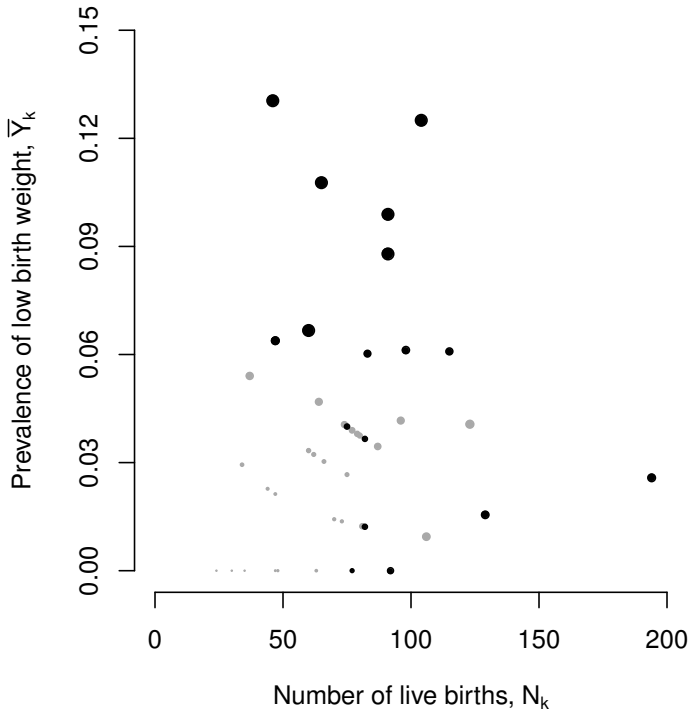

Figure 1.1: Scatterplot of health center-specific number of live births and prevalence of low birthweight births between April and June 2017, at each of K=44 health centers in two districts in northern Rwanda. Relative sizes of the circles are proportional to the probability of selection into the outcome-dependent scheme, π_k. Black shading indicates which of the 44 health centers were ultimately selected.

clinic k's standardized rank with respect to outcome prevalence and its rank with respect to size (in this case, 44 is the highest prevalence/largest size, 1 is the lowest prevalence/smallest size). The value of θ_0 was set to 15/38 to partially control the number of clusters sampled, and the value of θ_1 was set to 0.1.

Note, the relative sizes of the circles in Figure 1.1 are proportional to the value of π_k, while the black shading indicates which of the 44 health centers were ultimately selected. Intuitively, this strategy was adopted to balance inflating the prevalence of low birthweight in the sample (i.e. to artificially increase the prevalence of the outcome, a key feature of outcome-dependent sampling schemes) with maximizing the total sample size in the sub-sample used in the analysis. The particu-

lar design we used for sampling the health centers was selected based on the results of a simulation study conducted beforehand to compare the efficiency gains of various designs that seek to achieve these two objectives.

After visting several health centers, we observed that the high risk deliveries are referred to the district hospitals. In order to have representation of these births in the sample, yet with the constraint of only being able to visit one more location for data collection, we randomly sampled one of the four district hospitals for data collection. Following data abstraction, patient-level information was available on 1635 live facility-based births. The data was restricted to singleton births with complete data on the variables of interest, yielding 1572 observations. Table 1.1 describes the maternal and newborn characteristics of these births. The low birthweight prevalence is higher among mothers younger than 20 years of age, among mothers who weigh less than 56kg, and among women in their first pregnancy. The low birthweight prevalence is also higher among mothers with a history of abortion, c-section, or stillbirth. The low birthweight prevalence is slightly higher among female newborns, and the majority of preterm births are also low birthweight births.

1.3 Analysis Based on Complete Data

1.3.1 Marginal Model Specification

Consider the setting in which the scientific question of interest concerns learning about the relationship between some outcome Y and p-vector of covariates, $\mathbf{X}$ (which may include a 1.0 for the intercept). Furthermore, suppose the population of interest is naturally clustered, such as is the case where patients are clustered within health centers. Let K denote the number of clusters and N_k the number of individuals in the k^{th} cluster. In this paper we assume that the scientific question at hand corresponds to an analysis where estimation and inference will be performed with respect to the following marginal mean model for the outcome of the i^{th} individual in the k^{th} cluster as a function of

Table 1.1: Maternal and Newborn Characteristics

	LBW	Not LBW	Total	% LBW
Mother's Age				
<20	18	135	153	11.8
20-35	74	1060	1134	6.5
36-49	17	268	285	6.0
Mother's Weight				
< 56 kg	44	311	355	12.4
$56 - 59$ kg	33	344	377	8.8
$60 - 64$ kg	18	427	445	4.0
≥ 65 kg	14	381	395	3.5
Birth Order				
1	55	382	437	12.6
2-3	39	653	692	5.6
4+	15	428	443	3.4
HIV status at admission				
Positive	1	24	25	4.0
Negative	107	1433	1540	6.9
Unknown	1	6	7	14.3
Previous abortion				
Yes	10	98	108	9.3
No	99	1365	1464	6.8
Previous C-section				
Yes	14	20	34	41.2
No	95	1443	1538	6.2
Previous stillbirth				
Yes	7	73	80	8.8
No	102	1390	1492	6.8
District				
Gakenke District	73	719	792	9.2
Rulindo District	36	744	780	4.6
Sex of newborn				
Female	62	745	807	7.7
Male	47 12	718	765	6.1
Preterm birth				
Preterm	25	3	28	89.3
Not Preterm	84	1460	1544	5.4
Total	109	1463	1572	6.9

their covariates, $\mathbf{X}_{ki}$:

$$\mu_{ki} = E[Y_{ki}|\mathbf{X}_{ki}] = g^{-1}(\mathbf{X}_{ki}^T\beta), \tag{1.1}$$

where $g(\cdot)$ is a user-chosen link function and β a p-vector of regression parameters.

1.3.2 Complete Data Analysis

Given complete data on $(Y, \mathbf{X})$ for all $N = \sum_{k=1}^{K} N_k$ individuals in the K clusters, Liang and Zeger (1986)[33] proposed that estimation of β can be carried out by solving the following generalized estimating equations:

$$\mathcal{U}(\beta, \alpha) \;=\; \sum_{k=1}^{K} \mathcal{U}_k(\beta, \alpha) \;=\; \sum_{k=1}^{K} \mathbf{D}_k^T \mathbf{V}_k^{-1} \varepsilon_k \;=\; 0, \tag{1.2}$$

where $\varepsilon_k = (\mathbf{Y}_k - \mu_k)$, with $\mathbf{Y}_k = (Y_{k1}, \ldots, Y_{kN_k})$ and $\mu_k = (\mu_{k1}, \ldots, \mu_{kN_k})$, and with $\mathbf{D}_k = \partial\mu_k/\partial\beta$ denoting the $N_k \times p$ matrix of partial derivatives. Finally, $\mathbf{V}_k$, indexed by the unknown α, is an $N_k \times N_k$ working specification for $Cov[\mathbf{Y}_k]$. As will become clear below, it will be useful to write $\mathcal{U}(\beta, \alpha) = \mathbf{U}^T 1_{N\times 1}$ where $\mathbf{U} = diag\{\mathbf{Y} - \mu\}\mathbf{V}^{-1}\mathbf{D}$ is an $N \times p$ matrix, where $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_K)^T$, $\mu = (\mu_1, \ldots, \mu_K)^T$, $\mathbf{V}$ is an $N \times N$ block-diagonal matrix, with the $\mathbf{V}_k$ on the diagonal, and $\mathbf{D}$ is the $N \times p$ matrix obtained by stacking the K $\mathbf{D}_k$ matrices.

1.4 Cluster-Based Outcome-Dependent Sampling

In some settings analysts may not have access to complete data on all elements of $(Y, \mathbf{X})$ for all N individuals in the K clusters. They may, however, have access to resources that permit ascertainment of this information in a sub-sample of, say, $n < N$ individuals. Furthermore, they may have access to select components of $(Y, \mathbf{X})$, as well as other variables/information that are not of direct relevance to the scientific question, denoted here by $\mathbf{Z}$, that can, in principle, be used to make decisions re-

garding the sub-sampling. Moving forward we refer to this information as being available at the *design stage*. How researchers choose to make use of the information available at the design stage will depend, in part, on the precise nature of the information as well as on practical/logistical/financial considerations regarding how the otherwise unavailable data elements will be ascertained.

Motivated by the study we conducted in Rwanda, suppose the readily-available information is of the form $\mathcal{D}_k^* = (N_k, \mathbf{Y}_k^*, \mathbf{X}_k^*, \mathbf{Z}_k^*)$ where $\mathbf{Y}_k^*$ is a cluster-level summary of the outcomes (i.e. across the N_k individuals in the k^{th} cluster), $\mathbf{X}_k^*$ is a cluster-level feature or a cluster-level summary of elements of $\mathbf{X}$ that are readily-available at the outset, and $\mathbf{Z}_k^*$ is a cluster-level summary of $\mathbf{Z}$. For example, if Y is binary, as is low birthweight in the study reported on in Section 1.2, then $\mathbf{Y}_k^*$ may be the proportion of babies born at the health center in the last six months with low birthweight. Furthermore, $\mathbf{X}_k^*$ may be a feature of the cluster, such as whether the health center is in a rural or urban setting, and/or it may be an aggregated summary of individual-level data, such as the percentage of mothers that are less than 18 years of age. Finally, $\mathbf{Z}_k^*$ may be the prevalence of some other outcome or comorbidity that is routinely collected. Note, this data scenario is common in resource-limited settings where detailed information that is recorded and stored locally at health centers is only reported to a centralized agency/ministry after having been aggregated or otherwise summarized [22,39,47].

The information represented by $\mathcal{D}_k^*$ can, in principle, be used to inform a cluster-based outcome-dependent sampling design [12]. For this type of design, rather than selecting individuals directly, some sub-sample of $K_s < K$ clusters is initially selected. Then, the otherwise unavailable elements of $\mathbf{X}$ are ascertained on all individuals within the sampled clusters. For the contexts we consider, and presented in Section 1.2, this would correspond to selecting a certain number of health centers, and then having a member of the research team travel to the health center to extract from the local records all relevant information on patients satisfying the study inclusion/exclusion criteria.

Let R_k be a binary indicator of whether the k^{th} cluster is selected and $\pi_k = \Pr(R_k = 1|\mathcal{D}^*)$ the

corresponding (known) probability of being selected, where $\mathcal{D}^* = \{\mathcal{D}_1^*, \ldots, \mathcal{D}_K^*\}$ is the totality of the information available at the design stage. Towards operationalizing the sampling scheme, researchers may opt to use stratified sampling or Poisson sampling. For the first of these, the clusters are cross-classified on the basis of some variable S that is defined on the basis of one or more of the variables contained in $\mathcal{D}^*$ and is assumed to take on one of J levels. From this, suppose K_j clusters are classified as belonging to the j^{th} stratum. We assume that the stratification scheme is specified such that $K_j > 0$ for all $j = 1, \ldots, J$. Then $k_j \leq K_j$ clusters are randomly selected from those in the j^{th} stratum, such that $\sum_{j=1}^{J} k_j = K_s$. Note, for each of the clusters in the j^{th} stratum, we have that $\pi_k = k_j/K_j$. For those that are selected in this way we set R_k=1. Under the second option of Poisson sampling, one first pre-specifies each of the π_k as a function of elements of $\mathcal{D}^*$. For example, one could specify a logistic regression model for R_k as a function of the clusters' outcome prevalence. Whether a cluster is selected is then determined by an independent Bernoulli trial with probability π_k.

We conclude this section with a number of comments. First, we note that a key difference between the two approaches to selecting the sub-sample of clusters is that K_s is pre-determined under stratified sampling (and, therefore, fixed), but is random under the Poisson sampling. Second, as we elaborate upon below, when the sub-sampling is based on stratification the R_k indicators for clusters within a given stratum are negatively correlated (although they are independent across strata). This, in turn, has implications for the variance of the estimator of β. Under Poisson sampling, however, since the trials are taken to be independent, the R_k are also independent (i.e. both within and between clusters). Finally, the framework described above is sufficiently general that $\mathbf{Y}_k^*$ need not necessarily be used to inform the stratification scheme or the pre-specified model for π_k. Moving forward we assume, however, that $\mathbf{Y}_k^*$ will be used at the design stage.

1.5 Analysis Based on Data from a Cluster-Based ODS Design

When the sampling of the clusters is based, in part, on information on the outcome, the usual generalized estimating equations given by expression (1.2) are no longer guaranteed to be unbiased[54]. To resolve this, Cai et. al (2001)[12] proposed that β be estimated as the solution to the following weighted generalized estimating equations:

$$\mathcal{U}_w(\beta) \;=\; \sum_{k=1}^{K} R_k \pi_k^{-1} \mathbf{D}_k^T \mathbf{V}_k^{-1} \varepsilon_k \;=\; 0. \tag{1.3}$$

Note, following the development of Section 1.3.2, one can write $\mathcal{U}_w(\beta) = \mathbf{U}^T\mathbf{W}\mathbf{R}$, where $\mathbf{W}$ is an $N \times N$ diagonal matrix with diagonal entries equal to the vector $(\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_K)^T$, with $\mathbf{W}_k$ a vector of length N_k with each element equal to π_k^{-1}. Letting $\mathbf{R}_k$ denote the $N_k \times 1$ vector with all entries equal to R_k, $\mathbf{R}$ is the $N \times 1$ vector obtained by concatenating the $\mathbf{R}_k$.

Let $\widehat{\beta}_w$ denote the solution to (1.3). In the remainder of this section, we extend Cai et. al (2001)[12] by: (i) formally establishing the asymptotic properties of $\widehat{\beta}_w$; (ii) proposing an explicit formula for its asymptotic variance; and, (iii) proposing a suite of methods to correct for small-sample (i.e. when K_s is 'small') bias in the point and standard error estimates.

1.5.1 Asymptotic Properties

To establish the asymptotic properties of $\widehat{\beta}_w$ we consider the setting where $K \to \infty$ while $\max\{N_k; k = 1, \ldots, K\}$ is bounded above, and assume missingness at random (MAR), in other words that $\Pr(R_k|\mathcal{D}^*, \mathbf{X}_{ind}) = \Pr(R_k|\mathcal{D}^*)$, where $\mathbf{X}_{ind}$ are the individual-level covariate values for which only a summary is available at the design stage.

Detailed arguments in Appendix B.2, which build on those by Xie and Yang (2003)[81], show that

$\widehat{\beta}_w$ is consistent for β_0, the true value of β, and that

$$\mathbf{V}_T(\beta_0)^{-1/2}\mathbf{M}(\beta_0)^{-1/2}\mathbf{H}(\beta_0)(\widehat{\beta}_w - \beta_0) \;\rightarrow\; MVN(0, \mathbf{I}_{p\times p}),$$

where $\mathbf{M}(\beta) = E[\mathcal{U}(\beta)\mathcal{U}(\beta)^T]$, $\mathbf{H}(\beta) = -E[\partial\mathcal{U}(\beta)/\partial\beta]$, $\mathcal{F}_K$=$\{Y, \mathbf{X}, \mathbf{Z}, \mathbf{S}\}$, and $\mathbf{V}_T(\beta) = Var[\mathbf{M}(\beta)^{-1/2}\mathcal{U}(\beta)] + E[Var[\mathbf{M}(\beta)^{-1/2}\mathcal{U}_w(\beta)|\mathcal{F}_K]]$. Furthermore, from this we have that the asymptotic variance of $\widehat{\beta}_w$ is:

$$Var[\widehat{\beta}_w] \;=\; \mathbf{H}(\beta_0)^{-1}\mathbf{M}(\beta_0)^{1/2}\mathbf{V}_T(\beta_0)\mathbf{M}(\beta_0)^{1/2}\mathbf{H}(\beta_0)^{-1}. \tag{1.4}$$

Finally, $Var[\widehat{\beta}_w]$ can be reexpressed as $\mathbf{H}(\beta_0)^{-1}\{Var[\mathcal{U}_w(\beta)]|_{\beta=\beta_0}\}\mathbf{H}(\beta_0)^{-1}$, with $Var[\mathcal{U}_w(\beta)] = \mathbf{V}_I(\beta) + \mathbf{V}_{II}(\beta)$, where $\mathbf{V}_I(\beta) = Var[\mathbf{U}^T 1_{N\times 1}]$ represents the variance of the complete data estimating equations, and $\mathbf{V}_{II}(\beta) = E[\mathbf{U}^T\mathbf{W}Var[\mathbf{R}|\mathcal{F}_K]\mathbf{W}\mathbf{U}]$ the additional variance that arises due to having only complete data on the individuals in the sampled clusters.

1.5.2 Inference

Inference can be carried out in practice through use of the consistent plug-in estimator for the asymptotic variance of $\widehat{\beta}_w$:

$$\widehat{Var}[\widehat{\beta}_w] \;=\; \widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}\left\{\widehat{\mathbf{V}}_I(\widehat{\beta}_w) + \widehat{\mathbf{V}}_{II}(\widehat{\beta}_w)\right\}\widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}, \tag{1.5}$$

where $\widehat{\mathbf{H}}_w(\beta) = -\sum_{k=1}^K \widehat{\mathbf{H}}_{w,k}(\beta) = -\sum_{k=1}^K \mathbf{D}_k^T\mathbf{V}_k^{-1}\mathbf{W}_k diag(\mathbf{R}_k)\mathbf{D}_k$, and $\widehat{\mathbf{V}}_I(\beta) = \mathbf{U}(\beta)^T diag(\mathbf{R})\mathbf{W}_2^K diag(\mathbf{R})\mathbf{U}(\beta)$, and $\widehat{\mathbf{V}}_{II}(\beta) = \mathbf{U}(\beta)^T\mathbf{W}diag(\mathbf{R})\widetilde{\Delta}diag(\mathbf{R})\mathbf{W}\mathbf{U}(\beta)$.

In the expression for $\widehat{\mathbf{V}}_I(\beta)$, $\mathbf{W}_2^K$ is an $N \times N$ block-diagonal matrix where the entries of the k^{th} $N_k \times N_k$ block are all equal to π_k^{-1}, the inverse of the pairwise selection probability for any pair of individuals i and i' in cluster k: $P(R_{ki} = 1, R_{ki'} = 1) = P(R_{ki} = 1|R_{ki'} = 1)P(R_{ki'} = 1) =$

$1 * \pi_k = \pi_k$. In the expression for $\widehat{\mathbf{V}}_{II}(\beta)$, $\widetilde{\Delta}$ is an $N \times N$ matrix with all entries in the k^{th} block along the diagonal equal to $\frac{\pi_k - \pi_k^2}{\pi_k}$ and all entries in the off-diagonal block corresponding to clusters k and k' equal to $\frac{(\pi_{kk'} - \pi_k \pi_{k'})}{\pi_{kk'}}$. Given data from a cluster-stratified design, if clusters k and k' are both in stratum j, then $\pi_{kk'} = \frac{k_j}{K_j} \frac{k_j - 1}{K_j - 1}$, while if they belong to different strata j and j', it follows that $\pi_{kk'} = \frac{k_j}{K_j} \frac{k_{j'}}{K_{j'}}$. Under a Poisson sampling design, $\pi_{kk'} = \pi_k \pi_{k'}$ for individuals from different clusters. In this case, $\widetilde{\Delta}$ is a block-diagonal matrix, where the k^{th} block is an $N_k \times N_k$ matrix with all values equal to $\frac{\pi_k - \pi_k^2}{\pi_k}$.

In addition, we propose a second estimator of the variance motivated (in part) by the form proposed by Cai et. al (2001)[12]. Towards this, let $\widehat{\varepsilon}_{w,k} = (\mathbf{Y}_k - \widehat{\mu}_{w,k})$, where $\widehat{\mu}_{w,k}$ is an N_k-vector with elements $\widehat{\mu}_{w,ki} = g^{-1}(\mathbf{X}_{ki}^T \widehat{\beta}_w)$. We propose that the asymptotic variance of $\widehat{\beta}_w$ be estimated by:

$$\widehat{Var}^*[\widehat{\beta}_w] = \widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1} \left\{ \widehat{\mathbf{V}}_I(\widehat{\beta}_w) + \widehat{\mathbf{V}}^*_{II}(\widehat{\beta}_w) \right\} \widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}, \tag{1.6}$$

where $\widehat{\mathbf{V}}^*_{II}(\widehat{\beta}_w) = \sum_{k=1}^K B_{kk} \mathbf{D}_k^T \mathbf{V}_k^{-1} \widehat{\varepsilon}_{w,k} \widehat{\varepsilon}_{w,k}^T \mathbf{V}_k^{-1} \mathbf{D}_k$, with $B_{kk} = \pi_k^{-3}(\pi_k - \pi_k^2) R_k$. After some algebra, it can be shown that $\widehat{\mathbf{V}}_{II}(\widehat{\beta}_w) = \widehat{\mathbf{V}}^*_{II}(\widehat{\beta}_w) + \widehat{\mathbf{V}}^e_{II}(\widehat{\beta}_w)$, with

$$\widehat{\mathbf{V}}^e_{II}(\widehat{\beta}_w) = \sum_{k=1}^K \sum_{k' \neq k} \mathbf{D}_k^T \mathbf{V}_k^{-1} diag(\widehat{\varepsilon}_{w,k}) \mathbf{B}_{kk'} diag(\widehat{\varepsilon}_{w,k'}) \mathbf{V}_{k'}^{-1} \mathbf{D}_{k'}$$

where $\mathbf{B}_{kk'}$ is an $N_k \times N_{k'}$ matrix with all entries $\frac{(\pi_{kk'} - \pi_k \pi_{k'}) R_k R_{k'}}{\pi_{kk'} \pi_k \pi_{k'}}$. Thus, expression (1.6) differs from expression (1.5) in that the former ignores the impact of negative correlation between the selection indicators. Note that under Poisson sampling, $\widehat{Var}[\widehat{\beta}_w] = \widehat{Var}^*[\widehat{\beta}_w]$.

1.5.3 Estimation and Inference in Small Samples

As in the complete data setting, estimation and inference for β based on data from an outcome-dependent cluster-based design may be subject to small-sample bias. Indeed, it may be the norm for

studies based on such a design to have a small number of clusters in the dataset available for analysis. Motivated by this we propose bias-corrections to both the point estimates, $\widehat{\beta}_w$, and to $\widehat{Var}[\widehat{\beta}_w]$. For the former, the bias in the point estimates can be expanded to give $E[\widehat{\beta}_w - \beta_0] = \mathbf{B}_w(\beta_0) + O(K^{-2})$ and, building on the work of Paul and Zhang (2014)[51] and Lunardon and Scharfstein (2017)[36], we derive the form of $\mathbf{B}_w(\beta_0)$, which can be estimated by $\mathbf{B}_w(\widehat{\beta}_w)$. Since the resulting expressions are quite involved, we leave the details to Appendix B.3. The bias-corrected point estimate is then $\widehat{\beta}^c_w = \widehat{\beta}_w - \mathbf{B}_w(\widehat{\beta}_w)$.

We consider four corrections to $\widehat{Var}[\widehat{\beta}_w]$ for small-sample settings, adapted from corrections proposed in the complete data setting (see Appendices B.4-B.6 for details). The first is based on a simple 'degrees-of-freedom' adjustment, to give $\widehat{Var}_{DF}[\widehat{\beta}_w] = \frac{K_s}{K_s - p}\widehat{Var}[\widehat{\beta}_w]$. The second follows the approach taken by Mancl and DeRouen (2001)[38], and is given by

$$\widehat{Var}_{MD}[\widehat{\beta}_w] = \widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}\left\{\widehat{\mathbf{V}}_{I,MD}(\widehat{\beta}_w) + \widehat{\mathbf{V}}_{II,MD}(\widehat{\beta}_w)\right\}\widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}, \tag{1.7}$$

where $\widehat{\mathbf{V}}_{I,MD}(\widehat{\beta}_w)\sum_{k=1}^{K}\frac{R_k}{\pi_k}\mathbf{D}_k^T\mathbf{V}_k^{-1}\mathbf{C}_k\widehat{\boldsymbol{\varepsilon}}_{w,k}\widehat{\boldsymbol{\varepsilon}}_{w,k}^T\mathbf{C}_k^T\mathbf{V}_k^{-1}\mathbf{D}_k$ and $\widehat{\mathbf{V}}_{II,MD}(\widehat{\beta}_w) =$ $\sum_{k=1}^{K}B_{kk}\mathbf{D}_k^T\mathbf{V}_k^{-1}\mathbf{C}_k\widehat{\boldsymbol{\varepsilon}}_{w,k}\widehat{\boldsymbol{\varepsilon}}_{w,k}^T\mathbf{C}_k\mathbf{V}_k^{-1}\mathbf{D}_k + \widehat{\mathbf{V}}^e_{II}(\widehat{\beta}_w)$, where $\mathbf{C}_k = (\mathbf{I}_k - \widehat{\mathbf{A}}_{kk,w})^{-1}$ and $\widehat{\mathbf{A}}_{kk,w} = \mathbf{D}_k\widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}\mathbf{D}_k^T\mathbf{V}_k^{-1}\mathbf{W}_k diag(\mathbf{R}_k)$. The third is a Kauermann and Carroll (2001)[30]-type correction, $\widehat{Var}_{KC}[\widehat{\beta}_w]$, which is the same as (1.7), except with $\mathbf{C}_k = (\mathbf{I}_k - \widehat{\mathbf{A}}_{kk,w})^{-\frac{1}{2}}$. The fourth correction adapts the approach taken by Fay and Graubard (2001)[16], and is given by

$$\widehat{Var}_{FG}[\widehat{\beta}_w] = \widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}\left\{\widehat{\mathbf{V}}_{I,FG}(\widehat{\beta}_w) + \widehat{\mathbf{V}}_{II,FG}(\widehat{\beta}_w)\right\}\widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}, \tag{1.8}$$

where $\widehat{\mathbf{V}}_{I,FG}(\widehat{\beta}_w) = \sum_{k=1}^{K}\frac{R_k}{\pi_k}\widehat{\mathbf{F}}_{k,w}\mathbf{D}_k^T\mathbf{V}_k^{-1}\widehat{\boldsymbol{\varepsilon}}_{w,k}\widehat{\boldsymbol{\varepsilon}}_{w,k}^T\mathbf{V}_k^{-1}\mathbf{D}_k(\widehat{\mathbf{F}}_{k,w})^T$ and $\widehat{\mathbf{V}}_{II,FG}(\widehat{\beta}_w) =$ $\sum_{k=1}^{K}B_{kk}\widehat{\mathbf{F}}_{k,w}\mathbf{D}_k^T\mathbf{V}_k^{-1}\widehat{\boldsymbol{\varepsilon}}_{w,k}\widehat{\boldsymbol{\varepsilon}}_{w,k}^T\mathbf{V}_k^{-1}\mathbf{D}_k(\widehat{\mathbf{F}}_{k,w})^T + \widehat{\mathbf{V}}^e_{II}(\widehat{\beta}_w)$, where $\widehat{\mathbf{F}}_{k,w}$ is a $p \times p$ diagonal matrix with the jj^{th} element equal to $(1 - min(b, (\widehat{\mathbf{H}}_{w,k}(\widehat{\beta}_w)\widehat{\mathbf{H}}_w^{-1}(\widehat{\beta}_w))_{jj})^{-1/2}$, and $b = 0.75$.

Finally, we make two additional comments. First, if the data arise from a cluster-stratified design

then one could operationalize these corrections for the estimator of the variance that ignores the negative correlation in the sampling indicators (i.e. for $\widehat{Var}^{*}[\widehat{\beta}_w]$ given in Section 1.5.2), by dropping the $\widehat{\mathbf{V}}^{e}_{II}$ term in the expressions presented. Second, while the bias-corrected variance estimators $\widehat{Var}_{MD}[\widehat{\beta}_w]$, $\widehat{Var}_{KC}[\widehat{\beta}_w]$, and $\widehat{Var}_{FG}[\widehat{\beta}_w]$ are proposed for $\widehat{\beta}_w$, the same form could be adopted in practice using the bias-corrected point estimates $\widehat{\beta}^{c}_w$ instead.

1.6 Simulation Study

To evaluate the operating characteristics of the methods proposed in Section 1.5, we performed two sets of simulation studies. The first set was designed using characteristics of the birth dataset from Rwanda. The second was designed to evaluate the operating characteristics of the methods in a more general setting.

1.6.1 Data Generation

For the first simulation study, we suppose that interest lies in the following marginal logistic regression model for a binary response for the i^{th} individual in the k^{th} cluster:

$$\text{logit}P(Y_{ki} = 1) = \beta_0 + \beta_1 X_{1,ki} + \beta_2 X_{2,k},$$

where $X_{1,ki}$ is a binary individual-level variable with a cluster-specific prevalence, $p_k = Unif[0.1, 0.6]$, $X_{2,k}$ is a continuous cluster-level variable drawn from a Uniform[-2,2] distribution, and $\beta_0 = (\beta_0, \beta_1, \beta_2)$ = (-1.6, 0.5, 0.3). The value of β_0 was chosen so that the prevalence of the outcome is around 21%. Complete datasets with K = 44 clusters were generated with the same size distribution as the 44 health centers in the Rwanda study.

For the second set of simulations we suppose that interest lies in the model:

$$\text{logit}P(Y_{ki} = 1) \;=\; \beta_0 + \beta_1 X_{1,ki} + \beta_2 X_{2,ki} + \beta_3 X_{3,k} + \beta_4 X_{4,k},$$

with $X_{1,ki}$ a binary individual-level variable with the same distribution as in Simulation 1, $X_{3,k}$ a continuous cluster-specific variable drawn from a $N(0.2, 0.05^2)$ distribution, and $X_{4,k}$ a binary cluster-level variable with prevalence 0.20. Finally, $X_{2,ki}$ is an individual-level variable generated from a Normal distribution with cluster-specific mean $E[X_{2,ki}] \;=\; 1 + 0.5X_{4,k}$ and a variance of 1.0. We set $\boldsymbol{\beta}_0 \;=\; (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ = (-3.1, 0.5, 0.5, 0.5, 1), where β_0 was chosen so that the overall prevalence was about 13%. Complete datasets were generated with a total of K = 200 clusters with equal cluster sizes of N_k = 40.

In all simulations, the true correlation structure was exchangeable; we induced correlation among the observations using the `GenBinaryY()` function in the `MMLB` package for `R`, which implements a method for marginally-specified logistic-Normal models[25]. Briefly, the latter permits analysts to specify a marginal mean for the response while inducing within-cluster correlation via cluster-specific random intercepts that are taken to arise from a $N(0, \sigma_V^2)$ distribution. For Simulation 1, we varied $\sigma_V \in \{0.25, 0.5, 0.75\}$, and for Simulation 2, we set σ_V=0.5. The correlation parameter, α, was estimated for each of these scenarios by fitting the model of interest to the generated complete datasets using a working exchangeable correlation structure; this resulted in an estimated correlated parameter of $\widehat{\alpha} \in \{0.01, 0.04, 0.08\}$ for Simulation 1 and $\widehat{\alpha} = 0.01$ for Simulation 2.

1.6.2 Sampling Designs

For each of the simulation scenarios we generated 10,000 'complete' datasets. For each of these we considered three cluster-based designs for selecting $K_s < K$ clusters: (#1) simple random sampling of the clusters; (#2) an outcome-dependent cluster-stratified design, where clusters were stratified

on Y^*_q, the q^{th} quantile of the distribution of the outcome prevalence across the clusters (with $q = 0.75$ in Simulation 1 and $q = 0.80$ in Simulation 2); and, (#3) an outcome-dependent cluster-stratified design where in Simulation 1 clusters were stratified on the basis of $Y^*_{0.75}$ and X^*_2, where $X^*_2 = \mathrm{I}(X_2 > 1)$, while in Simulation 2 clusters were stratified on the basis of $Y^*_{0.80}$ and X_4. Both designs (#2) and (#3) were 'balanced' with respect to the stratification; an equal number of clusters was sampled from each stratum, with the exception of the cases in which K_s could not be evenly divided by the number of strata. In these cases, more clusters were sampled from the strata with higher outcome prevalence (Y^*_q=1). In Simulation 1, we considered designs with $K_s \in \{12, 18, 24\}$, and in Simulation 2 we considered $K_s \in \{15, 20, 30, 50, 100\}$.

1.6.3 Analyses

For each dataset generated through the processes described in Sections 1.6.1 and 1.6.2, we computed three estimates of β: $\widehat{\beta}^{(s)}$, based on (naïvely) solving the unweighted estimating equations given in Section 1.3.2 for the K_s sampled clusters; $\widehat{\beta}_w$, based on solving the weighted estimating equations given at the start of Section 1.5; and, the bias-corrected estimator, $\widehat{\beta}^c_w$, defined in Section 1.5.3. Note, under simple random sampling of the clusters, $\widehat{\beta}^{(s)}$ and $\widehat{\beta}_w$ will be numerically the same. Furthermore, we note that the working independence correlation structure was adopted in the specification of $\mathbf{V}_k$ when computing each of these estimators. The reason for this is that if $\mathbf{X}_{ki}$ includes at least one covariate that varies across units within a cluster (as will almost always be the case), the complete data estimating equations, given by expression (1.2), are not guaranteed to be unbiased unless working independence is adopted [52]. We ran one scenario (Simulation 1, $\sigma_V = 0.5$) with working exchangeable as well, in order to investigate the degree to which using a working correlation structure other than working independence affects the results.

To evaluate the proposed approaches to inference we focus attention on $\widehat{\beta}^c_w$ as a point estimate for β, as it generally exhibits little-to-no bias even when K_s is as small as 12 or 15. We then calculated

six estimates of $Var[\widehat{\beta}_w^c]$, and thus the standard errors, using the methods described in Sections 1.5.2 and 1.5.3. The first three were: $\widehat{Var}[\widehat{\beta}_w^c]$, the unadjusted estimator; the degrees-of-freedom adjusted estimator, $\widehat{Var}_{DF}[\widehat{\beta}_w^c]$; and the Mancl and DeRouen-type estimator, $\widehat{Var}_{MD}[\widehat{\beta}_w^c]$. Another three were as these but ignoring the negative correlation in the sampling indicators: $\widehat{Var}^*[\widehat{\beta}_w^c]$, $\widehat{Var}^*_{DF}[\widehat{\beta}_w^c]$, and $\widehat{Var}^*_{MD}[\widehat{\beta}_w^c]$. In Simulation 1, we also computed the the Kauermann and Carroll-type estimator, $\widehat{Var}_{KC}[\widehat{\beta}_w^c]/\widehat{Var}^*_{KC}[\widehat{\beta}_w^c]$ and the Fay and Graubard-type estimator, $\widehat{Var}_{FG}[\widehat{\beta}_w^c]/\widehat{Var}^*_{FG}[\widehat{\beta}_w^c]$, as well as all of the above variance estimates using the uncorrected point estimates, $\widehat{\beta}_w$, in place of $\widehat{\beta}_w^c$. The final approach towards carrying out inference that we considered was a smoothed bootstrap approach adapted from the work of Li and Wang (2008)[32], with details given in Appendix B.7. Using each of the standard error estimates, we constructed 95% confidence intervals using both the normal distribution and the t_{K_s-p} distribution, estimating coverage as the proportion of iterations in which the constructed confidence intervals contained the true parameter values.

1.6.4 Results: Point Estimation

Figures 1.2 and 1.3 summarize the mean absolute bias in the point estimates (across the 10,000 such estimates) for each of the three sampling designs, under the scenarios considered in Simulation 1 and Simulation 2. Based on these results, we make the following observations:

First, in general, $\widehat{\beta}^{(s)}$ exhibits bias except, as is to be expected, under simple random sampling when K_s is relatively large. Second, for the two outcome-dependent schemes, the weighted estimator $\widehat{\beta}_w$ generally exhibits less bias than $\widehat{\beta}^{(s)}$. While there appear to be some exceptions to this (see Figure 1.2(e), and Figures 1.3(f), (h) and (i)), the extent of bias is very small and it is unclear if any meaning can be attributed to the relative ordering.

Third, the bias in $\widehat{\beta}_w^c$ is generally lower than the bias in $\widehat{\beta}_w$; the additional use of the bias-correction method proposed in Section 1.5.3 generally reduces bias, albeit to a small degree. An exception to this arises in Figure 1.2(b), although the difference is very small. Additional exceptions are in Fig-

ures 1.3(j), (k), (m), and (o), each of which correspond to one of the cluster-level covariates, X_3 and X_4. This may speak to a fundamental difficulty of estimating associations for cluster-level covariates when K_s is small, one that may only be resolved through the collection of additional data (rather than through analytic means). From Figure 1.4 we also see that the extent of bias may depend on the prevalence of a cluster-level covariate that is used to define S in a cluster-stratified design.

Finally, the bias-corrected weighted point estimates, $\widehat{\beta}^c_w$, generally exhibit little-to-no bias across all parameters, even when K_s is a low as 12 or 15.

1.6.5 Results: Coverage

Tables 1.2 and 1.3 report select results regarding coverage of Wald-based 95% confidence intervals in Simulations 1 and 2, respectively. Comprehensive results, specifically for coverage under a broader range of values for K_s and σ_V, as well as coverage using the t distribution, use of the uncorrected point estimates in standard error estimation, and use of working exchangeable correlation matrix, are given in Appendix A.9.

From the first column of results in Tables 1.2 and 1.3, use of the plug-in estimator, $\widehat{Var}[\widehat{\beta}^c_w]$, generally leads to undercoverage when K_s=12 or 15, respectively. This is especially the case for design #3 where the coverage ranges from 0.85-0.87 in Table 1.2, and is even more pronounced in the more complex setting of Simulation 2 with coverage ranging from 0.79-0.84. When K_s is increased to 24 or 50 the performance of confidence intervals based on $\widehat{Var}[\widehat{\beta}^c_w]$ is improved, though there is still undercoverage for all parameters.

Use of any of the proposed small-sample corrections yields improved coverage compared to the unadjusted approach. Among these, the Mancl and DeRouen-type estimator, $\widehat{Var}_{MD}[\widehat{\beta}^c_w]$, generally exhibits the closest to nominal coverage when the standard normal distribution is used for confidence interval construction. When the t_{K_s-p} distribution is used for confidence interval construction (see Appendix A.9), the Mancl and DeRouen-type estimator tends to be conservative, par-

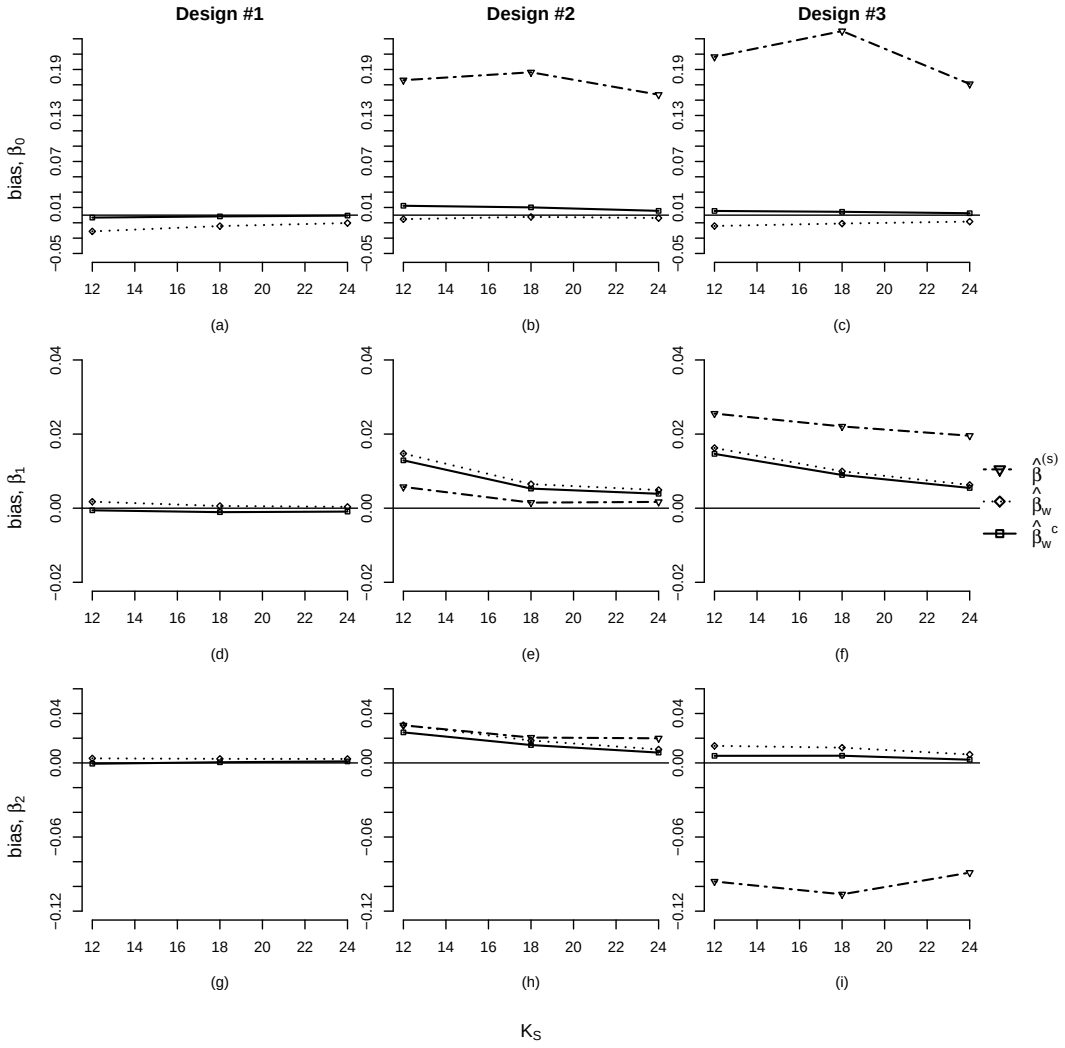

Figure 1.2: The absolute bias in the mean of the unweighted point estimates $\widehat{\boldsymbol{\beta}}^{(s)}$, weighted uncorrected point estimates $\widehat{\boldsymbol{\beta}}_w$, and the bias-corrected weighted point estimates $\widehat{\boldsymbol{\beta}}_w^c$ in Simulation 1, as a function of $K_s \in \{12, 18, 24\}$, under designs #1, #2, and #3. The degree of correlation was determined by $\sigma_V = 0.5$. The true parameter values are given by $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ = (-1.6, 0.5, 0.3). Under design #1, $\widehat{\boldsymbol{\beta}}^{(s)}$ and $\widehat{\boldsymbol{\beta}}_w$ are equivalent.

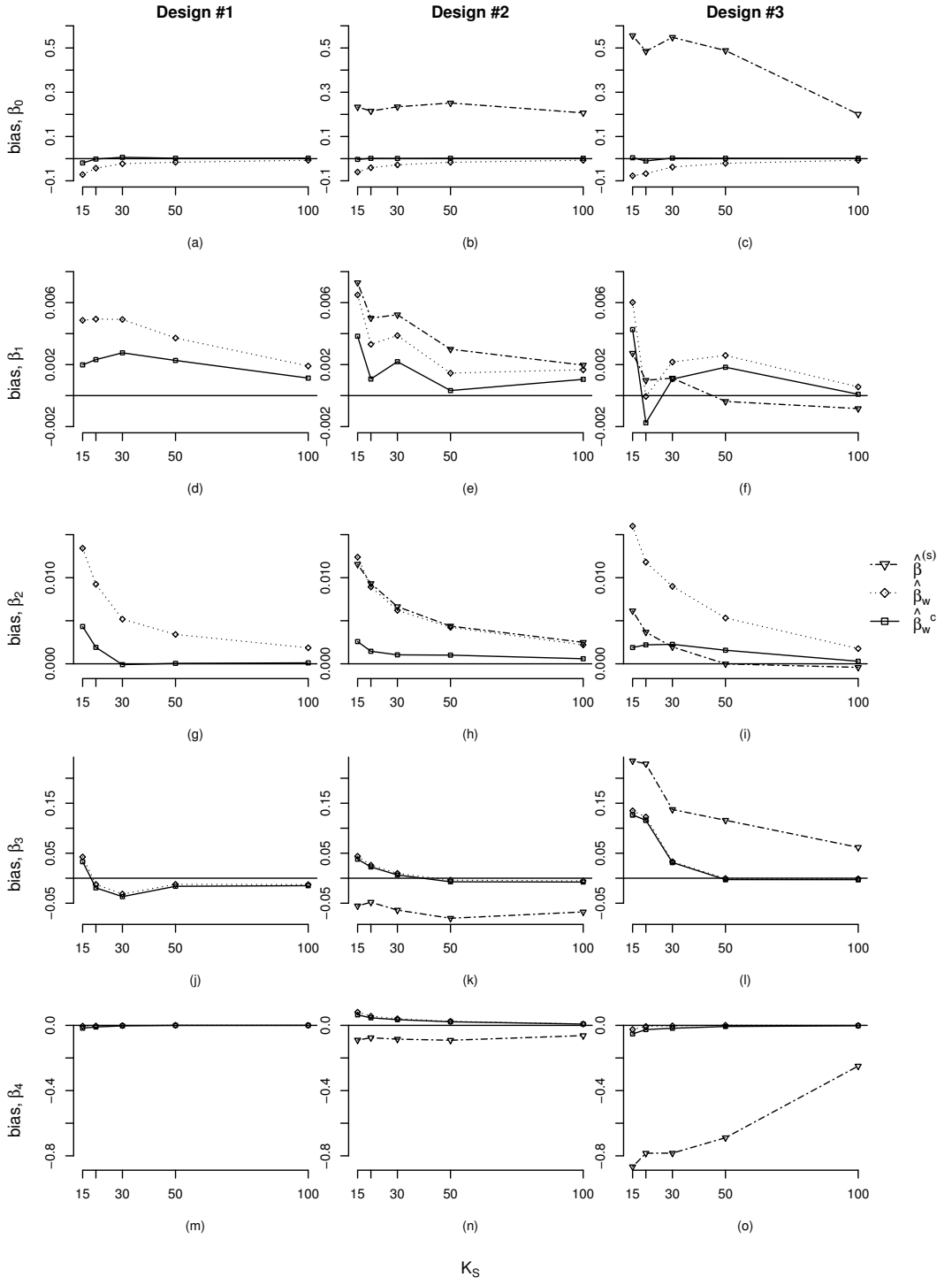

Figure 1.3: The absolute bias in the mean of the unweighted point estimates $\widehat{\boldsymbol{\beta}}^{(s)}$, weighted uncorrected point estimates $\widehat{\boldsymbol{\beta}}_w$, and the bias-corrected weighted point estimates $\widehat{\boldsymbol{\beta}}_w^c$ in Simulation 2, as a function of $K_s \in \{15, 20, 30, 50, 100\}$, under designs #1, #2, and #3. The degree of correlation was determined by $\sigma_V = 0.5$. The true parameter values are given by $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ = (-3.1, 0.5, 0.5, 0.5, 1). Under design #1, $\widehat{\boldsymbol{\beta}}^{(s)}$ and $\widehat{\boldsymbol{\beta}}_w$ are equivalent.

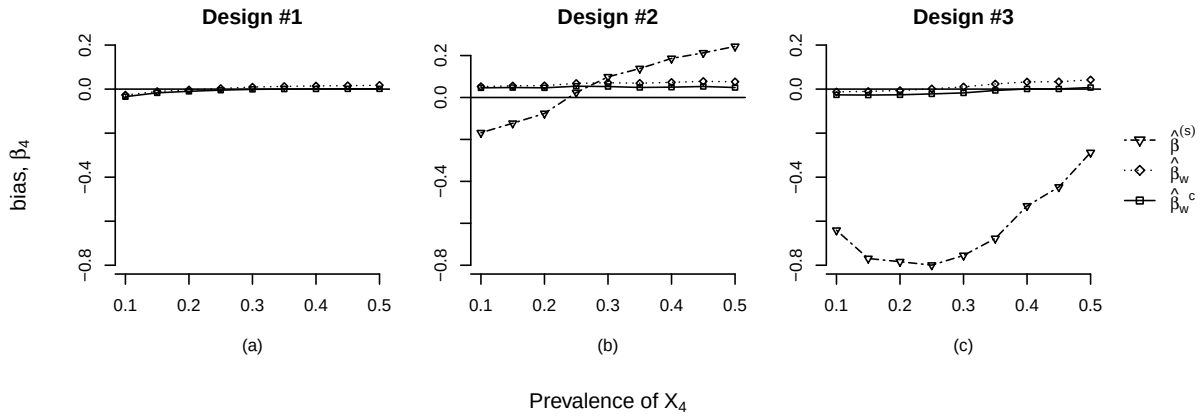

Figure 1.4: The absolute bias in the mean of the unweighted point estimate $\widehat{\beta}_4^{(s)}$, weighted uncorrected point estimate $\widehat{\beta}_{4,w}$, and the bias-corrected weighted point estimate $\widehat{\beta}_{4,w}^{c}$ in Simulation 2 as a function of the prevalence of $X_4 \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$, with the degree of correlation determined by $\sigma_V = 0.5$.

ticularly when working exchangeable is used or $\sigma_V = 0.75$, while the other bias-corrected variance estimators see improved coverage, though generally remain anti-conservative. Across all variance estimators, regardless of whether the normal or t distribution is used for confidence interval construction, the coverage resulting from using estimators with $\widehat{\boldsymbol{\beta}}_w$ is generally very similar to the coverage when using $\widehat{\boldsymbol{\beta}}_w^c$. Furthermore, the difference in coverage when using the variance estimators ignoring the negative correlation in the selection indicators generally decreases as K_s increases, with the directionality of the difference for smaller K_s depending on the parameter in question, design used, and the degree of within-cluster correlation. The findings also hold when working exchangeable correlation structure is used as opposed to working independence.

Table 1.2: Estimated coverage of Wald-based 95% confidence intervals in Simulation 1 using methods proposed in Sections 1.5.2 and 1.5.3, using the bias-corrected weighted estimator, $\widehat{\beta}^{c}_{w}$, as the point estimate. See Section 1.6 for details.

Parameter	K_s/ Design	$\widehat{Var}[\widehat{\beta}^{c}_{w}]$	$\widehat{Var}_{DF}[\widehat{\beta}^{c}_{w}]$	$\widehat{Var}_{MD}[\widehat{\beta}^{c}_{w}]$	$\widehat{Var}_{KC}[\widehat{\beta}^{c}_{w}]$	$\widehat{Var}_{FG}[\widehat{\beta}^{c}_{w}]$
β_0	12/ #1	0.89	0.93	0.94	0.91	0.90
	#2	0.89	0.92	0.95	0.92	0.91
	#3	0.85	0.89	0.93	0.89	0.88
	24/ #1	0.92	0.94	0.95	0.93	0.93
	#2	0.92	0.94	0.95	0.94	0.93
	#3	0.91	0.92	0.94	0.92	0.92
β_1	12/ #1	0.90	0.94	0.94	0.92	0.92
	#2	0.89	0.93	0.94	0.92	0.92
	#3	0.87	0.91	0.93	0.90	0.90
	24/ #1	0.92	0.94	0.95	0.94	0.94
	#2	0.92	0.94	0.94	0.93	0.93
	#3	0.92	0.94	0.95	0.93	0.93
β_2	12/ #1	0.87	0.91	0.94	0.91	0.90
	#2	0.86	0.90	0.93	0.90	0.88
	#3	0.85	0.89	0.94	0.90	0.88
	24/ #1	0.91	0.93	0.94	0.93	0.92
	#2	0.91	0.93	0.95	0.93	0.92
	#3	0.91	0.93	0.95	0.93	0.92

Table 1.3: Estimated coverage of Wald-based 95% confidence intervals in Simulation 2 using methods proposed in Sections 1.5.2 and 1.5.3, using the bias-corrected weighted estimator, $\widehat{\beta}^{c}_{w}$, as the point estimate. Shown are results for estimators that acknowledge negative correlation in the selection indicators. See Section 1.6 for details.

Parameter	$K_s = 15$			$K_s = 50$		
/Design	$\widehat{Var}[\widehat{\beta}^{c}_{w}]$	$\widehat{Var}_{DF}[\widehat{\beta}^{c}_{w}]$	$\widehat{Var}_{MD}[\widehat{\beta}^{c}_{w}]$	$\widehat{Var}[\widehat{\beta}^{c}_{w}]$	$\widehat{Var}_{DF}[\widehat{\beta}^{c}_{w}]$	$\widehat{Var}_{MD}[\widehat{\beta}^{c}_{w}]$
β_0/ #1	0.86	0.92	0.94	0.92	0.93	0.94
#2	0.86	0.92	0.94	0.92	0.94	0.95
#3	0.80	0.87	0.94	0.91	0.93	0.94
β_1/ #1	0.91	0.96	0.94	0.94	0.95	0.95
#2	0.91	0.96	0.94	0.94	0.95	0.95
#3	0.84	0.90	0.91	0.93	0.94	0.94
β_2/ #1	0.91	0.96	0.94	0.94	0.95	0.95
#2	0.91	0.96	0.94	0.94	0.95	0.95
#3	0.84	0.90	0.91	0.93	0.94	0.94
β_3/ #1	0.85	0.91	0.94	0.92	0.93	0.95
#2	0.85	0.91	0.94	0.92	0.94	0.95
#3	0.79	0.86	0.94	0.91	0.92	0.94
β_4/ #1	0.78	0.85	0.88	0.92	0.93	0.95
#2	0.87	0.93	0.95	0.91	0.92	0.93
#3	0.84	0.90	0.95	0.93	0.94	0.95

1.7 Data Application

1.7.1 Analysis

For the purpose of applying the methods proposed in Section 1.5, we focus attention on a logistic regression model for the binary outcome of low birthweight as a function of seven covariates: maternal age, in years (categorized as 1: $<$20; 2: 20-35; 3: 36-49); maternal weight, in kg (categorized as 1: $<$56; 2: 56-59; 3: 60-64; 4: $\geq$ 65); birth order (categorized as 1: 1^{st}; 2: 2^{nd} or 3^{rd}; 3: 4^{th} or higher); sex of the newborn (1: female; 0: male); whether the mother had a previous abortion (1: yes; 0: no), whether the woman had a previous stillbirth (1: yes; 0: no); and, the district the health center/district hospital is located in (1: Rulindo; 0: Gakenke). For this model we estimated and report adjusted odds ratios (OR) based on: $\widehat{\beta}^{(s)}$, the solution to the unweighted estimating equations given in Section 1.3.2; $\widehat{\beta}_w$, based on solving the weighted estimating equations given at the start of Section 1.5; and, the bias-corrected estimator, $\widehat{\beta}^c_w$, defined in Section 1.5.3. In addition, we calculated Wald-based 95% confidence intervals (CI) for the bias-corrected estimator of each OR based on $\widehat{Var}[\widehat{\beta}^c_w]$, the plug-in estimator given by expression (1.5); the degrees-of-freedom adjusted estimator, $\widehat{Var}_{DF}[\widehat{\beta}^c_w]$; and the Mancl and DeRouen-type estimator, $\widehat{Var}_{MD}[\widehat{\beta}^c_w]$. Finally, as in the simulation studies, working independence was adopted throughout (see Section 1.6.3).

1.7.2 Results

Table 1.4 provides a summary of the results. We generally see a difference in the unweighted, weighted, and bias-corrected point estimates, though the story (direction) remains the same: among women weighing less than 56 kg, giving birth to a male as their first child in Gakenke district, with no history of abortion or previous stillbirth, younger women ($<$20) and older women (ages 36-49) have a higher odds of having a low birthweight baby than women ages 25-34. After adjusting for the other

covariates in the model, the higher the mother's weight the lower the odds of having a low birthweight baby. Similarly, after adjusting for the other covariates in the model, as birth order increases women have a lower odds of having a low birthweight baby than women having their first child, female newborns have a higher odds of being low birthweight compared to male newborns, and women who have had a previous abortion or have had a previous stillbirth have a higher odds of having a low birthweight birth compared to women who have no history of previous abortions or no history of previous stillbirths, respectively.

The confidence intervals using both the degrees-of-freedom adjusted standard errors and the Mancl and DeRouen-type adjusted standard errors generally result in wider confidence intervals than those constructed using the unadjusted standard errors; a significant effect of age vanishes when either correction is applied. In some instances, one correction maintains significance while the other does not: for the effect of stillbirth, the confidence interval using the degrees-of-freedom adjusted standard errors contains 1, suggesting that the effect is not statistically significant, whereas the confidence interval using the Mancl and DeRouen-type adjustment does not contain 1, suggesting a statistically significant effect; conversely, for the effect of maternal age (36-39), the confidence interval using the degrees-of-freedom adjusted standard errors suggests a statistically significant effect, while that constructed using the Mancl and DeRouen-type adjustment does not. Given the conflicting results, we recommend the confidence intervals using the Mancl and DeRouen-type correction to the variance estimator based on the simulation study results. The simulations, in particular Simulation 2/K_s=15, suggest that sometimes $\widehat{Var}_{DF}[\hat{\beta}^c_w] > \widehat{Var}_{MD}[\hat{\beta}^c_w]$ while at other times $\widehat{Var}_{DF}[\hat{\beta}^c_w] < \widehat{Var}_{MD}[\hat{\beta}^c_w]$, but that in general, $\widehat{Var}_{MD}[\hat{\beta}^c_w]$ yields the closest to nominal coverage.

Table 1.4: Results from the analysis of the Rwandan birth dataset. See Section 1.7.1 for details.

	Point estimate			95% CI for $\widehat{OR}^{c}_{w}$		
	$\widehat{OR}$	$\widehat{OR}_w$	$\widehat{OR}^{c}_{w}$	$\widehat{Var}[\widehat{\beta}^{c}_{w}]$	$\widehat{Var}_{DF}[\widehat{\beta}^{c}_{w}]$	$\widehat{Var}_{MD}[\widehat{\beta}^{c}_{w}]$
Maternal age						
<20 years	1.15	1.11	1.12	(0.78, 1.60)	(0.62, 2.02)	(0.78, 1.61)
36-49 years	1.90	2.08	2.01	(1.40, 2.88)	(1.11, 3.63)	(0.97, 4.18)
Maternal weight						
$56-59$ kg	0.70	0.60	0.59	(0.36, 0.97)	(0.26, 1.34)	(0.26, 1.34)
$60-64$ kg	0.32	0.26	0.26	(0.14, 0.49)	(0.09, 0.74)	(0.10, 0.71)
≥ 65 kg	0.29	0.23	0.24	(0.11, 0.51)	(0.07, 0.84)	(0.08, 0.74)
Birth order						
2 - 3	0.42	0.44	0.49	(0.32, 0.75)	(0.25, 0.99)	(0.33, 0.73)
4+	0.15	0.11	0.13	(0.08, 0.22)	(0.05, 0.31)	(0.05, 0.36)
Female newborn	1.40	1.68	1.72	(1.07, 2.75)	(0.73, 3.74)	(0.93, 3.19)
Previous abortion	1.95	2.25	2.17	(1.32, 3.57)	(0.96, 4.93)	(0.97, 4.87)
Previous stillbirth	1.98	2.51	2.31	(1.18, 4.52)	(0.76, 6.98)	(1.05, 5.09)
Rulindo District	0.51	0.28	0.26	(0.08, 0.92)	(0.03, 2.08)	(0.05, 1.38)

1.8 Discussion

In this paper, we consider cluster-based ODS in resource-limited settings. Within this context, we extend the work of Cai et. al (2001) [12] by formally establishing the asymptotic properties of the IPW-GEE estimator, and provide an explicit expression for the asymptotic variance. In addition, motivated by our own study in Rwanda, we propose several small-sample bias corrections to both the point estimates and estimates of the asymptotic variance. Our simulation results suggest that there is no clear overarching story in terms of one approach consistently outperforming the others.

With regard to the point estimates, our simulations indicate that the bias-corrected point estimates generally reduce the bias in the point estimates, though the improvement is small in most cases. The impact of the bias-corrected standard errors, on the other hand, is much more substantial. The unadjusted standard errors result in severe undercoverage when K_s is small, and all of the bias-corrected standard errors yield closer to nominal coverage. Among the bias-corrections to the

variance estimator that we considered, the Mancl and DeRouen-type correction generally yielded the closest to nominal coverage when the normal distribution was used for confidence interval construction. When the t_{K_s-p} distribution was used for confidence interval construction, the Mancl and DeRouen-type correction could be conservative, particularly when the degree of correlation was increased. While the use of the t_{K_s-p} distribution also improved the coverage of the other bias-corrections, there was no method that consistently performed better than the others, and these methods still suffered from some undercoverage and overcoverage, depending on the design and parameter. Other degrees of freedom have been proposed when using the t distribution for confidence interval construction, and may yield improved coverage in this setting as well. Furthermore, once K_s was at least 50, the standard errors taking into account the covariance of the selection indicators and the standard errors ignoring the covariance of the selection indicators, generally resulted in similar coverage, indicating that asympototically, even the naïve method does not perform so poorly. This result held for both the unadjusted and the adjusted standard errors. When K_s was small, however, the narrative was mixed.

For these reasons, while acknowledging that there is not one superior method, and that the decision regarding which approach to take is in the hands of the researcher, we offer our perspective on how we would proceed: bias-corrected point estimates, and confidence intervals constructed using the normal distribution together with the variance taking into account the covariance of the selection indicators, adjusted with the Mancl and DeRouen-type correction. We stress the importance of using a bias-correction to the variance estimator, while using uncorrected (but weighted) point estimates will, based on our results, generally not have a substantial impact on the analysis. The observation that one approach does not consistently outperform all of the others is consistent with results from papers comparing different small-sample bias corrections to the robust sandwich estimator in the complete data setting - no single small-sample correction consistently outperforms other small-sample adjustments. Which correction performs better depends on a variety of factors, including

the number of clusters, the size of the clusters, the degree of variability in the cluster sizes, the degree of correlation among the clusters, and the type (cluster-level vs. individual-level) and distribution of the covariates in the model of interest. The theory that is presented relies on asymptotics, though in small samples, we are not aware of theory establishing small-sample behavior. Researchers must make decisions on what to do based on the characteristics of their data, and the objectives of their analysis. Finally, the simulations in this paper solely consider binary outcomes; more research is needed to determine the performance of these methods for continuous or count outcomes.

The focus of this paper is how to carry out estimation and inference for a dataset collected through a cluster-based outcome-dependent sampling design, and on considerations that a researcher must make when the number of clusters in the sub-sample is small. An open question, however, is how to identify an optimal sampling design for selecting the clusters at the design stage. As explained in Section 1.2.1 ,the specific design we adopted in Rwanda arose through efforts to balance financial/logistic considerations with statistical power/efficiency, with the latter assumed to be driven, in part at least, by the prevalence of the outcome and the sample size. While this informal strategy may intuitively be reasonable, how to make optimal (or at the very least wise) choices regarding a stratification scheme or a choice of model for π_k in this context is a topic of our on-going research. Another approach to increasing efficiency is to use pieces of information known at the design stage to calibrate the weights used in the weighted estimating equations (Breidt and Opsomer (2017)[8], Rivera-Rodriguez et. al (2019)[57]); this, too, is an area for future research.

Acknowledgements

Ms. Sauer was funded by the Harvard T.H. Chan School of Public Health NIEHS-sponsored Environmental Training Grant T32ES007142, and the Rose Traveling Fellowship Program in Chronic Disease Epidemiology and Biostatistics. Dr. Haneuse was supported by NIH grant R01

HL094786. Dr. Rivera-Rodriguez was supported by University of Auckland-Science/FRDF New Staff - 3716994. The Rwanda example included data collected as part of the All Babies Count program funded by Grand Challenges Canada Saving Lives at Birth, implemented by Partners In Health/Inshuti Mu Buzima and the Ministry of Health. The authors would like to thank Alphonse Nshimyiryo, Catherine Kirk, Ibrahim Hakizimana, and Robert Mutsinzi for their support in coordination and implementation of data collection in Rwanda.

2

Optimal allocation in stratified cluster-based outcome-dependent sampling designs

Sara Sauer[1], Bethany Hedt-Gauthier[1,2], Sebastien Haneuse[1]

1 *Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA USA*

2 *Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA USA*

Abstract

In public health research, finite resources often require that decisions be made at the study design stage regarding which individuals to sample for detailed data collection. At the same time, when study units are naturally clustered, as patients are in clinics, it may be preferable to sample clusters rather than the study units, especially when the costs associated with travel between clusters is high. In this setting, aggregated data on the outcome and select covariates is sometimes routinely-available, through, for example, a country's Health Management Information System (HMIS). If used wisely, this information can be used to guide decisions regarding which clusters to sample, and potentially obtain gains in efficiency over simple random sampling. In this paper we derive a series of formulae for optimal allocation of resources when a single-stage stratified cluster-based outcome-dependent sampling design is to be used and a marginal mean model is specified to answer the question of interest. Specifically, we consider two settings: (i) when a particular parameter in the mean model is of primary interest; and, (ii) when multiple parameters are of interest. We investigate the finite population performance of the optimal allocation framework through a comprehensive simulation study. Our results show that there are trade-offs that must be considered at the design stage: optimizing for one parameter yields efficiency gains over balanced and simple random sampling, while resulting in losses for the other parameters in the model. Optimizing for all parameters simultaneously yields smaller gains in efficiency, but mitigates the losses for the other parameters in the model.

2.1 Introduction

In public health research, finite resources, particularly in low-and-middle-income countries (LMICs), often require decisions to be made regarding which individuals are to be sampled for detailed data collection. For settings where individual study units in the population of interest are naturally clustered, as patients are in clinics or residents within districts, it may be preferred to sample clusters (as opposed to the individuals directly) especially if the cost associated with travel between clusters is high compared to the cost of sampling individuals within a cluster. Towards this, single-stage cluster-based sampling proceeds by first identifying a sub-sample of clusters and then performing detailed data collection on all individuals within those clusters.[59] When designing such a study, the key decision is that of which *clusters* to select. One straightforward approach is to take a simple random sample of the clusters. Doing so, however, will likely result in inefficient estimation if the outcome and/or the exposure of interest is rare. Furthermore, if some information is available at the design stage, either aggregate or individual-level, using simple random sampling forgoes the potential benefits associated with incorporating that information in the design. For instance, the Health Management Information System (HMIS), which has been implemented in a growing number of LMICs,[1,2] stores routinely-collected group or cluster-level summaries on a range of health indicators, such as the proportion of low birthweight births in health centers or the proportion of women who had 4 standard antenatal care (ANC) visits. While this aggregated data has primarily been used for the monitoring and evaluation of health system performance, as well as public health decision-making,[3,49,77] it may be cost-efficient to use relevant pieces of this group-level information to guide the sampling of the clusters at the design stage of a study.

The optimal design of experiments has been widely studied, and has been extended to the setting with correlated responses.[4,27,79] That work, however, has focused on settings that differ from

the ones we consider, specifically in that they assume complete control over the values of the covariate vectors that are included in the design, and how these covariate vectors are grouped into the K clusters. Such a setting is appropriate in the design of experiments in science and engineering in which the researcher has control over the experimental units, a level of control that public health researchers studying clusters of individuals do not have. Optimal design has also been extensively studied in the survey sampling context, though much of this work was originally limited to the estimation of simple quantities such as the mean or total of the measure of interest.[6,59] More recently, the ideas of optimal design have been extended to the regression context. For example, Zaslavsky et. al (2008)[82] propose a framework for optimal sampling of individuals when it is of interest to estimate the parameters in a linear regression model. Tao et. al (2019)[70] develop optimal two-phase designs for nonparametric maximum likelihood estimators, Han et. al (2020)[21] propose optimal sampling designs for survival models using the mean score estimator, and Zhong and Cook (2020)[83] develop optimal selection models in the context of family studies. McIsaac and Cook (2014)[41], building on the work of Reilly (1996)[55] develop optimal two-phase designs in the independent data setting with Bernoulli sampling, for analyses using maximum likelihood, mean score, inverse-probability weighted, and augmented inverse-probability weighted estimating equations. Finally, McIsaac and Cook (2013)[40] evaluate a proposed framework for optimal allocation in the clustered data setting where: the clusters are small and equally sized; selection is via Bernoulli sampling, so that the number of selected clusters is random; and, where primary interest lies in a cluster-level covariate. As one reviewer pointed out, in many cases, the quantity of interest, for example the parameter estimate in a regression model, can be expressed as the total of its influence functions,[11] so that these recent extensions of optimal allocation procedures are in fact related to the classic problems in survey sampling that were originally developed.

In this paper we consider the class of *single-stage stratified cluster-based outcome-dependent sampling* (ODS) designs, in which cluster-level information on the outcome (and possibly other covari-

ates) is taken to be readily-available. Furthermore, we suppose that within each stratum, clusters are sampled via simple random sampling, and that resources allow only a fixed number of clusters to be sampled. Depending on the nature of the available data, designs in this class can be operationalized in a number of ways. For example, Cai et al (2001)[12] proposed the *cluster-based* case-control design, in which clusters are sampled from strata defined by some threshold of the outcome prevalence of the clusters. They showed that such a design has the potential to yield substantial efficiency gains relative to a strategy based on taking a simple random sample of clusters, particularly when the outcome of interest is rare. Additionally, research on study design in the longitudinal data setting in which an expensive or difficult to measure exposure can only be ascertained for a subsample of subjects, showed that sampling from strata defined by the individuals' outcomes yields efficiency gains.[61,62,63,64,67] More generally, efficiency gains under ODS designs depend on various aspects of the specific design, such as what pieces of information are used to stratify the clusters, the degree of variability in the outcomes and covariates across the clusters, and the allocation of the sample size across the strata.

While the aforementioned research demonstrates that leveraging information that is available at the design stage through thoughtful stratification of the clusters has the potential for efficiency gains, it does not address how one can optimally allocate finite resources across strata. In this paper we address this gap. In doing so, we also note that the aforementioned has focused on settings where the cluster sizes are small (as is the norm in family studies or longitudinal studies). In contrast, motivated by a study of birth outcomes in Zanzibar, Tanzania, this paper considers settings where the cluster sizes are medium-to-large, a distinction we make for two reasons. First, it is not clear that the efficiency gains observed from efficient sampling designs in the setting with small cluster sizes would hold in the setting with larger cluster sizes, in particular due to greater heterogeneity of the subjects within larger clusters. Second, with larger cluster sizes, using the outcome to define the strata would require discretizing the cluster-level summaries of the outcome (e.g. the cluster-specific prevalences

or counts of the outcome), rather than using the individual-level values of the outcome directly in defining the strata, which we assume to be unknown at the design stage.

Whether a design is optimal depends on: 1) the optimality criterion; 2) what the target parameters are; and, 3) the method of analysis. In this paper, we assume that interest is in estimating one or more particular associations in a marginal mean model and that the data collected through the cluster-stratified outcome-dependent sampling scheme will be analyzed using inverse-probability weighted generalized estimating equations. Within this framing, this paper extends previous research in that it: (i) considers the potential for efficiency gains for individual-level as well as cluster-level covariates; and, (ii) considers settings in which there are multiple covariates of interest. As will become clear, the optimal allocation formulae we derive depend on quantities that, at the outset, will be unknown, including the parameters of the target marginal model. Thus, our primary goal in this paper is to develop a comprehensive understanding of the potential value (in terms of efficiency) of pursuing an optimal allocation strategy. That is, we seek to provide insight into how the potential for efficiency gain is impacted by different factors such as the optimality criterion and the relationship between the covariate of interest and the stratification variable. In the Discussion we return to this, and speak to how we believe this will spur creative solutions to how optimal designs can be operationalized in practice.

The remainder of this paper is organized as follows: In Section 2, we describe a hypothetical study that illustrates the objective of this work. In Sections 3 and 4 we give an overview of the methods. In Section 5, we present a comprehensive simulation study that examines the finite population operating characteristics of the optimal allocation sampling strategy proposed in this paper. Section 6 describes the results and Section 7 provides a discussion and concluding remarks.

2.2 Increasing Facility-Based Births in Zanzibar

2.2.1 Safer Deliveries Zanzibar

Safer Deliveries is a program in Zanzibar, Tanzania, implemented in collaboration between the local Ministry of Health and D-tree International, with the goal to reduce the maternal mortality rate in Zanzibar by increasing the number of health facility deliveries. [18] As of May 31, 2017, 42,056 women had been enrolled in the program. For these women, demographic and health information, obstetric history, the number of antenatal visits, and the location of delivery was collected by program-supported community health workers (CHWs). In addition, each woman's shehia of residence was collected; shehias, of which there are K=280 represented in the data set, are the lowest official administrative units in Zanzibar.

2.2.2 A Hypothetical Study

Given the aim of the program, researchers may be interested in investigating factors that may be associated with a woman delivering outside of a health facility, the prevalence of which in the available data is approximately 0.25. Here we consider a hypothetical study, one for which the following marginal mean model is of interest:

$$logit(P(Y_{ki} = 1)) = \beta_0 + \beta_1 X_{loc,k} + \beta_2 X_{ANC,ki} + \boldsymbol{\beta}_A^T \boldsymbol{X}_{A,ki}. \tag{2.1}$$

In model (2.1), Y_{ki} denotes the binary outcome of delivery outside of a health facility (1/0=Yes/No), $X_{loc,k}$ is a binary cluster-level variable indicating which island the shehia of residence is located in (1/0=Pemba/Unguja) and $X_{ANC,ki}$ indicates whether or not the woman had 4 standard ANC visits. Additionally, the model includes $\boldsymbol{X}_{A,ki}$, a collection of individual-level, woman-specific covariates

such as the previous location of delivery (if the woman has previously delivered) and whether the woman received a visit from a CHW at 8-9 months of pregnancy.

2.2.3 Potential for Efficiency Gains

Although complete data is available on the women enrolled in the Safer Deliveries program, for the purposes of this paper we consider the hypothetical situation in which this is not the case. In this setting, as motivated in the Introduction, one can imagine resource constraints limiting the number of shehias that can be visited for data collection. While the remainder of the paper addresses how to optimally select shehias, a natural first question to ask is whether there is, in fact, any potential for efficiency gain. Moreover, while it is well-known that wise selection of individuals in standard ODS designs can result in substantial efficiency gains over simple random sampling, it is less clear whether selecting clusters in a wise manner can yield efficiency gains over simple random sampling of the clusters. This is particularly the case when the analysis takes place at the individual level.

To investigate this, and further motivating the remainder of the paper, we conducted a simulation study, the details of which can be found in Appendix B.1. Briefly, we generated 1000 complete datasets with 28,789 women (the number of women in the data set who had given birth, with complete data on all covariates in the model of interest) from K=280 shehias according to (2.1). We then defined 500 'designs', where each design is a set of K_s=40 shehia ids that was obtained via simple random sampling. For each of the 1000 generated complete datasets, we took 500 samples corresponding to the 500 designs; for each of the resulting 500 samples, we computed the point estimates according to (2.1). Figure 2.1 shows a histogram of the standard errors (i.e. the standard deviation of the 1000 point estimates) of $\widehat{\beta}_{ANC}$ across the 500 designs; each value represented in the histogram of Figure 2.1 corresponds to a particular design/set of cluster ids. From the variation in the standard errors it is clear that some designs are substantially more efficient than others: the standard deviation of $\widehat{\beta}_{ANC}$ under the most efficient design is 0.078, while it is 0.129, under the least efficient design.

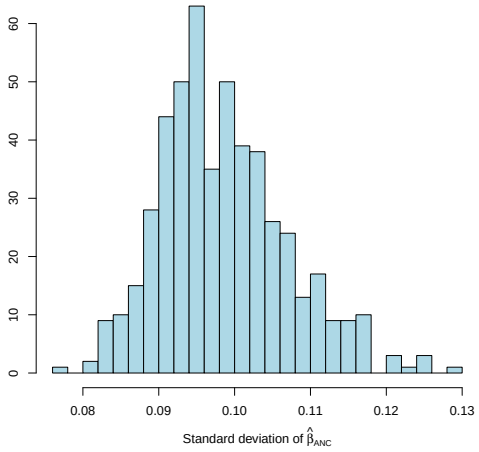

Figure 2.1: Distribution of the standard errors of $\widehat{\beta}_{ANC}$ from the 500 simple random sampling designs across 1000 iterations.

2.2.4 The Inadequacies of Rules of Thumb

In thinking about developing guidance for the selection of an efficient design, it might be natural to posit a rule of thumb that recommends sampling clusters with 1) a higher number of outcome cases, 2) a higher number of exposure cases, and 3) a higher overall individual-level sample size n. To investigate whether such a general rule of thumb is valid, we looked at the characteristics of the 'best' design for the estimation of β_{ANC}, which we defined as follows: we restricted attention to the designs which yielded a mean point estimate of β_{ANC} within 5% of the gold standard, where the gold standard was taken to be the mean point estimate obtained from running the analysis on the 1000 generated complete data sets (N=28,789). Among this set of designs, we then defined the 'best' design to be that with the lowest standard deviation in the point estimates of β_{ANC}, and the 'worst' design to be that with the highest standard deviation in the point estimates of β_{ANC}.

Table B.1/Scenario 1 displays summary information on the best and worst design for the estimation of β_{ANC}. We see that the best design has a substantially larger overall sample size ($n = 4606$ vs.

Table 2.1: Characteristics of the 'best' (lowest standard deviation of $\widehat{\beta}_{ANC}$) and 'worst' (highest standard deviation of $\widehat{\beta}_{ANC}$) designs among the set of 'unbiased' designs for the estimation of β_{ANC}. In Scenario 1, the number of outcome cases is proportional to the cluster size, while in Scenario 2, the number of outcome cases is inversely proportional to the cluster size. See Section 2.2.4 for details.

	Design	n	X_{ANC}	Y
Scenario 1				
Lowest sd($\widehat{\beta}_{ANC}$)	#431	4606	1284	1337
Highest sd($\widehat{\beta}_{ANC}$)	#284	3160	672	773
Scenario 2				
Lowest sd($\widehat{\beta}_{ANC}$)	#156	3846	1130	368
Highest sd($\widehat{\beta}_{ANC}$)	# 316	4621	1234	211

$n = 3160$), a larger number of exposure cases (1284 vs. 672), and a larger mean number of outcome cases (1337 vs. 773). These results are intuitive, and abide by the rule of thumb, in that the best design has a higher average number of outcome cases, a higher number of X_{ANC} cases, and a higher overall individual-level sample size, n.

However, it may not necessarily be the case that a cluster simultaneously satisfies the three criteria listed above; it may be the case, for example, that the clusters with a higher number of outcome cases have a lower number of exposure cases. In the original data set-up we described, the number of outcome cases and the number of X_{ANC} cases is proportional to the shehia size. We therefore also consider a scenario in which the number of outcome cases in a cluster is *inversely* proportional to cluster size. This was done by introducing into the data generation model an indicator for cluster size, where the indicator is equal to 1 if the cluser size is greater than the median cluster size, and 0 otherwise, with a coefficient of -3.

Table B.1/Scenario 2 gives the characteristics of the best and worst design for the estimation of β_{ANC}, in the setting where the number of outcome cases in a cluster is inversely proportional to the cluster size. We see that in this scenario, the best (lowest standard deviation of $\widehat{\beta}_{ANC}$) design has a smaller overall sample size than the design with the highest standard deviation of $\widehat{\beta}_{ANC}$ ($n = 3846$

vs. $n = 4621$), as well as a lower number of exposure cases, but a higher average number of outcome cases: the rule of thumb breaks down. This, together with the findings from Section 2.2.3, suggests that there is practical (and not just theoretical) value in developing a formalized framework for the selection of shehias in such a way as to maximize efficiency for estimation of the covariate of interest.

2.2.5 Optimal Allocation

Within the framing of this paper, suppose that select shehia-level data is available for all of the K=280 shehias, specifically the shehia-specific count of the outcome and $X_{loc,k}$. One could proceed by stratifying the clusters according to, say, $Y^*_{0.80}$, the 80^{th} quantile of the number of women delivering outside of a health facility across shehias, and $X_{loc,k}$. Doing so yields the following 2×2 stratification of the K=280 shehias:

	$X_{loc,k} = 0$	$X_{loc,k} = 1$
$Y^*_k < Y^*_{0.80}$	134	87
$Y^*_k \geq Y^*_{0.80}$	21	38

How best to then use this information will depend on the particular research question; here, building on model (2.1) we illustrate the three settings considered in this paper. In the first, we suppose particular interest lies in estimating the association between the cluster-specific $X_{loc,k}$ and the outcome, adjusting for other covariates in the model. While complete information is available on $X_{loc,k}$ in our hypothetical study, woman-level data on the outcome and the other covariates is not, so that it would be necessary to perform additional data collection on at least some individuals in a subsample. In the second setting, we suppose that it is of particular interest to estimate the association between $X_{ANC,ki}$ and the outcome. Note, in the first of these settings information is available at the design stage on the covariate of interest, while in the second, information is only available on a covariate that may be related to the covariate of interest. In the final setting we consider, we assume it is of interest to estimate all parameters in (2.1) with precision.

For each of these settings, the key question is how many shehias should be sampled from each stratum in the above 2×2 stratification in order to minimize the asymptotic variance of the estimator of the parameter(s) of interest. In the next sections, we propose a framework for deriving the optimal stratum-specific sample sizes, when the analysis proceeds via inverse-probability-weighted generalized estimating equations.

2.3 Setting

2.3.1 Model of Interest

In this paper, generalizing beyond model (2.1), we suppose that the research question of interest concerns learning about the relationship between an outcome Y and a p-vector of covariates, X (which may include a 1.0 for the intercept), in a population where the study units are naturally clustered in some way. Let K denote the number of clusters and N_k the number of individuals in the k^{th} cluster. Furthermore, we assume that estimation and inference will be performed with respect to the following marginal mean model for the outcome of the t^{th} individual in the k^{th} cluster as a function of their covariates, X_{ki}:

$$\mu_{ki} = E[Y_{ki}|X_{ki}] = g^{-1}(X_{ki}^T\beta), \tag{2.2}$$

where $g(\cdot)$ is a user-chosen link function and β a p-vector of regression parameters.

2.3.2 Stratified Cluster-Based ODS

Assuming complete information is not available on all elements of (Y, X) for all N individuals in the K clusters, we suppose that a stratified cluster-based ODS design will be employed and that budgetary or logistical constraints limit the number of clusters that can be visited to $K_s < K$. Towards this, we suppose that summary measures of (Y, X) for the K clusters are readily-available, as may

be other variables/information that are readily available but not of direct relevance to the scientific question, denoted by Z. Notationally, we denote the information available for cluster k at the design stage by $\mathcal{D}_k^* = (N_k, \mathbf{Y}_k^*, \mathbf{X}_k^*, \mathbf{Z}_k^*)$, where $\mathbf{Y}_k^*$ is a cluster-level summary of the outcomes (i.e. across the N_k individuals in the k^{th} cluster), $\mathbf{X}_k^*$ is a cluster-level feature or a cluster-level summary of elements of X, and $\mathbf{Z}_k^*$ is a cluster-level summary of Z. For example, if the outcome is Y_{ki} in Section 2.2 then $\mathbf{Y}_k^*$ might be the number of women from the k^{th} shehia who did not deliver in a health facility. Furthermore, $\mathbf{X}_k^*$ may be a cluster-level feature, such as $X_{loc,k}$ in Section 2.2 or it may be an aggregated summary of individual-level data, such as the proportion of mothers in the shehia who had 4 standard ANC visits. Finally, $\mathbf{Z}_k^*$ may, for example, be the prevalence of some other outcome such as the birth outcome.

Let R_k be a binary indicator of whether the k^{th} cluster is selected by the design and $\pi_k = \Pr(R_k = 1|\mathcal{D}^*)$ the corresponding probability of being selected, where $\mathcal{D}^* = \{\mathcal{D}_1^*, \ldots, \mathcal{D}_K^*\}$ is the totality of the information available at the design stage. In a cluster-stratified ODS design, the clusters are cross-classified according to some variable S that is defined on the basis of one or more of the variables contained in $\mathcal{D}^*$ and is assumed to take on one of J levels. From this, suppose K_j clusters are classified as belonging to the j^{th} stratum. We assume that the stratification scheme is specified such that $K_j > 0$ for all $j = 1, \ldots, J$. The design then proceeds by randomly selecting $k_j \leq K_j$ clusters from those in the j^{th} stratum, such that $\sum_{j=1}^{J} k_j = K_s$. Note, because of the random sampling, we have that $\pi_k = k_j/K_j$ for each cluster in the j^{th} stratum. Finally, the otherwise unavailable elements of (Y, X) are ascertained on all individuals within the sampled clusters.

2.3.3 Estimation and Inference

When information is only available on a subset of clusters that have been selected through a cluster-based ODS scheme, β can be estimated as the solution to the following weighted generalized esti-

mating equations[12]:

$$\mathcal{U}_w(\beta) \;=\; \sum_{k=1}^{K} R_k \pi_k^{-1} \mathbf{D}_k^T \mathbf{V}_k^{-1} (\mathbf{Y}_k - \mu_k) \;=\; 0. \tag{2.3}$$

This can be rewritten as $\mathcal{U}_w(\beta) \;=\; \mathbf{U}^T \mathbf{W} \mathbf{R}$, where $\mathbf{U} \;=\; diag\{\mathbf{Y} - \mu\}\mathbf{V}^{-1}\mathbf{D}$ is an $N \times p$ matrix where $\mathbf{Y} \;=\; (\mathbf{Y}_1, \ldots, \mathbf{Y}_K)^T$, $\mu \;=\; (\mu_1, \ldots, \mu_K)^T$, $\mathbf{V}$ is an $N \times N$ block-diagonal matrix, with the $\mathbf{V}_k$ on the diagonal, and $\mathbf{D}$ is the $N \times p$ matrix obtained by stacking the K $\mathbf{D}_k \;=\; \partial \mu_k / \partial \beta$ matrices. Furthermore, $\mathbf{W}$ is an $N \times N$ diagonal matrix with diagonal entries equal to the vector $(\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_K)^T$, with $\mathbf{W}_k$ a vector of length N_k with each element equal to π_k^{-1}. Letting $\mathbf{R}_k$ denote the $N_k \times 1$ vector with all entries equal to R_k, $\mathbf{R}$ is the $N \times 1$ vector obtained by concatenating the $\mathbf{R}_k$ together. Following arguments similar to those presented by Xie and Yang (2003)[81] in the complete data setting, Sauer et. al (2021)[60] showed that under regularity conditions, $\widehat{\beta}_w$, the solution to (2.3), is consistent for β_0, the true value of β, and is asymptotically multivariate normal, with the asymptotic variance given by

$$Var[\widehat{\beta}_w] = \mathbf{H}(\beta_0)^{-1} \left\{ Var[\mathcal{U}_w(\beta)]|_{\beta=\beta_0} \right\} \mathbf{H}(\beta_0)^{-1}, \tag{2.4}$$

where $\mathbf{H}(\beta) = -E[\partial \mathcal{U}(\beta)/\partial \beta]$, and $\mathcal{U}(\beta) = \mathbf{U}^T 1_{N \times 1}$.

2.4 Optimal Allocation

In the event that the research team has the opportunity to inform the design of the study (i.e the data has not already been collected), the simulation in Section 2.2.3 highlighted the potential efficiency gains associated with a wise choice of which clusters to select. Towards that, suppose primary scientific interest lies in estimating a treatment or intervention effect. Then the optimal design will be the one that maximizes the precision of the estimate of that parameter. In other settings,

there may be multiple covariates of interest, as may be the case in situations in which the research question concerns identifying risk factors for a health outcome. In this section, we first derive the optimal allocation formula for the setting in which there is just one parameter of interest. We then extend this to address the situation in which there are multiple parameters of interest, in particular the situation in which all parameters are of interest, which involves minimizing the trace of the variance-covariance matrix.

2.4.1 One Parameter of Interest

In Section 2.2.5, we described two scenarios in which interest was taken to be in investigating the association between the outcome of whether a woman delivered outside of health facility and which island the woman's shehia of residence is on (a cluster-level covariate) and whether the woman had 4 standard ANC vistis (a woman-specific covariate). More generally, suppose that primary interest lies in estimating β_q, one specific element of $\boldsymbol{\beta}$ in model (2.2). Under a cluster-stratified design, the optimal allocation for the estimation of β_q involves determining the $k_j, j = 1, \ldots, J$ such that the variance of $\widehat{\beta}_q$ is minimized, subject to the constraint that $\sum_{j=1}^{J} k_j = K_s$. In determining the optimal sample sizes for each stratum, it is useful to rewrite (2.4) as:

$$Var[\widehat{\boldsymbol{\beta}}_w] = \mathbf{H}(\boldsymbol{\beta}_0)^{-1} \left\{ Var[\mathbf{U}^T 1_{N\times 1}] + E[\mathbf{U}^T \mathbf{W} \boldsymbol{\Delta} \mathbf{W} \mathbf{U}] \right\} \mathbf{H}(\boldsymbol{\beta}_0)^{-1}.$$

where $\boldsymbol{\Delta} = Var[\mathbf{R}|\mathcal{F}_K]$ is an $N \times N$ matrix with the entries on the diagonal equal to $Var[R_k|\mathcal{F}_K]$ = $\pi_k - \pi_k^2 = \frac{k_j}{K_j} - (\frac{k_j}{K_j})^2$ for all N_k individuals in the k^{th} cluster and $\mathcal{F}_K = \{\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{S}\}$. On the off-diagonal, the entries are equal to $Cov[R_k, R_{k'}|\mathcal{F}_K] = \pi_{kk'} - \pi_k \pi_{k'}$, with $\pi_{kk'}$ equal to the joint probability that clusters k and k' are selected by the outcome-dependent sampling scheme. If clusters k and k' are not in the same stratum, then $\pi_{kk'} = \pi_k \pi_{k'}$, and $Cov[R_k, R_{k'}|\mathcal{F}_K]$=0, while if clusters k and k' are in the same stratum, $\pi_{kk'} = \frac{k_j}{K_j} \frac{k_j - 1}{K_j - 1}$, and $Cov[R_k, R_{k'}|\mathcal{F}_K] = \frac{k_j}{K_j} \frac{k_j - 1}{K_j - 1} - (\frac{k_j}{K_j})^2$. In deter-

mining the optimal allocation across strata, it suffices to determine the allocation that minimizes the $[q,q]^{th}$ entry of $E[\mathbf{H}^{-1}\mathbf{U}^T\mathbf{W}\boldsymbol{\Delta}\mathbf{W}\mathbf{U}\mathbf{H}^{-1}]$, the term which depends on the sampling of clusters, and is given by:

$$E[\mathbf{H}^{-1}\mathbf{U}^T\mathbf{W}\boldsymbol{\Delta}\mathbf{W}\mathbf{U}\mathbf{H}^{-1}]_{[q,q]} = \sum_{j=1}^{J} \frac{K_j - k_j}{k_j}[A_{q,j} - \frac{B_{q,j}}{K_j - 1}] = \sum_{j=1}^{J} \frac{K_j - k_j}{k_j} C_{q,j}$$

where $A_{q,j} = \sum_{k \in S_j} \sum_{i=1}^{N_k} \sum_{i'=1}^{N_k} E[h_{ki}^{[q]} h_{ki'}^{[q]}]$, $B_{q,j} = \sum_{k \in S_j} \sum_{k' \neq k \in S_j} \sum_{i=1}^{N_k} \sum_{i'=1}^{N_{k'}} E[h_{ki}^{[q]} h_{k'i'}^{[q]}]$, $h_{ki}^{[q]}$ is the entry in the $(q+1)^{th}$ column of $\mathbf{U}\mathbf{H}^{-1}$ corresponding to the i^{th} individual in the k^{th} cluster, and $S_j = \{k : \text{cluster } k \in \text{stratum } j\}$. The optimization problem involves minimizing $f_q(k_1, k_2, \ldots, k_J) = \sum_{j=1}^{J} \frac{K_j - k_j}{k_j} C_{q,j}$ subject to the constraint that $\sum_{j=1}^{J} k_j = K_s$, and can be solved using the method of Lagrange multipliers, which yields:

$$k_j = K_s \frac{(K_j^{1/2} C_{q,j}^{1/2})}{\sum_{j=1}^{J} K_j^{1/2} C_{q,j}^{1/2}} \tag{2.5}$$

From expression (2.5), the optimal sample size for stratum j, k_j, therefore depends on the interplay between K_j, the number of clusters in stratum j, and the value of $C_{q,j}$, the contribution of the elements in stratum j to the variance of $\widehat{\beta}_q$, relative to the totality of these two quantities across all of the strata.

2.4.2 Multiple Parameters of Interest

In certain situations, researchers may be interested in estimating more than one parameter with precision. Let w_q be a $p \times 1$ vector of weights, with the q^{th} element denoting the weight assigned to the q^{th} parameter. Of interest is to minimize $f(k_1, k_2, \ldots, k_J) = \sum_{q=1}^{p} w_q \sum_{j=1}^{J} \frac{(K_j - k_j)}{k_j} C_{q,j}$ subject to the constraint that $\sum_{j=1}^{J} k_j = K_s$. For example, one may be equally interested in estimating every parameter in the model with precision. Towards this, one approach would be to minimize the trace of the variance-covariance matrix of the parameter estimates, which corresponds to setting all of

the w_q to 1, and is known as A-optimality[4]. Again, the optimal stratum-specific sample sizes can be solved using the method of Lagrange multipliers:

$$k_j = K_s \frac{K_j^{1/2}(\sum_{q=1}^{p} w_q C_{q,j})^{1/2}}{\sum_{j=1}^{J} K_j^{1/2}(\sum_{q=1}^{p} w_q C_{q,j})^{1/2}}. \tag{2.6}$$

In both expressions (2.5) and (2.6), the $C_{q,j}$ depend on the expectations $E[h_{ki}^{[q]} h_{ki'}^{[q]}]$ and $E[h_{ki}^{[q]} h_{k'i'}^{[q]}]$. These expectations are not known, and can be replaced by $h_{ki}^{[q]} h_{ki'}^{[q]}$ and $h_{ki}^{[q]} h_{k'i'}^{[q]}$ when computing the stratum-specific sample sizes.

2.4.3 Practical Issues

In conducting simulation studies to investigate the performance of the allocation schemes given by expressions (2.5) or (2.6), a number of practical issues/challenges arise that we summarize here. First, we note that the k_js computed using (2.5) or (2.6) may not be integers. They must therefore be rounded if they are to serve as meaningful sample sizes. In doing so, however, the rounded stratum-specific sample sizes may not add up to the constraint K_s. To resolve this, we introduce a small rounding threshold, τ, which can be increased until the sum of the rounded sample sizes is equal to K_s. For example, if K_s=40, $J = 4$, and (k_1, k_2, k_3, k_4)=(20.18, 7.01, 6.49, 6.32), using the standard rounding threshold of 0.5 would yield $(k_1^{r_{0.50}}, k_2^{r_{0.50}}, k_3^{r_{0.50}}, k_4^{r_{0.50}})$=(20, 7, 6, 6), the sum of which is 39. We would therefore set the rounding threshold at $\tau = 0.001$ and increase the threshold by a small increment such as 0.0001 until $\sum_{j=1}^{J} k_j^{r_\tau} = 40$. In this example, this approach would yield a rounding threshold of $\tau = 0.3201$, which gives $(k_1^{r_\tau}, k_2^{r_\tau}, k_3^{r_\tau}, k_4^{r_\tau})$=(20, 7, 7, 6) and satisfies $\sum_{j=1}^{J} k_j^{r_\tau} = 40$. Other approaches have been suggested that directly yield integer solutions, such as the algorithm propose by Wright (2017),[80] and used with good results by Chen and Lumley (2020);[15] we do not consider those here, as our simulations indicate good performance of the continuous allocation strategy combined with the rounding procedure described above.

Second, it is possible for one or more of the resulting (rounded) k_j to equal zero or to be larger than the number of clusters in the stratum, K_j. We refer to these as 'edge cases'. When this occurs, we fix the stratum-specific sample size at the boundary (i.e. set $k_j = 1$ if $k_j = 0$, set $k_j = K_j$ if $k_j > K_j$), and recalculate the other k_j with an updated constraint, a strategy also used by Reilly (1996)[55] :

$$K_s^* = K_s - \sum_{j:k_j=0} 1 - \sum_{j:k_j>K_j} K_j.$$

Finally, we note that the calculation of $C_{q,j}$ in expressions (2.5) and (2.6) relies on having complete data on the outcome and all of the covariates, which is not the case in practice. As indicated in the Introduction, however, the primary goal of this paper is to establish and understand the potential for efficiency gains; doing so will, we believe, motivate creative solutions to implementing these optimal designs in practice.

2.5 Simulation Study

2.5.1 Model Specification

We conducted a simulation study to examine the finite sample performance of the optimal allocation strategies presented in Sections 2.4.1 and 2.4.2 for a single parameter (expression (2.5)) and for multiple parameters (expression (2.6)), respectively. In particular, we sought to evaluate the validity of the approach for estimation and inference as well as the performance of the optimal allocation designs relative to the simple random sampling and balanced stratified sampling designs in terms of efficiency. Throughout, we assume that interest lies in the following marginal mean model for the i^{th} individual in the k^{th} cluster:

$$logit(P(Y_{ki} = 1)) = \beta_0 + \beta_1 X_{1k} + \beta_2 X_{2k} + \beta_3 X_{3ki} + \beta_4 X_{4k} + \beta_5 X_{5ki} \tag{2.7}$$

Table 2.2: Covariate distributions for nine simulation scenarios considered in Section 2.5.

Covariate	Baseline Scenario	Eight scenarios in which X_{1k} and each of the remaining covariates are dependent	
		Positive dependence	Negative dependence
X_{1k}	$Ber(0.3)$	–	–
X_{2k}	$N(1, 0.25)$	$\mu = 1 + 0.5I_{(X_{1k}=1)}$ $\sigma = 0.25 + 0.75I_{(X_{1k}=1)}$	$\mu = 1 + 0.5I_{(X_{1k}=0)}$ $\sigma = 0.25 + 0.75I_{(X_{1k}=0)}$
X_{3ki}	$N(1, 0.25)$	$\mu = 1 + 0.5I_{(X_{1k}=1)}$ $\sigma = 0.25 + 0.75I_{(X_{1k}=1)}$	$\mu = 1 + 0.5I_{(X_{1k}=0)}$ $\sigma = 0.25 + 0.75I_{(X_{1k}=0)}$
X_{4k}	$Ber(p_k)$, $p_k = expit(-0.9)$	$p_k = 0.6I_{(X_{1k}=1))} +$ $0.156I_{(X_{1k}=0)}$	$p_k = 0.156I_{(X_{1k}=1)} +$ $0.346I_{(X_{1k}=0)}$
X_{5ki}	$Ber(p_k)$, $p_k = 0.25$	$p_k = 0.6I_{(X_{1k}=1))} +$ $0.1I_{(X_{1k}=0)}$	$p_k = 0.1I_{(X_{1k}=1))} +$ $0.314I_{(X_{1k}=0)}$

where $\beta_0 = (-3.1, 0.3, 0.7, 0.7, 0.7, 0.7)$. The value of the intercept was chosen so that the prevalence of the binary outcome in the baseline scenario (defined in the next paragraph) is about 0.23, close to the prevalence of the outcome of interest in the D-tree dataset.

As summarized in Table 2.2, we consider nine data scenarios. In the baseline scenario, X_{1k} is a binary cluster-level covariate with prevalence of 0.30, $X_{2k} \sim N(1, \sigma = 0.25)$ is a continuous cluster-level covariate, $X_{3ki} \sim N(1, \sigma = 0.25)$ is a continuous individual-level covariate, $X_{4k} \sim Ber(p_k)$ is a binary cluster-level covariate with p_k = expit(-0.9) and $X_{5ki} \sim Ber(p_k)$ is a binary individual-level covariate with $p_k = 0.25$. Note, in this scenario the covariates are all independent of each other. Building on this we consider eight additional scenarios where X_{1k} is taken to be positively and negatively associated with each of the other four covariates (individually).

2.5.2 Designs

For each data scenario, we generated 10000 complete datasets of K=280 clusters with equal cluster sizes of $N_k = 40 \,\forall\, k = 1, \ldots, K$. Correlation between the outcomes under the marginal model (2.7) was induced using the `GenBinaryY()` function in the `MMLB` package for R, which implements a method for marginally-specified logistic-Normal models[25]. Given a specification for a marginal mean model for the response, this method induces within-cluster correlation via cluster-specific random intercepts that are taken to arise from a N(0, σ_V^2) distribution. In our simulations, we set σ_V equal to 0.5.

For each generated complete dataset, the 280 clusters were stratified according to $Y^*_{0.80}$, which as in Section 2.2.5 is defined to be the 80^{th} quantile of the number of outcome cases across the K clusters, and X_1. Then, K_s=80 clusters were sampled from the four strata according to eight designs. The first two, which serve as current-practice comparators are: SRS, simple random sampling of the K_s clusters; and, BalX_1, stratified balanced sampling in which 20 clusters were sampled from each of the four $Y^*_{0.80} \times X_1$ stratum. The next five each consider optimal allocation with respect to one parameter using expression (2.5) proposed in Section 2.4.1: OptX_1, for the estimation of β_1; OptX_2, for the estimation of β_2; OptX_3, for the estimation of β_3; OptX_4, for the estimation of β_4; and, OptX_5, for the estimation of β_5. Finally, Opt$_A$ represents the optimal allocation based on expression (2.6) proposed in Section 2.4.2 when interest lies in estimating all parameters with precision.

The nine data scenarios combined with eight designs for each data scenario resulted in seventy-two simulations. Further simulations, presented in Appendix B.3 , looked at the impact of: (i) varying cluster sizes; (ii) increasing the degree of correlation; and, (iii) decreasing the number of sampled clusters K_s. In the setting of unequal cluster sizes, the distribution of the K clusters was taken to be the same as the distribution of the 280 cluster sizes in the D-tree dataset, adjusted so that the overall

population size N is about the same as the the equal cluster size scenarios described. In the setting of increased within cluster-correlation, we set $\sigma_V = 1$. In the setting of smaller K_s, 40 clusters were sampled under each design.

2.5.3 Analyses

For each sample taken, we computed the point estimates by solving expression (2.3) and estimated the standard errors using an estimator of (2.4), the form of which is given in Sauer et. al (2021).[60] Due to the potential downward bias of the robust sandwich variance estimator in finite samples, we computed four estimates of the standard errors: i) unadjusted standard errors, ii) degrees-of-freedom adjusted standard errors, in which the variance-covariance matrix is multplied by the factor $K_s/(K_s - p)$,[37,60] iii) a Mancl and DeRouen-type bias correction[38] that is adapted to accomodate the weights needed to account for the sampling design,[60] and iv) a Kauermann and Carroll-type[30] bias correction that is similarly adapted.[60] For each design, across the 10000 iterations, we computed the mean point estimates and the relative uncertainty of each design for the estimation of each parameter, defined as the ratio:

$$\frac{sd(\widehat{\beta}^{1:R}_{w,q})_{Design}}{sd(\widehat{\beta}^{1:R}_{w,q})_{SRS}} \; q = 1, \ldots, p$$

where $R = 10000$ is the number of simulation iterations. Furthermore, to check the validity of the inference, we estimated the 95% Wald coverage probabilities by computing the proportion of constructed confidence intervals that contain the true parameter value.

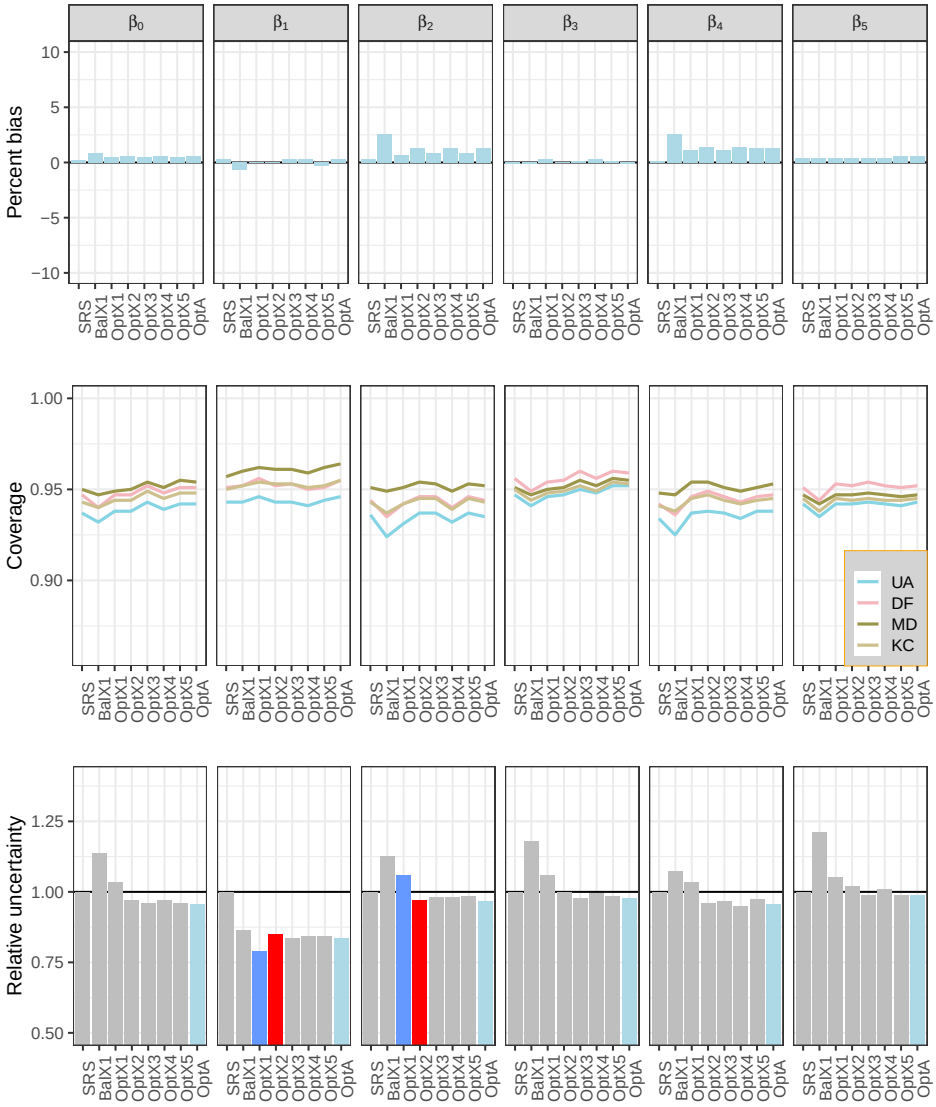

Figure 2.2: Baseline scenario: K=280, $N_k = 40 \; \forall \; k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. Shown is (i) the percent bias in the mean point estimates (top panel), (ii) the estimated coverage probabilities for confidence intervals (middle panel) using unadjusted estimated standard errors (UA), estimated standard errors with a degrees-of-freedom adjustment (DF), estimated standard errors with a Mancl and DeRouen-type correction (MD), and estimated standard errors with a Kauermann and Carroll-type correction (KC), and (iii) the uncertainty relative to simple random sampling (bottom panel) with the dark blue, red, and light blue bars corresponding to scenarios 1, 2, and 3 described in Section 2.2.5 respectively.

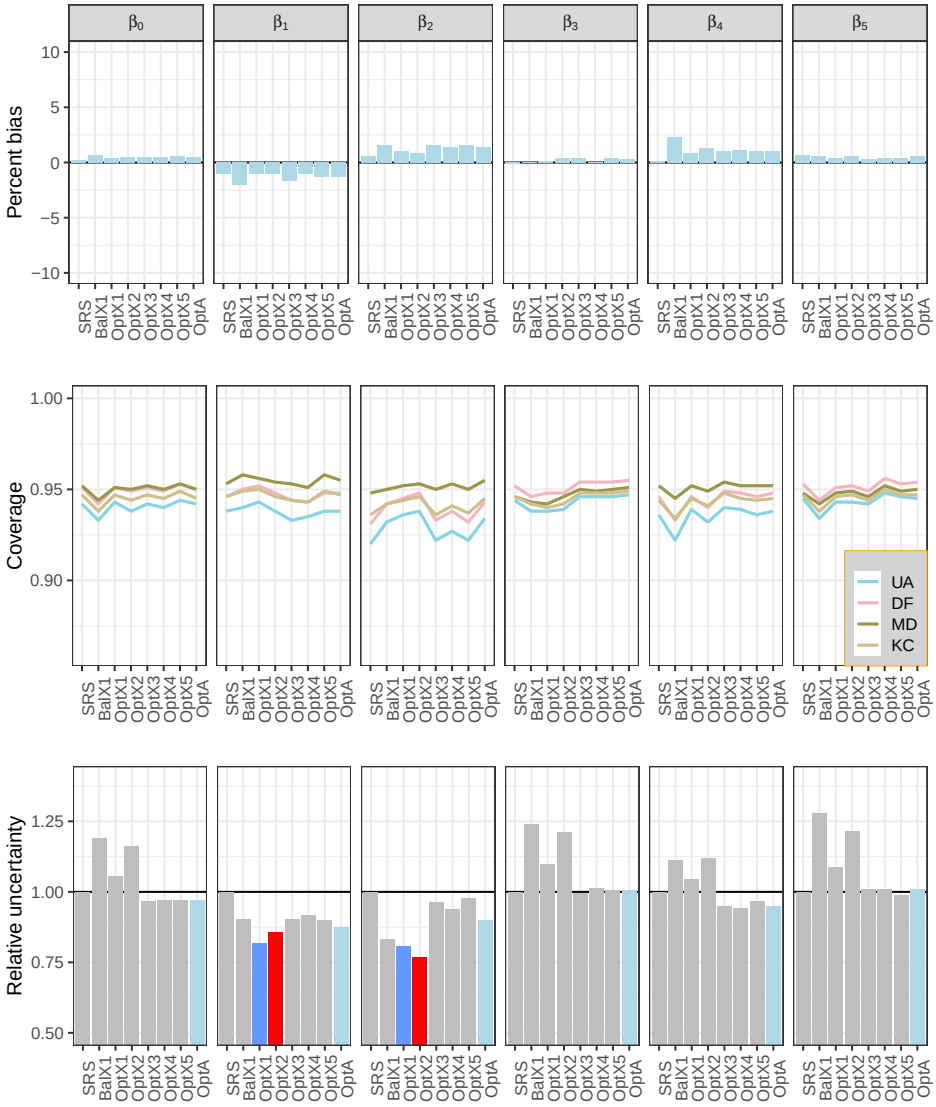

Figure 2.3: Positive association X_1 and X_2: K=280, $N_k = 40 \; \forall \; k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. Shown is (i) the percent bias in the mean point estimates (top panel), (ii) the estimated coverage probabilities for confidence intervals (middle panel) using unadjusted estimated standard errors (UA), estimated standard errors with a degrees-of-freedom adjustment (DF), estimated standard errors with a Mancl and DeRouen-type correction (MD), and estimated standard errors with a Kauermann and Carroll-type correction (KC), and (iii) the uncertainty relative to simple random sampling (bottom panel) with the dark blue, red, and light blue bars corresponding to scenarios 1, 2, and 3 described in Section 2.2.5 respectively.

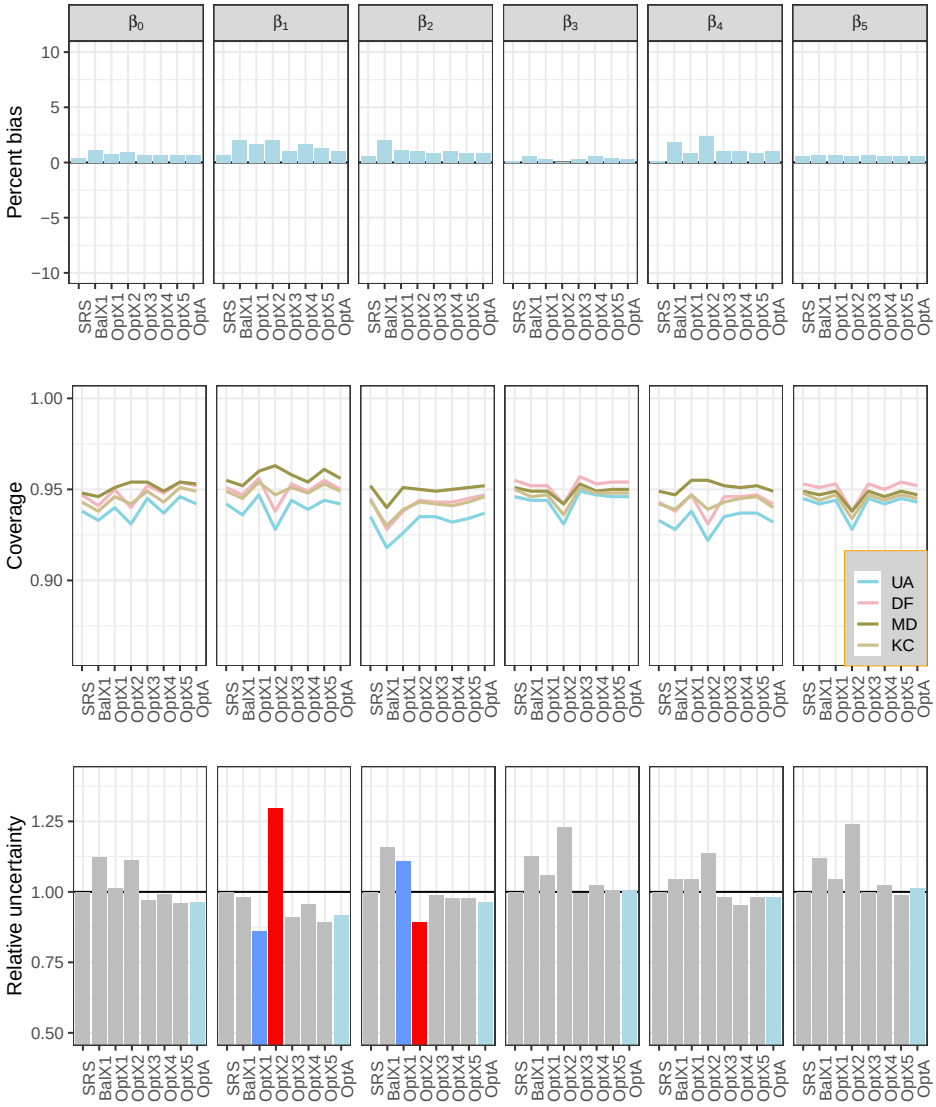

Figure 2.4: Negative association X_1 and X_2: K=280, $N_k = 40 \, \forall \, k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$. Shown is (i) the percent bias in the mean point estimates (top panel), (ii) the estimated coverage probabilities for confidence intervals (middle panel) using unadjusted estimated standard errors (UA), estimated standard errors with a degrees-of-freedom adjustment (DF), estimated standard errors with a Mancl and DeRouen-type correction (MD), and estimated standard errors with a Kauermann and Carroll-type correction (KC), and (iii) the uncertainty relative to simple random sampling (bottom panel) with the dark blue, red, and light blue bars corresponding to scenarios 1, 2, and 3 described in Section 2.2.5 respectively.

2.6 Results

2.6.1 Relative Uncertainty

Figures 2.2 - 2.4 show the results for the baseline data scenario (Figure 2.2), the setting in which there is a positive relationship between X_1 and X_2 (Figure 2.3), and the setting in which there is a negative relationship between X_1 and X_2 (Figure 2.4). Note, complete results for all scenarios considered can be found in Appendix C of the Supporting Information. Each figure shows: 1) percent bias in the mean point estimates; 2) estimated coverage probabilities; and, 3) relative uncertainty. Collectively, we find that there is little to no bias (< 5%) in the mean point estimates and the estimated coverage probabilities are near the nominal level, in particular when the Mancl and DeRouen-type standard error correction is used. The conclusions regarding relative uncertainty are more nuanced, however, with the key results being:

- In general, the optimal design for the parameter of interest estimates that parameter with the greatest precision. For example, across all scenarios, OptX_1 yields the highest efficiency gain for the estimation of β_1, outperforming even the balanced stratified design that stratifies on X_1 (BalX_1): in the baseline data scenario (Figure 2.2), for instance, the gain in efficiency compared to simple random sampling is 13.8% under BalX_1, while it is 21.1% under OptX_1.

- In the baseline scenario (Figure 2.2), in which there is no relationship between X_1 and any of the other covariates, we do not see substantial efficiency gain for any of the parameters under any of the optimal allocation schemes aside from β_1, which is to be expected.

- In the setting in which X_1 and the continuous cluster-level covariate X_2 are positively associated (Figure 2.3), the OptX_2 design yields an efficiency gain of 23.1% over simple random sampling in the estimation of β_2. That is, even though the stratification is based on X_1, efficiency gains regarding β_2 are obtained through the proposed optimal allocation scheme.

Furthermore, even though OptX_2 optimizes with respect to β_2, using this design when X_1 and X_2 are positively associated also yields an efficiency gain of 14.5% for β_1. This design, however, yields substantial efficiency losses for the other parameters, ranging from 11.8% to 21.5%. Finally, in this setting, the Opt$_A$ design results in a smaller efficiency gain for the estimation of β_1 and β_2 (12.7% and 10%, respectively), but also yields slight gains for β_0, β_1, and β_4, while mitigating the losses in efficiency for β_3 (0.6% loss) and β_5 (1.1% loss).

- When there is a negative dependence between X_1 and X_2, the optimal design still yields efficiency gains relative to simple random sampling, though the magnitude of the gain is attenuated (10.8% in the estimation of β_2). It is important to recognize, however, that while the efficiency gain is not as high under this setting, if β_2 is the sole parameter of interest, it appears to be even more important to optimize for this parameter, as other designs may lead to substantial losses in efficiency for the estimation of that parameter: for example, while both BalX_1 and OptX_1 also yield substantial efficiency gains for the estimation of β_2 when there is a positive relationship between X_1 and X_2 (Figure 2.3), these allocations result in efficiency losses of 15.7% and 10.8%, respectively, when the relationship between X_1 and X_2 is negative (Figure 2.4). The OptX_2 design yields substantial losses for all other parameters in this scenario in the range of 11.2% to 29.8%. Again, the Opt$_A$ design, while resulting in a smaller efficiency gain for β_2 (3.6%), only yields slight losses for β_3 and β_5 (0.6 - 1.1%).

- Although not presented here, we see similar results for OptX_3 in the settings where there is a relationship (positive or negative) between X_1 and the individual-level covariate X_3; see Appendix B.3. The story is also similar for the settings in which there is a relationship between X_1 and the binary cluster-level X_4 or the binary individual-level covariate X_5, though the efficiency gains and losses in these scenarios are not as large as those considering the continuous covariates X_2 and X_3: 7.9% and 6.4% efficiency gains in the estimation of β_4 under OptX_4 in

the positive and negative dependence settings, respectively, and 4.5% and 2.1% gains in the estimation of β_5 for OptX_5 in the positive and negative dependence settings, respectively. This may be due to the fact that the difference in the distributions of X_4/X_5 are not as different across levels of the stratification variable X_1 as are the distributions of X_2/X_3. Note, however, that even in the situations with small efficiency gain relative to simple random sampling under the OptX_4 (OptX_5) design for the estimation of β_4 (β_5), the loss in efficiency for the estimation of β_4 (β_5) under the BalX_1 can be substantial (1.7% and 10.9% loss compared to SRS for estimation of β_4 in the positive and negative dependence settings, respectively; 12.6% and 24.7% loss for the estimation of β_5 in the positive and negative dependence settings, respectively), thereby making the optimal allocation preferable, particularly if one is interested in estimating β_1 with precision without the loss of efficiency in the estimation of β_4 (β_5).

- Finally, while the balanced stratified design yields substantial efficiency gains for the estimation of β_1, the gain is generally not as much as that under OptX_1. Moreover, BalX_1 often results in substantial losses in the estimation of the other parameters relative to simple random sampling. The inconsistent performance of the balanced stratified design has been noted in other settings as well.[40] This suggests that the balanced stratified design, while commonly used due to the fact that it is simple to implement in practice, may not be a wise choice of design when interest lies in estimating parameters other than that corresponding to the stratification variable with precision.

2.6.2 Design Characteristics

Turning attention to the simulation study in which the cluster sizes vary, Table 2.3 shows the average stratum-specific, cluster-level sample sizes (i.e. the k_js), and the average overall individual-level sample size, n, across the 10000 iterations for the eight data scenarios in which there is a dependence

between X_1 and one of the other covariates in the model. Based on these summary characteristics, we make several observations:

- *The design that on average yields the largest overall individual-level sample size, n, is not necessarily the most efficient design for the estimation of the parameter of interest.* For example, under the data scenario in which there is a negative dependence between X_1 and X_2, the average sample size under the BalX_1 design is 4105, while the average sample size under OptX_1 is 3597, and that under OptX_2 is 4016. Even though the average sample size under BalX_1 is larger than that under OptX_1 (4105 vs. 3597), OptX_1 is more efficient for the estimation of β_1 ($\text{sd}(\widehat{\beta}^{1:R}_{1,OptX_1})/\text{sd}(\widehat{\beta}^{1:R}_{1,BalX_1}) \sim 0.968$). Similarly, even though the average sample size under BalX_1 is larger than that under OptX_2, OptX_2 is more efficient for the estimation of β_2 ($\text{sd}(\widehat{\beta}^{1:R}_{2,OptX_2})/\text{sd}(\widehat{\beta}^{1:R}_{2,BalX_1}) \sim 0.762$). Finally, even though the average sample size under OptX_2 is larger than that under OptX_1, OptX_1 is more efficient for the estimation of β_1 ($\text{sd}(\widehat{\beta}^{1:R}_{1,OptX_1})/\text{sd}(\widehat{\beta}^{1:R}_{1,OptX_2}) \sim 0.720$).

- *When the covariate of interest is continuous, more clusters are generally selected from the strata in which the variability of the covariate of interest is greater.* For example, in the setting in which there is a negative association between X_1 and X_2, more clusters are sampled from the strata with $X_1 = 0$ (on average, 72 clusters from the strata with $X_1 = 0$ vs. 8 clusters from the strata with $X_1 = 1$), due to the fact that there is more variation in the values of X_2 among clusters with $X_1 = 0$; in contrast, in the setting in which there is a positive association between X_1 and X_2, more clusters are sampled from the strata with $X_1 = 1$ (48 clusters on average vs. 33 clusters from the strata with $X_1 = 0$), because in this scenario there is more variation in the values of X_2 among clusters with $X_1 = 1$.

- *In the setting in which the values of the covariate of interest are homogeneous within strata (as is the case with the stratification variable X_1 in all scenarios), the k_js are generally distributed*

across the strata in such a way that results in a decrease in the variability of the weights, K_j/k_j, compared to balanced stratified sampling. For example, under the setting in which there is a negative relationship between X_1 and X_2, in comparing the weights under BalX_1 and OptX_1, we see that the standard deviation in the average sampling weights under OptX_1 is smaller than that under BalX_1 (1.51 vs. 2.82). This may be one of the reasons why OptX_1 is slightly more efficient than BalX_1 for the estimation of β_1, as greater variability in the weights is known to impact the efficiency of IPW-GEE estimation.[64]

- *The distribution of the stratum-specific sample sizes across levels of the stratification variable X_1 is more extreme for negative dependence vs. positive dependence of the covariate of interest with X_1.*

2.7 Discussion

In this paper, we provide a formalized framework for determining optimal stratum-specific sample sizes when interest lies in estimating one or multiple associations in a marginal mean model and the analysis is conducted using inverse-probability-weighted generalized estimating equations. Through a comprehensive simulation study, we showed that the optimal allocation of sample sizes across strata generally yields the greatest gain in efficiency among all designs considered, performing better than the straightforward approach of taking a simple random sample of clusters, and better than the balanced stratified design. Our results indicate that when one parameter is of particular interest, the optimal allocation for that parameter yields the greatest efficiency gain for that parameter; furthermore, if the covariate of interest is positively associated with the stratification covariate, the optimal allocation for the parameter of interest also results in substantial efficiency gains for the parameter associated with the stratification covariate. However, optimizing for one parameter in particular can result in losses for other parameters in the model. This demonstrates a trade-off that is

Table 2.3: $K = 280$, varying N_k, K_s=80, σ_V=0.5. Shown are the average stratum-specific cluster-level sample sizes, and the average overall individual-level sample size n, under the eight data scenarios in which there is a dependence between X_1 and one of the other covariates in the model, across 10000 iterations. The average number of clusters sampled from stratum $(Y^* \geq Y^*_{0.80}, X_1 = 1)$ is k_{11}, the average number of clusters sampled from stratum $(Y^* < Y^*_{0.80}, X_1 = 1)$ is k_{01}, the average number of clusters sampled from stratum $(Y^* \geq Y^*_{0.80}, X_1 = 0)$ is k_{10}, and the average number of clusters sampled from stratum $(Y^* < Y^*_{0.80}, X_1 = 0)$ is k_{00}. See Sections 2.5.2 and 2.5.3 for details.

Covariate	Dependence with X_{1k}	Design	k_{11}	k_{01}	k_{10}	k_{00}	K_s	n	Most efficient for β_1
X_{2k}	Positive	BalX_1	20	20	20	20	80	4173	
		OptX_1	17	23	13	27	80	3704	✓
		OptX_2	22	26	11	22	80	3671	
	Negative	BalX_1	17	21	21	21	80	4105	
		OptX_1	11	24	15	30	80	3597	✓
		OptX_2	3	5	31	41	80	4016	
X_{3ki}	Positive	BalX_1	20	20	20	20	80	4208	
		OptX_1	21	19	12	28	80	3758	✓
		OptX_3	24	27	8	22	80	3544	
	Negative	BalX_1	17	21	21	21	80	4188	
		OptX_1	11	26	17	26	80	3741	✓
		OptX_3	3	6	24	47	80	3724	
X_{4k}	Positive	BalX_1	20	20	20	20	80	4184	
		OptX_1	16	21	15	29	80	3760	✓
		OptX_4	15	18	17	30	80	3873	
	Negative	BalX_1	20	20	20	20	80	4166	
		OptX_1	15	23	15	26	80	3736	✓
		OptX_4	10	13	22	35	80	3942	
X_{5ki}	Positive	BalX_1	20	20	20	20	80	4174	
		OptX_1	19	19	12	30	80	3717	✓
		OptX_5	15	17	12	36	80	3599	
	Negative	BalX_1	20	20	20	20	80	4160	
		OptX_1	14	634	16	26	80	3732	✓
		OptX_5	7	12	19	42	80	3670	

mitigated by the design that optimizes for all parameters simultaneously: under such a design, there are generally small to moderate efficiency gains across all parameters. Researchers must therefore identify the primary objective of the research question at the design stage, in order to select the most appropriate optimality criterion. While the simulation study described in the paper corresponds to a specific data setup, we also expanded upon the setting presented here, the results of which can be found in Appendix B.3. In general, the story remains the same, with differences related to the degree of efficiency gain observed.

As mentioned in the Introduction, the degree of efficiency gain depends upon a variety of factors, including which piece(s) of information are used to stratify the clusters, the degree of variability of the outcome and the covariates across clusters, the method of analysis, and the sample size allocations across the strata. In the setting we consider in this paper, only cluster-level summaries of the outcome would be available at the design stage. A threshold for the outcome summary measures must therefore be selected for stratification (as must be done for an individual-level covariate that is to be stratified upon). In the simulation study we conducted, we chose as the threshold, $Y^*_{0.80}$, the 80^{th} quantile of the distribution of the number of outcome cases across the K clusters. This may not have been the optimal stratification, and determining the best threshold for stratification is an area for future research. In the survey sampling context, when interest lies in estimating a mean or a total of a measure of interest, efficiency may be gained by creating homogenous strata. Making strata homogeneous for estimation of regression coefficients may be less straightforward.[82] With regard to the method of analysis, we assumed in this paper that the analysis would proceed via inverse-probability-weighted generalized estimating equations. While inverse-probability-weighted analysis methods are known to be inefficient compared to other approaches such as maximum likelihood or mean-score methods, the advantage of such an approach is robustness to model misspecification[9,62,68] and the fact that one does not need to model the relationship between the exposure of interest and the stratification variable(s) at the design stage.[40]

The results of our simulation study demonstrate the value in sampling clusters according to the optimal allocation - in other words, sampling clusters wisely has the potential for efficiency gains over designs such as simple random sampling or balanced stratified sampling, even when the analysis takes place at the level of the individual. There are several direct extensions of this work. First, as has been mentioned, the formulae for the optimal stratum-specific sample sizes depend on quantities that, at the outset, will be unknown, including the true parameters of the target marginal mean model, β_0, as well as the values of the variance components. Now that we have established that there are indeed potential benefits to be gained, developing and evaluating creative strategies to operationalize these designs in practice is, therefore, an important avenue for future research. This research, for example, could build on adaptive strategies that have been proposed for other instances in which key parameters are unknown at the outset, such as that proposed in the context of sample size and power calculations,[24] or that presented for approximating optimal allocation in other settings.[15,21,42,69] Another potential way forward would be to use imputation, following the approaches taken in the analysis of longitudinal data.[62,66] Furthermore, with regard to optimizing for multiple parameters, our simulation study only evaluated the approach in which every parameter is given equal weight ($w_q = 1$ for $q = 1, \ldots, p$). There may be instances in which researchers are interested in estimating multiple but not all parameters with high precision. For example, one may be interested in estimating two of the main effects and their interaction term with precision. Even if interest lies in estimating all parameters with precision (i.e. all w_q non-zero), it may be preferable to assign different weights to different parameters. A topic for future research therefore concerns expanding the study of optimal allocation in the setting when multiple parameters are of interest. In other cases, researchers may not know the exact specification of the model in the final analysis, and may prefer to move forward with a sampling design that optimizes for more than one model, presenting another area for future research. Finally, in this paper we considered single-stage cluster-based sampling designs. A perhaps more common design is a two-stage cluster-based sampling de-

sign. In such a setting, there may be a specific cost associated with travel to clusters, and a certain cost associated with collecting information on individuals once at a cluster. Given a budgetary constraint, researchers must determine which clusters to sample, and subsequently, which individuals to sample within the sampled clusters, based on information that is available at the design stage. Extending the framework for optimal allocation to this setting, and investigating the potential for efficiency gains, is yet another area for future research.

Acknowledgements

Ms. Sauer was funded by the Harvard T.H. Chan School of Public Health NIEHS-sponsored Environmental Training Grant T32ES007142. Dr. Haneuse was supported by NIH grant R01 HL094786. We are gratefulto D-tree International and Zanzibar Ministry of Health for access to the Safer Deliveries (*Uzazi Salama*) program data, which informed the design of the simulation studies in this paper.

3

Practical strategies for operationalizing optimal allocation in stratified cluster-based outcome-dependent sampling designs

Sara Sauer[1], Bethany Hedt-Gauthier[1,2], Sebastien Haneuse[1]

1 *Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA USA*

2 *Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA USA*

Abstract

Cluster-based outcome-dependent sampling (ODS) has the potential to yield efficiency gains when the outcome of interest is relatively rare, and resource constraints allow only a certain number of clusters to be visited for data collection. Previous research has shown that when the intended analysis is inverse-probability weighted generalized estimating equations (IPW-GEE), and the number of clusters that can be sampled is fixed, optimal allocation of the (cluster-level) sample size across strata defined by auxiliary variables readily available at the design stage has the potential to increase efficiency in the estimation of the parameter(s) of interest. In such a setting, the optimal allocation formulae depend on quantities that are unknown, currently making such designs difficult to implement in practice. In this paper, we consider an adaptive sampling approach, in which a first wave sample is collected using balanced stratified sampling, and subsequently used to compute the optimal second wave stratum-specific sample sizes. We consider two strategies for estimating the necessary components using the first wave data: an inverse-probability weighting (IPW) approach and a multiple imputation (MI) approach. In a comprehensive simulation study, we show that the adaptive sampling approach performs well, and that the MI approach yields designs that are near-optimal, regardless of the covariate type. The IPW approach, on the other hand, has mixed results, with better performance for parameters associated with individual-level covariates compared to those corresponding to cluster-level covariates. Finally, we illustrate the proposed adaptive sampling procedures with data on maternal characteristics and birth outcomes among women enrolled in the Safer Deliveries program in Zanzibar, Tanzania.

3.1 Introduction

Outcome-dependent sampling (ODS) can be a cost-efficient sampling strategy when resources for data collection are limited. When the outcome of interest is rare, and sampling occurs at the level of the individual, sampling designs that use information on the outcomes of the individuals in the population of interest, such as the case-control design, have demonstrated potential for efficiency gains over alternative sampling schemes such as simple random sampling. In public health settings, individuals are often clustered. When resource constraints prevent researchers from being able to carry out data collection in every cluster, a cluster-based ODS design that leverages information on a summary measure of the outcomes in a cluster and possibly some other auxiliary variable(s) known at the design stage, can be implemented instead.

One way to operationalize a cluster-based ODS design is to stratify the clusters based on pieces of information known at the design stage, and to then sample a certain number of clusters from each of the strata. Data that has been collected in such a way can be analyzed using inverse-probability weighted generalized estimating equations (IPW-GEE), [12,60] where the weights are the inverse of the cluster-specific probabilities of selection. The number of clusters sampled from each of the defined strata can influence the efficiency gain or loss for a particular parameter in the analysis model, and in Chapter 2 we derived formulae for the optimal allocation of the (cluster-level) sample size across the strata, when the intended analysis is IPW-GEE. We showed that such an optimal allocation strategy yields efficiency gains for the parameter of interest relative to simple random sampling of clusters or balanced sampling of clusters across strata.

One major obstacle to implementing such a design in practice, however, is the fact that the formulae for the optimal stratum-specific sample sizes depend on quantities that are, at the outset, unknown. Such quantities include the true parameter values and the variance components, which

rely on having complete information on the outcome and the covariates in the analysis model for all individuals in the population. If external pilot data are available, one could in principle obtain estimates of the various design components needed. Such pilot studies can be prohibitively expensive, however, or the efficiency to be gained under the optimal allocation design may not justify the additional cost of conducting the external pilot study.[42] As an alternative, we propose a two-stage adaptive sampling scheme, in which the data collected at the first stage serves as an internal pilot study that is used to obtain estimates of the quantities needed to calculate the stratum-specific sample sizes for the clusters that can be sampled with the remaining resources.

The use of an internal pilot study has been shown to be effective in other settings.[17,24,34,42,78] For example, Haneuse et. al (2011) proposed a two-stage strategy for conducting power and sample size calculations that account for patterns of confounding in observational studies. This approach involves the use of data collected to-date as an internal pilot study, which is used to estimate necessary components that would otherwise be unknown, such as the joint distribution of the covariates of interest. In a setting adjacent to the one we consider, McIsaac et. al (2015) presented an adaptive sampling approach to operationalize optimal allocation in two-phase designs, when the intended analysis is the mean score method. More recently, the use of an internal pilot study through adaptive, or multi-wave sampling, has also been proposed by Han et. al (2020)[21], Tao et. al (2020),[69] and Chen and Lumley (2020),[71] and with good results.

In this paper, we consider two-wave adaptive sampling in settings where the target population is clustered into K clusters, but resources are available to sample only $K_s < K$ clusters. Such a strategy involves sampling $K_{s,1}$ clusters in the first wave, using this first wave data to estimate the optimal allocation of the remaining resources, and sampling the remaining $K_{s,2} = K_s - K_{s,1}$ clusters according to the (approximated) optimal allocation. We develop and evaluate two approaches to estimating the variance components given the first wave data: an inverse-probability weighting (IPW) approach and a multi-level multiple imputation (MI) approach that employs the imputation approach pro-

posed by Jolani et. al (2015).[29] In a comprehensive simulation study, we evaluate the performance of the adaptive sampling approach. The rest of this paper is organized as follows: in Section 2, we detail the optimal allocation formulae; in Section 3, we describe our proposed adaptive sampling strategies; Section 4 describes the simulation study we conducted, and Section 5 applies the proposed methods to a dataset on women enrolled in the Safer Deliveries program, the goal of which was to to reduce the rate of maternal mortality in Zanzibar, Tanzania. Finally, Section 6 provides a discussion and concluding remarks.

3.2 Review of Optimal Allocation in Stratified Cluster-Based ODS

In Chapter 2 we proposed a framework for the optimal allocation of finite resources across defined strata, when the final analysis is expected to proceed via IPW-GEE. Those results are theoretical in the sense that the proposed allocation formulae assumed knowledge of unknown quantities, including the true parameter values. As such, while the results of Chapter 2 are important, they do not immediately translate into practice. Our goal in this paper is therefore to propose a strategy that enables the operationalization of such a design in practice. The rest of this section reviews the framework laid out in Chapter 2, while the following section describes the proposed adaptive sampling strategies.

3.2.1 Model of Interest

Suppose that interest lies in learning about the relationship between an outcome Y and a p-vector of covariates, X, in a population where the individuals exhibit cluster-correlation in their outcomes. Specifically, the population of interest is made up of K clusters, with N_k individuals in the k^{th} cluster, such that the total number of individuals in the target population is $N = \sum_{k=1}^{K} N_k$. Furthermore, we assume that estimation and inference will be performed using the following marginal

mean model for the outcome of the i^{th} individual in the k^{th} cluster:

$$\mu_{ki} = E[Y_{ki}|X_{ki}] = g^{-1}(X_{ki}^T\beta), \tag{3.1}$$

where $g(\cdot)$ is a user-chosen link function and β a p-vector of regression parameters.

3.2.2 Estimation and Inference

Given data that has been collected through a single-stage cluster-stratified outcome-dependent sampling design, estimation and inference can proceed via IPW-GEE. [12,60] In particular, the regression parameters β can be estimated as the solution to the following:

$$\mathcal{U}_w(\beta) = \sum_{k=1}^{K} \frac{R_k}{\pi_k} \mathbf{D}_k^T \mathbf{V}_k^{-1} \boldsymbol{\varepsilon}_k = 0. \tag{3.2}$$

where R_k is equal to 1 if cluster k is sampled and is equal to 0 otherwise; π_k is the probability of cluster k being sampled; $\boldsymbol{\varepsilon}_k = (\mathbf{Y}_k - \boldsymbol{\mu}_k)$, with $\mathbf{Y}_k = (Y_{k1}, \ldots, Y_{kN_k})$ and $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kN_k})$. The $N_k \times p$ matrix of partial derivatives is denoted by $\mathbf{D}_k = \partial\boldsymbol{\mu}_k/\partial\beta$, and $\mathbf{V}_k$, indexed by the unknown α, is an $N_k \times N_k$ working specification for $Cov[\mathbf{Y}_k]$.

Note, expression (3.2) can be rewritten as $\mathcal{U}_w(\beta) = \mathbf{U}^T\mathbf{W}\mathbf{R}$, where $\mathbf{U} = diag\{\mathbf{Y} - \boldsymbol{\mu}\}\mathbf{V}^{-1}\mathbf{D}$ is an $N \times p$ matrix, $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_K)^T$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)^T$, $\mathbf{V}$ is an $N \times N$ block-diagonal matrix, with the $\mathbf{V}_k$ on the diagonal, and $\mathbf{D}$ is the $N \times p$ matrix obtained by stacking the K $\mathbf{D}_k$ matrices; $\mathbf{W}$ is an $N \times N$ diagonal matrix with diagonal entries equal to the vector $(\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_K)^T$, with $\mathbf{W}_k$ a vector of length N_k with each element equal to π_k^{-1}. Letting $\mathbf{R}_k$ denote the $N_k \times 1$ vector with all entries equal to R_k, $\mathbf{R}$ is the $N \times 1$ vector obtained by concatenating the $\mathbf{R}_k$ together. Sauer et. al (2021)[60] showed that $\widehat{\beta}_w$, the solution to (3.2), is consistent for β_0, the true value of β, and asymptotically multivariate Normal, with

$$Var[\widehat{\beta}_w] = \mathbf{H}(\beta_0)^{-1} \left\{ Var[\mathcal{U}_w(\beta)]|_{\beta=\beta_0} \right\} \mathbf{H}(\beta_0)^{-1}.$$

In the expression above, $\mathbf{H}(\beta) = E[-\partial\mathcal{U}/\partial\beta], \mathcal{U} = \mathbf{U}^T 1_{N\times 1}$, $Var[\mathcal{U}_w(\beta)] = Var[\mathbf{U}^T 1_{N\times 1}] + E[\mathbf{U}^T\mathbf{W}\boldsymbol{\Delta}\mathbf{W}\mathbf{U}]$, $\boldsymbol{\Delta} = Var[\mathbf{R}|\mathcal{F}_K]$, and $\mathcal{F}_K = \{\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{S}\}$, where $\mathbf{Z}$ is a collection of variables available at the design stage that, while not included in the model, may be used to define the stratification of the clusters, and $\mathbf{S}$ is a vector of values of the variable S which defines the stratification of the clusters into J strata.

3.2.3 Optimal Allocation

The optimal allocation for the estimation of β_q, the q^{th} parameter in (3.1), involves determining the stratum-specific sample sizes $k_j, j = 1, \ldots, J$ such that the variance of $\widehat{\beta}_q$ is minimized, subject to the constraint that $\sum_{j=1}^{J} k_j = K_s$. This corresponds to determining the allocation that minimizes the $[q,q]^{th}$ entry of $E[\mathbf{H}^{-1}\mathbf{U}^T\mathbf{W}\boldsymbol{\Delta}\mathbf{W}\mathbf{U}\mathbf{H}^{-1}]$, the term of $Var[\widehat{\beta}_w]$ which depends on the selection indicators $\mathbf{R}$. This is given by:

$$E[\mathbf{H}^{-1}\mathbf{U}^T\mathbf{W}\boldsymbol{\Delta}\mathbf{W}\mathbf{U}\mathbf{H}^{-1}]_{[q,q]} = \sum_{j=1}^{J} \frac{K_j - k_j}{k_j}[A_{q,j} - \frac{B_{q,j}}{K_j - 1}] = \sum_{j=1}^{J} \frac{K_j - k_j}{k_j} C_{q,j}$$

where $A_{q,j} = \sum_{k\in S_j} \sum_{i=1}^{N_k} \sum_{i'=1}^{N_k} E[h_{ki}^{[q]} h_{ki'}^{[q]}]$, $B_{q,j} = \sum_{k\in S_j} \sum_{k'\neq k\in S_j} \sum_{i=1}^{N_k} \sum_{i'=1}^{N_{k'}} E[h_{ki}^{[q]} h_{k'i'}^{[q]}]$, $h_{ki}^{[q]}$ is the entry in the $(q+1)^{th}$ column of $\mathbf{U}\mathbf{H}^{-1}$ corresponding to the i^{th} individual in the k^{th} cluster, and $S_j = \{k : \text{cluster } k \in \text{stratum } j\}$. The optimization problem involves minimizing $f_q(k_1, k_2, \ldots, k_J) = \sum_{j=1}^{J} \frac{K_j - k_j}{k_j} C_{q,j}$ subject to the constraint that $\sum_{j=1}^{J} k_j = K_s$, and can be solved using the method of Lagrange multipliers, from which it follows that

$$k_j = K_s \frac{(K_j^{1/2} C_{q,j}^{1/2})}{\sum_{j=1}^{J} K_j^{1/2} C_{q,j}^{1/2}}. \tag{3.3}$$

The formula for the k_j given in expression (3.3) depends on knowledge of the true parameter values β_0, as well as the variance components through the $C_{q,j}$. To implement the optimal allocation in practice, we propose a two-wave adaptive sampling approach. In the first wave, $K_{s,1}$ clusters are sampled through a simple to implement but likely sub-optimal strategy, such as balanced stratified sampling. In the second wave, $K_{s,2} = K_s - K_{s,1}$ clusters are then sampled through the *approximately* optimal design, using estimates of β_0 and the variance components obtained from analyzing the data collected from the $K_{s,1}$ clusters sampled in the first wave. By approximately optimal, we mean a design that yields stratum-specific sample sizes that are close to the optimal allocation one would get with complete knowledge of the quantities needed to compute (3.3). The overall number of clusters sampled from stratum j is $k_j = k_{j,1}+k_{j,2}$, where $\sum_j k_{j,1} = K_{s,1}$, $\sum_j k_{j,2} = K_{s,2}$, and $K_{s,1}+K_{s,2} = K_s$. The probability of selection for cluster k in stratum j is given by:

$$\begin{aligned}
\pi_k &= P(R_k = 1|\mathcal{D}^*) = P(R_{k,1} = 1 \cup R_{k,2} = 1|\mathcal{D}^*) \\
&= P(R_{k,1} = 1|\mathcal{D}^*) + P(R_{k,2} = 1|R_{k,1} = 0, \mathcal{D}^*) * P(R_{k,1} = 0|\mathcal{D}^*) \\
&= \frac{k_{j,1}}{K_j} + \frac{k_{j,2}}{K_j - k_{j,1}} \times \frac{K_j - k_{j,1}}{K_j} = \frac{k_{j,1}}{K_j} + \frac{k_{j,2}}{K_j} = \frac{k_j}{K_j}
\end{aligned}$$

where $\mathcal{D}^*$ is the totality of the information available at the design stage, $R_{k,1}$ is the selection indicator for cluster k in the first sampling wave, and $R_{k,2}$ denotes the selection indicator for cluster k in the second sampling wave. The overall probability of selection under adaptive sampling is therefore the same as that under the non-adaptive setting. We can then determine the second wave sample sizes, $k_{j,2}$ by minimizing

$$E[\mathbf{H}^{-1}\mathbf{U}^T\mathbf{W}\boldsymbol{\Delta}\mathbf{W}\mathbf{U}\mathbf{H}^{-1}]_{q,q} = \sum_{j=1}^{J} \frac{K_j - k_{j,1} - k_{j,2}}{k_{j,1}+k_{j,2}} C_{q,j}$$

subject to the constraint that $\sum_{j=1}^{J} k_{j,1} + \sum_{j=1}^{J} k_{j,2} = K_s$. Defining the Lagrangian

$$\mathcal{L} := \sum_{j=1}^{J} \frac{K_j - k_{j,1} - k_{j,2}}{k_{j,1} + k_{j,2}} C_{q,j} + \eta(\sum_{j=1}^{J} k_{j,1} + \sum_{j=1}^{J} k_{j,2} - K_s),$$

taking the partial derivatives with respect to the $k_{j,2}$ and η, and setting them equal to 0 yields a system of $J+1$ equations. Solving the system of equations for $k_{j,2}$ and η yields

$$k_{j,2} = K_s \frac{(K_j^{1/2} C_{q,j}^{1/2})}{\sum_{j=1}^{J} K_j^{1/2} C_{q,j}^{1/2}} - k_{j,1},$$

which must be estimated by estimating $C_{q,j}$. In the following two sections we describe two different approaches to estimating the $C_{q,j}$.

3.3.1 Inverse-Probability Weighting

Given the first wave data, one can estimate $E[\mathbf{H}^{-1}\mathbf{U}^T\mathbf{W}\boldsymbol{\Delta}\mathbf{W}\mathbf{U}\mathbf{H}^{-1}]$ with

$$\widehat{\mathbf{H}}^{-1}\mathbf{U}^T\mathbf{W}diag(\mathbf{R}_1)\widetilde{\boldsymbol{\Delta}}_1 diag(\mathbf{R}_1)\mathbf{W}\mathbf{U}\widehat{\mathbf{H}}^{-1},$$

where $\widehat{\mathbf{H}} = \sum_{k=1}^{K} \frac{R_{k,1}}{\pi_{k,1}} \mathbf{D}_k^T \mathbf{V}_k^{-1} \mathbf{D}_k$ and $\pi_{k,1}$ is the probability of cluster k being sampled in the first wave, $\frac{k_{j,1}}{K_j}$. The $N \times 1$ vector $\mathbf{R}_1$ is made of up of the K first wave N_k-vectors of the selection indicators concatenated together. Finally, $\widetilde{\boldsymbol{\Delta}}_1$ is an $N \times N$ matrix with entries in the k^{th} diagonal block equal to $\frac{\pi_k - \pi_k^2}{\pi_{k,1}}$ and entries in the off-diagonal block corresponding to clusters k and k' equal to $\frac{\pi_{kk'} - \pi_k \pi_{k'}}{\pi_{kk',1}}$, where $\pi_{kk',1}$ is the joint probability of clusters k and k' being sampled in the first wave. If clusters k and k' belong to the same stratum, $\pi_{kk',1} = \frac{k_{j,1}}{K_j} \frac{k_{j,1}-1}{K_j-1}$, whereas the joint probability is equal to $\frac{k_{j,1}}{K_j} \frac{k_{j,1}}{K_j}$ if these two clusters belong to different strata.

3.3.2 Multiple Imputation

One of the drawbacks of the IPW approach is that only information on the individuals in the $K_{s,1}$ clusters sampled in the first wave is used in the estimation of the $C_{q,j}$. The IPW approach therefore

discards some pieces of information that are available for the individuals in all K clusters.[66] For example, the values of the cluster-level covariates in the model, and the cluster-level summaries of the outcome, which we assume to be available for all K clusters from the outset. We therefore also consider a second approach to estimating the $C_{q,j}$.

Here we assume that after the first wave data has been collected, researchers have the following information available to them: the values of cluster-level variables for all K clusters, as well as the proportion of outcome cases and size of each cluster, and the values of individual-level covariates for all of the individuals sampled in the first wave. One must therefore impute the individual-level covariates in the model of interest for the individuals *not* sampled in the first wave. In this setting, the individual-level covariates are *systematically* missing,[56] which in this context means that they are missing completely for the clusters not sampled in the first wave. To impute these individual-level covariates, we use the multilevel imputation approach of Jolani et. al (2015),[5,29] which takes into account the hierarchical structure of the data. Briefly, this approach uses a fully conditional specification (FCS),[72] where a conditional model is specified for every variable that has some missingness. Using a generalized linear mixed model for the imputation model, this approach first draws the imputation model parameters from their posterior predictive distributions using a noninformative Jeffreys prior; the missing covariate values are then drawn from its posterior predictive distribution given the imputation model parameters. The multiple imputation approach we propose involves the following steps:

1. generate M imputed complete datasets
2. for each of these M datasets, compute the $C_{q,j}$
3. compute $\widetilde{C}_{q,j} = \frac{1}{M} \sum_{m=1}^{M} \widetilde{C}_{q,j}^{m}$
4. compute the $k_{j,2}$ using $\widetilde{C}_{q,j}$ as an estimate for $C_{q,j}$.

3.3.3 Other Practical Issues

It is possible for the computed second wave sample sizes to yield $k_{j,2} > K_j - k_{j,1}$; such a setting is problematic, as there are not enough clusters remaining in stratum j after the first wave in order to sample the $k_{j,2}$ clusters in the second wave. On the other hand, it is also possible for the computed $k_{j,2}$ to be negative; such a scenario suggests that fewer clusters should have been sampled from stratum j, but as the $k_{j,1}$ clusters have already been sampled in the first wave, it is not possible to sample fewer clusters at this point. We refer to the cases in which either $k_{j,2} > K_j - k_{j,1}$ or $k_{j,2} < 0$ as *edge cases*, and develop a strategy for handling these scenarios, the details of which are given in Appendix C.1. Briefly, we define a threshold τ, which denotes the number of edge cases we are willing to tolerate. If the number of edge cases is greater than or equal to τ, we sample $K_{s,1_{inc}}$ more clusters via the first wave sampling strategy and recompute the $k_{j,2}$ using the larger first wave sample. If the number of edge case is less than τ, on the other hand, we fix the edge cases at the boundary (i.e. set $k_{j,2} = K_j - k_{j,1}$ if $k_{j,2} > K_j - k_{j,1}$ and set $k_{j,2} = 0$ if $k_{j,2} < 0$), and recompute the other $k_{j,2}$ with an updated constraint:

$$K_s^* = K_s - \sum_{j:k_{j,2}<0} k_{j,1} - \sum_{j:k_{j,2}>K_j-k_{j,1}} K_j$$

We note that if the number of edge cases is greater than τ, the sampling strategy becomes one of $s > 2$ waves. This may not be desirable in certain settings, in which getting permission to do data collection at health centers (clusters) is a long process. This is therefore one of the features we compare between the two proposed estimation procedures in our simulation study, which is described in the next section.

3.4 Simulation Study

We conducted a simulation study to evaluate the performance of the proposed adaptive sampling approach, using both estimation (IPW and MI) strategies. The model of interest for the i^{th} individual in the k^{th} cluster used in the simulation study is given by:

$$logit(P(Y_{ki} = 1)) = \beta_0 + \beta_1 X_{1k} + \beta_2 X_{2k} + \beta_3 X_{3ki} + \beta_4 X_{4k} + \beta_5 X_{5ki}$$

where $\beta_0 = (-3.1, 0.3, 0.7, 0.7, 0.7, 0.7)$. In order to investigate how the performance of the adaptive sampling procedure is affected by the type of the covariate of interest, and the direction of the association with the stratification variable(s), we consider nine data scenarios, which are based on the simulation study in Chapter 2; Table 3.1 provides a summary. In the baseline scenario, X_{1k} is a binary cluster-level covariate with prevalence of 0.30. $X_{2k} \sim N(1, \sigma = 0.25)$ is a continuous cluster-level covariate, while $X_{3ki} \sim N(1, \sigma = 0.25)$ is a continuous individual-level covariate. $X_{4k} \sim Ber(p_k)$ is a binary cluster-level covariate with p_k = expit(-0.9), while $X_{5ki} \sim Ber(p_k)$ is a binary individual-level covariate with $p_k = 0.25$. In the baseline scenario, there is no dependence between the covariate used for stratification, X_1, and the other covariates in the model. We build on this scenario by changing the relationship between X_1 and each of the other covariates in turn, considering both positive and negative associations with X_1.

3.4.1 Designs

For each data scenario, we generated 10000 complete datasets of $K = 280$ clusters with varying cluster sizes. The number of clusters and the variation of the cluster sizes was set equal to these in the dataset on maternal characteristics and birth outcomes among women enrolled in the Safer Deliveries program, a dataset which we discuss in more detail in Secton 3.6. Correlation between the

Table 3.1: Covariate distributions for nine simulation scenarios considered in Section 3.4.

Covariate	Baseline Scenario	Eight scenarios in which X_{1k} and each of the remaining covariates are dependent	
		Positive dependence	Negative dependence
X_{1k}	$Ber(0.3)$	–	–
X_{2k}	$N(1, 0.25)$	$\mu = 1 + 0.5I_{(X_{1k}=1)}$ $\sigma = 0.25 + 0.75I_{(X_{1k}=1)}$	$\mu = 1 + 0.5I_{(X_{1k}=0)}$ $\sigma = 0.25 + 0.75I_{(X_{1k}=0)}$
X_{3ki}	$N(1, 0.25)$	$\mu = 1 + 0.5I_{(X_{1k}=1)}$ $\sigma = 0.25 + 0.75I_{(X_{1k}=1)}$	$\mu = 1 + 0.5I_{(X_{1k}=0)}$ $\sigma = 0.25 + 0.75I_{(X_{1k}=0)}$
X_{4k}	$Ber(p_k)$, $p_k = expit(-0.9)$	$p_k = 0.6I_{(X_{1k}=1))} +$ $0.156I_{(X_{1k}=0)}$	$p_k = 0.156I_{(X_{1k}=1)} +$ $0.346I_{(X_{1k}=0)}$
X_{5ki}	$Ber(p_k)$, $p_k = 0.25$	$p_k = 0.6I_{(X_{1k}=1))} +$ $0.1I_{(X_{1k}=0)}$	$p_k = 0.1I_{(X_{1k}=1))} +$ $0.314I_{(X_{1k}=0)}$

outcomes was induced using the `MMLB` package for R. For each generated complete dataset, the 280 clusters were stratified according to $Y^*_{0.80}$, the 80^{th} quantile of the number of outcome cases across the K clusters, and X_1. Then, $K_s = 80$ clusters were sampled from the four strata according to the following five designs: *(i)* optimal allocation for the estimation of β_1 (OptX_1), *(ii)* optimal allocation for the estimation of β_2 (OptX_2), *(iii)* optimal allocation for the estimation of β_3 (OptX_3), *(iv)* optimal allocation for the estimation of β_4 (OptX_4), and *(v)* optimal allocation for the estimation of β_5 (OptX_5).

3.4.2 Design Approximations

For each of the scenarios described in the previous section, we determined the optimal allocation using the complete data set and true parameter values (gold standard), and approximated the op-

timal design using i) adaptive sampling with IPW estimation and ii) adaptive sampling with MI estimation. For both estimation approaches, we investigated the impact of the relative sizes of the first wave and second wave samples (i.e. $K_{s,1}$ and $K_{s,2}$) on the performance of the optimal allocation approximation. In particular, we varied $K_{s,1}/K_{s,2} \in \{20/60, 40/40, 60/20\}$. Furthermore, we set the first wave increment size to 20 clusters (see Section 3.3.3) and set the threshold for the number of tolerated edge cases to $\tau = 3$. For the MI approach, we took the cluster-level variables X_1, X_2, and X_4, as well as the outcome Y, to be known. We then used the multilevel imputation approach of Jolani et. al (2015)[28] to impute the continuous individual-level covariate X_3 and the binary individual-level covariate X_5 for the clusters not sampled in the first wave, using the `micemd` and `mice` packages in R. For the setting in which X_1 and X_3 are associated, due to the fact that the variance of X_3 depends on X_1, we imputed X_3 and X_5 separately for the clusters with $X_1 = 1$ and $X_1 = 0$. We set M, the number of imputed data sets within one iteration, to 5 in all settings. The first wave sample size was selected using balanced stratified sampling (BalX_1).

3.4.3 Analyses

For each sample obtained under the various design options, we computed the point estimates by solving expression (3.2). For each design, across the 10000 iterations, we computed the mean point estimates, as well as the standard deviation of the point estimates. In order to evaluate the performance of the adaptive sampling strategies in approximating the optimal design, we computed the change relative to the optimal design in the standard deviation of the point estimates, expressed as a percentage:

$$\left(\frac{sd(\widehat{\beta}^{1:R}_{w,q})_{Adapt} - sd(\widehat{\beta}^{1:R}_{w,q})_{Opt}}{sd(\widehat{\beta}^{1:R}_{w,q})_{Opt}} \right) \times 100 \text{ for } q = \{1, \ldots, p\}.$$

where R is the number of simulation iterations, in this case 10000. Furthermore, we computed the difference between the stratum-specific sample sizes under the adaptive designs and the optimal

design, as a fraction of the overall stratum size, a measure presented in McIsaac et. al (2015)[42]:

$$\frac{k^i_{jAdapt} - k^i_{jOpt}}{K^i_{jOpt}} \text{ for } i = \{1, \ldots, 10000\},\ j = \{1, \ldots, J\}.$$

The latter can be thought of as a distance measure, by which to evaluate the accuracy and precision of the adaptive sampling strategies.

3.5 Results

Figure 3.1 shows the percent change in the standard deviation of the point estimates (first measure described in Section 3.4.3) for the settings in which there is a positive relationship between X_1 and the continuous cluster-level covariate X_2 (top panel) and the setting in which there is a negative relationship between X_1 and X_2 (bottom panel); similarly, Figure 3.2 shows the results corresponding to the setting in which there is a dependence (positive and negative) between X_1 and the continuous individual-level covariate, X_3.

3.5.1 Impact of Estimation Strategy

Based on Figures 3.1 and 3.2, we see that the efficiency losses under the adaptive sampling approach with MI estimation are generally lower than those under the IPW approach, particularly when the allocation is optimal with respect to a cluster-level covariate. In fact, under the MI approach in the settings in which there is a positive association between X_1 and X_2 (Figure 3.1) or X_1 and X_3 (Figure 3.2), the percent change in the standard deviation of the cluster-level parameter estimates $\widehat{\beta}_1$, $\widehat{\beta}_2$, and $\widehat{\beta}_4$, under the designs that are optimal with respect to those parameters, is less than 2.7% regardless of the first wave sample size. On the other hand, the percent loss under the IPW approach is as much as 11.7%. The superior performance of the MI approach for allocation to a cluster-level parameter is intuitive, as the MI approach leverages the cluster-level information that is available for

all K clusters in the population of interest, while the IPW approach uses only the cluster-level information available for the $K_{s,1}$ clusters sampled at the first wave. When the allocation is optimal with respect to an individual-level covariate, the performance of the IPW approach improves, though we see that the losses in efficiency are in general still greater than those under the MI approach, particularly when the first wave sample size is 20 or 40.

Additional results (not shown here) show boxplots of the stratum-specific differences between the adaptive and optimal sample sizes across the 10000 iterations (the second measure described in Section 3.4.3). Regardless of the data scenario and the first wave sample size, the variability of the IPW approach is substantially higher than that of the MI approach.

3.5.2 Impact of Imputation Model Misspecification

The results described in Section 3.5.1 seem to suggest that the MI approach combined with an appropriate first wave sample size is always the preferable strategy for operationalizing the optimal allocation in practice. However, we note that care must be taken in carrying out the imputation of variables with missing values. For instance, in the setting in which there is a relationship (positive or negative) between X_1 and X_3, imputing the missing values of X_3 separately for the clusters with $X_1 = 0$ and $X_1 = 1$ is necessary in order to capture the fact that the variance of X_3 depends on the value of X_1 (for example, $X_3 \sim \text{Normal}(\mu, \sigma^2)$ with $\sigma = 0.25 + 0.75I(X_1 = 1)$ in the positive dependence setting) - failing to do so results in efficiency losses that exceed those under the IPW approach: the percent loss in efficiency under the MI/OptX_3 approach ranges from 13.6% to 18.6% for the estimation of β_3, compared to losses of 3.4 to 6.8% under the IPW/OptX_3 approach. On the other hand, the stratified imputation approach yields losses of only 0 to 3.4 % under OptX_3 for the estimation of β_3, as is shown in the top panel of Figure 3.2 in this paper. The reason for the loss stems from the fact that in this setting, the variance of X_3 is higher in the strata with $X_1 = 1$. The optimal allocation therefore samples more clusters from these strata; imputing the values of X_3 without

carrying out the imputation separately according to levels of X_1 fails to capture the difference in the variation of the X_3 values according to X_1, and therefore undersamples from the more informative strata, i.e. those with $X_1 = 1$: under the optimal allocation approach, the average stratum-specific sample sizes for $(I_{(Y^* \geq Y^*_{0.80})}, X_1) = (1, 1)$ and $(I_{(Y^* \geq Y^*_{0.80})}, X_1 = (0, 1)$ under OptX_3 are $(k_{11}, k_{01}) = (22, 28)$, while under the unstratified MI adaptive approach with a first wave sample size of 20, the average sample sizes are $(k_{11}, k_{01}) = (13, 15)$.

3.5.3 Impact of First Wave Sampling Strategy

The performance of BalX_1 as the first wave sampling strategy differs depending on the nature of the dependence between the covariate of interest and the stratification covariate, X_1. In our simulation study, balanced stratified sampling was appropriate in the settings with a positive dependence. On the other hand, in the settings where the covariate of interest is negatively associated with X_1, balanced stratified sampling in the first wave resulted in too many clusters being sampled from the strata with $X_1 = 1$. The consequence is a loss in efficiency, as can be seen for the estimation of β_2 under OptX_2 in the bottom panel of Figure 3.1 (1.2 - 18.1% loss under the MI approach, 10.8 - 20.5% loss under the IPW approach) and for the estimation of β_3 under OptX_3 in the bottom panel of Figure 3.2 (0 - 14.9% loss under the MI approach, 6.4 - 14.9% loss under the IPW approach). Additional results (not shown here) in which the first wave sampling strategy involves sampling 40%, as opposed to 50%, of the first wave clusters from the strata with $X_1 = 1$ when $K_{s,1}$=20, and sampling 20% of the first wave clusters from these strata when $K_{s,1}$=40 or 60, show that such a strategy performs better in the negative dependence setting. For example, in the setting of negative dependence between X_1 and X_2, the loss in efficiency for the estimation of β_2 under OptX_2 is reduced to 1.2 - 2.4% under the MI approach and 3.6 - 4.8% under the IPW approach. Similarly, in the setting of negative dependence between X_1 and X_3, the loss in efficiency for the estimation of β_3 under OptX_3 is reduced to 1.2 - 2.4% under the MI approach and 4.3-6.4 % under the IPW approach. The

strategy of sampling fewer clusters from the strata with $X_1 = 1$ in the first wave under the negative dependence setting works well because this results in fewer clusters being sampled from the less informative strata.

3.5.4 Impact of Varying First Wave Sample Size

The impact of varying the first wave sample size depends on the estimation procedure (IPW or MI), whether the allocation is optimal with respect to a cluster-level or individual-level covariate, and the direction of the dependence between X_1 and the covariate of interest. Under the MI approach, the first wave sample sizes of 20 or 40 generally resulted in very close approximations to the optimal allocation, while the performance using a first wave sample of 60 clusters depended on the gold standard/optimal distribution of the k_js. When using IPW estimation, a first wave sample size of 20 clusters generally resulted in large efficiency losses. This seems to indicate that a first wave sample of 20 clusters is too small to yield reliable IPW estimates of the design components needed to compute the optimal allocation of the second wave sample.

The relative performance of a first wave sample of 40 or 60 clusters under the IPW approach depends upon the distribution of the k_js that the adaptive sampling strategy is meant to approximate. For example, across all settings considered with a positive dependence of one of the covariates with X_1, Opt$X_1/K_{s,1} = 60$ was the most efficient of the adaptive designs for the estimation of β_1, while Opt$X_4/K_{s,1} = 60$ was most efficient for estimation of β_4. On the other hand, across all settings of positive dependence with X_1, Opt$X_2/K_{s,1} = 40$ and Opt$X_3/K_{s,1} = 40$ were most efficient for the estimation of β_2 and β_3, respectively. This seems to suggest that when using the IPW approach, a larger first wave sample size yields better estimates of the stratum-specific sample sizes, as long as the first wave sample has not already 'overreached' (sampled too many clusters) in certain strata. For example, under positive dependence of X_1 and X_2, the average stratum-specific sample sizes using the optimal allocation design was $(k_{11}, k_{01}, k_{10}, k_{00}) = (17, 22, 15, 26)$ for OptX_1, and $(11, 14, 15, 40)$ for

OptX_3; under the adaptive approach for OptX_1, a first wave sample of 60 clusters permits a larger sample with which to estimate the second wave allocation, without oversampling too many clusters from any of the strata, while for OptX_3, a first wave sample of 40 clusters is more appropriate. In the settings of negative dependence with X_1, the story was similar.

3.5.5 Impact of Threshold for Edge Cases

The value of τ did not matter under the MI approach, as there were generally less than two edge cases for every iteration under this estimation strategy. We ran additional simulations to investigate the impact of varying the threshold $\tau \in \{2, 3, 4\}$ for the number of tolerated edge cases under the IPW approach in the scenario with a positive dependence between X_1 and X_2. The most pronounced differences arise when the first wave sample size is $K_{s,1}$=60, with only slight differences when $K_{s,1}$ = 20 or 40. When $K_{s,1}$ = 60, a threshold of $\tau = 2$ results in greater losses in efficiency for the estimation of β_0, as well as $\beta_3 - \beta_5$. This is most likely due to the fact that the lower threshold of $\tau = 2$ results in more clusters being sampled via the non-optimal first wave sampling strategy (BalX_1), which differs from OptX_3 - OptX_5 more than from OptX_1 and OptX_2.

3.5.6 Other Comments

We close this section with a few additional comments pertaining to the simulation results:

1. Though rare, there are a few instances in which the adaptive strategy is more efficient for the parameter of interest than the optimal allocation design. This is an observation that is also raised by both McIsaac et. al (2015)[42] and Chen and Lumley (2020),[71] and is most likely due to the fact that the optimal design is only optimal asymptotically (though as we see, still efficient in finite samples). In other instances, we see that the adaptive sampling strategy is more efficient for the other parameters in the model that do not correspond to the covariate

of interest. This is most likely due to the fact that the adaptive sampling strategy, while losing some efficiency for the parameter of interest, results in a design that mitigates the losses for some of the other parameters in the model.

2. Under the IPW approach, there was not much of a difference between using an increment of 4 clusters as opposed to an increment of 20 clusters.

3. The results for the settings in which there is a dependence between a binary covariate and X_1 are similar to those described in the previous subsections.

3.6 Data Application

3.6.1 Safer Deliveries Program

The Safer Deliveries program was an effort by the Zanzibar Ministry of Health and D-tree International to reduce the high rate of maternal mortality in Zanzibar, Tanzania, by increasing the rate of deliveries that occur in health facilities. [18,19] Initially piloted in 2011-2012, expanded from 2013 to 2014, and finally implemented at scale in 10 of Zanzibar's 11 districts between January 2016 and September 2019, the program involved enlisting and training community health workers (CHWs) to enroll pregnant women into the program and to subsequently provide guidance and support during the woman's pregnancy, all with the aid of a mobile app. [19]

As part of this process, the CHWs collected data on the women enrolled in the program, such as demographic and health information, obstetric history, the number of antenatal care (ANC) visits, and the number of visits the woman received by a CHW during pregnancy. Each woman's shehia (lowest offical administrative unit) of residence was also recorded, as was the location of delivery, after the woman had given birth.

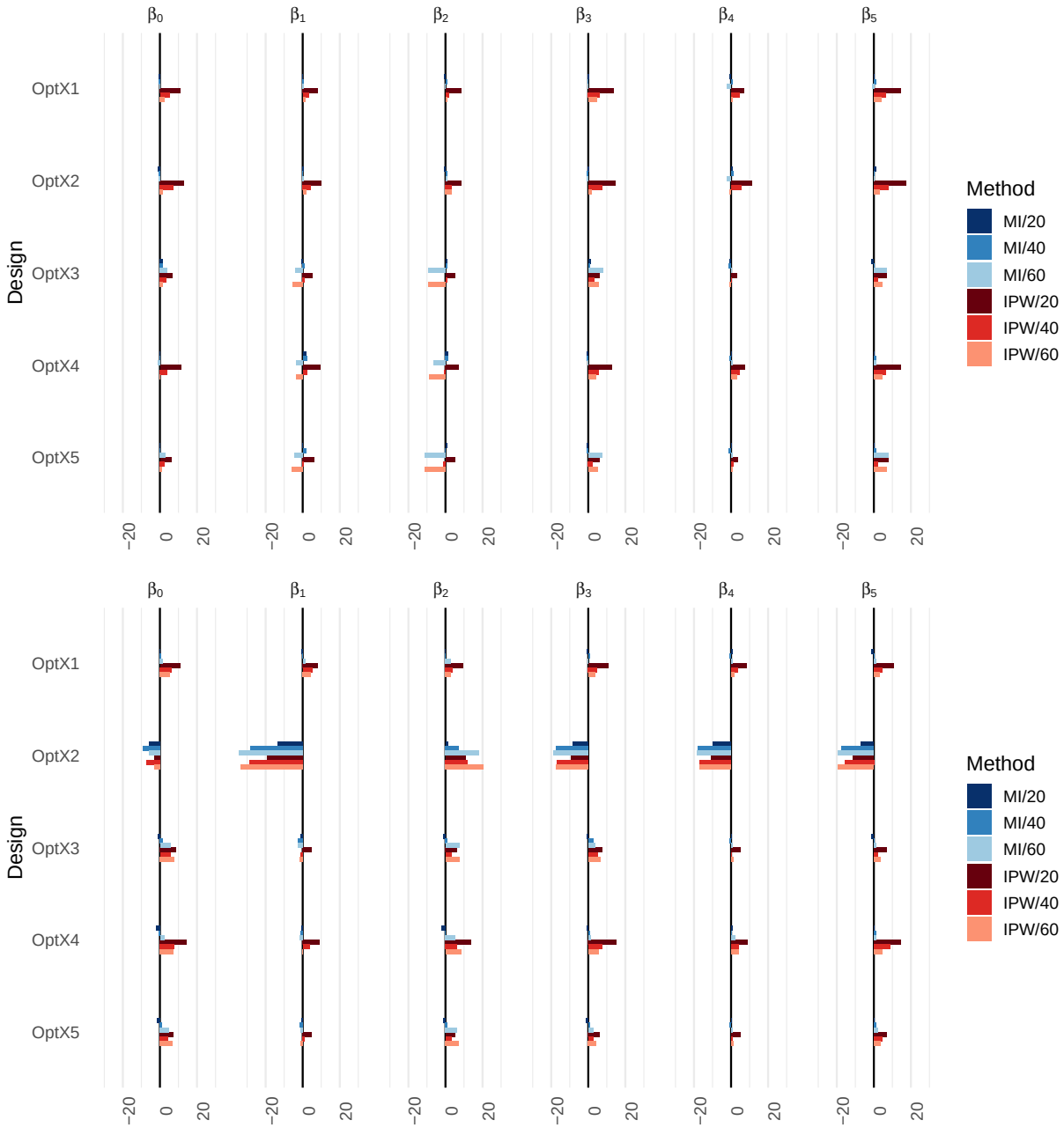

Figure 3.1: Shown is (sd($\widehat{\beta}^{1:R}_{q,Adapt}$) - sd($\widehat{\beta}^{1:R}_{q,Opt}$))/sd($\widehat{\beta}^{1:R}_{q,Opt}$), $q = 1, \ldots, p$ under the adaptive sampling strategy using IPW estimation (reds) and multiple imputation (blues), for $K_s = 80$ and $K_{s,1} \in \{20, 40, 60\}$ under positive dependence X_1 and X_2 (top panel), and negative dependence X_1 and X_2 (bottom panel).

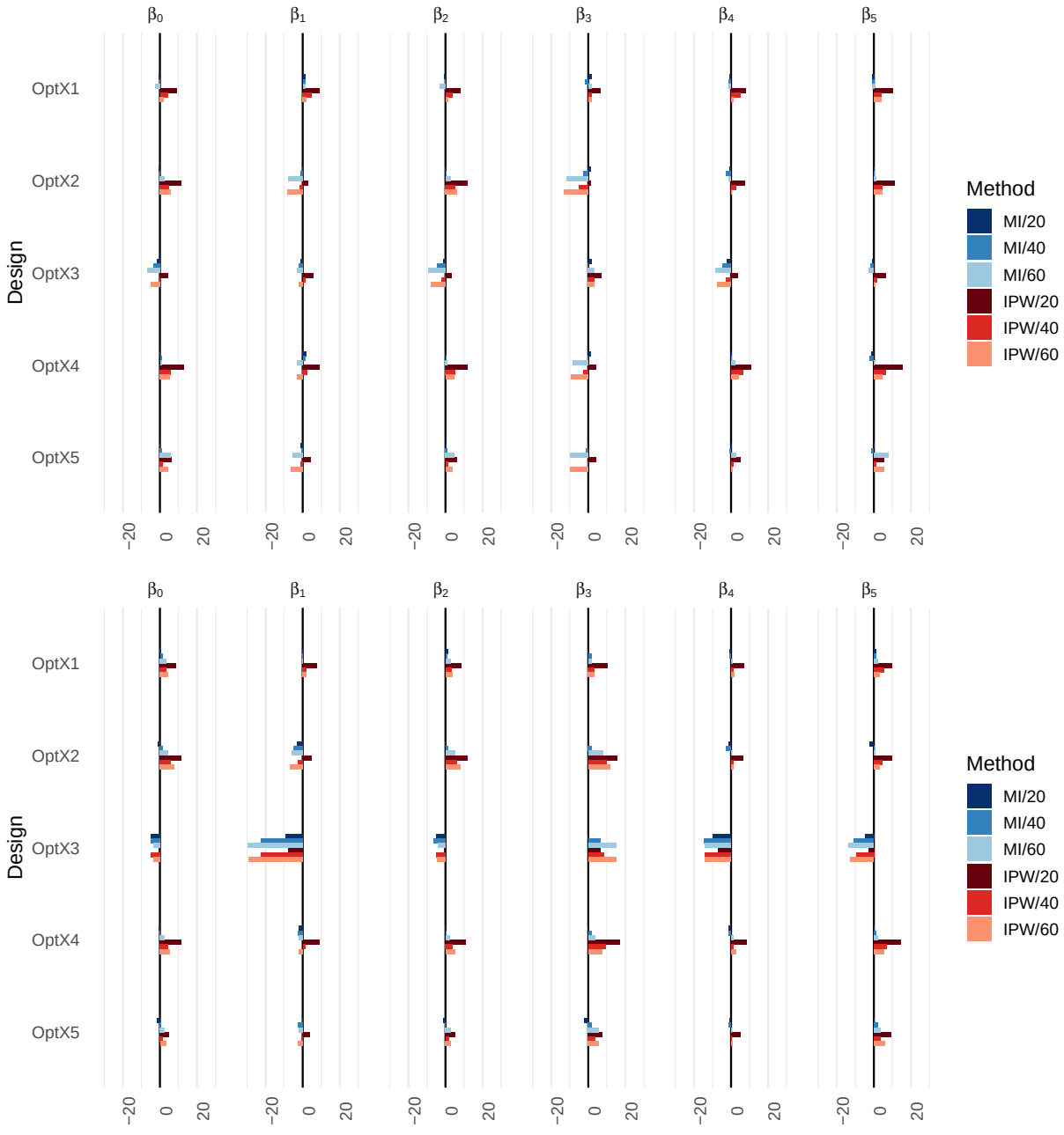

Figure 3.2: Shown is (sd($\widehat{\beta}^{1:R}_{q,Adapt}$) - sd($\widehat{\beta}^{1:R}_{q,Opt}$))/sd($\widehat{\beta}^{1:R}_{q,Opt}$), $q = 1, \ldots, p$ under the adaptive sampling strategy using IPW estimation (reds) and multiple imputation (blues), for $K_s = 80$ and $K_{s_1} \in \{20, 40, 60\}$ under positive dependence X_1 and X_3 (top panel), and negative dependence X_1 and X_3 (bottom panel).

Table 3.2 shows the characteristics of a subset of 28,789 women enrolled in the program who gave birth between 2014 and May 2017. This subset includes women from 280 shehias, with the number of women from each shehia ranging from 1 to 509. We see that a higher percentage of women in the $36-49$ age category delivered outside of a health facility (27.2%) compared to women in the $\leq$ 20 age category (21.0%) and women in the $21-35$ age category (26.1%). Furthermore, women from the island of Pemba are seemingly more likely to deliver outside of a health facility than women from the island of Unguja (34.5% vs. 18.4%). Finally, a larger proportion of women who did not have 4 ANC visits during pregnancy, did not receive a visit from a CHW at 8-9 months of pregnancy, and did not complete secondary school delivered outside of a health facility compared to women who did have 4 ANC visits, were visited by a CHW at 8-9 months of pregnancy, and did complete secondary school, respectively.

3.6.2 Hypothetical Study

Given that one of the primary aims of the Safer Deliveries program is to increase the rate of health facility deliveries, researchers may be interested in investigating the relationship between delivering outside of a health facility and potentially relevant factors. In the remainder of this section, we consider a hypothetical study, in which the following marginal mean model is of interest:

$$logit(P(Y_{ki}=1)) = \beta_0+\beta_1 X_{loc,k}+\beta_2 X_{age,k}+\beta_3 X_{ANC4,ki}+\beta_4 X_{educ,ki}+\beta_5 X^*_{educ,ki}+\beta_6 X_{ANC4,ki}\times X_{educ,ki} \tag{3.4}$$

where Y_{ki} is the indicator for whether the i^{th} woman in the k^{th} shehia delivered outside of a health facility (1=Yes/0=No), X_{1k} is a binary cluster-level covariate indicating which island the woman is from (1=Pemba/0=Unguja), X_{2ki} is a continuous individual-level covariate representing the

Table 3.2: Maternal Characteristics

	Delivered outside of a facility		
	No	Yes	%
Mother's Age			
≤20	3336	886	21.0
21-35	15689	5536	26.1
36-49	2433	909	27.2
Region			
Unguja	13115	2986	18.6
Pemba	8343	4345	34.2
4 ANC visits			
Yes	5968	1274	17.6
No	15490	6057	28.1
Previous abortions			
Yes	3101	1175	27.5
No	18357	6156	25.1
HIV status			
Positive	371	139	27.3
Negative	21087	7192	25.4
Previous location of delivery			
Outside of facility	3046	3519	53.6
Facility	12745	2855	18.3
No previous delivery	5667	957	14.4
CHW visit at < 6 months			
Yes	17994	6247	25.8
No	3464	1084	23.8
CHW visit at 6-8 months			
Yes	4576	1481	24.5
No	16882	5850	25.7
CHW visit at 8-9 months			
Yes	14461	4548	23.9
No	6997	2783	28.5
Education level			
Low	17847	6591	27.0
High	3611	740	17.0
Recommended facility type			
Cottage hospital	7700	1915	19.9
PHCU+	4550	1728	27.5
Referral hospital	9208	3688	28.6

mother's (standardized) age, X_{3ki} is a binary individual-level variable representing whether the mother in the k^{th} shehia had 4 ANC visits during pregnancy (1=Yes/0=No), X_{4ki} is a binary individual-level variable indicating whether the woman completed secondary school (1=Yes/0=No), and X_{5k} is cluster-level covariate representing the proportion of women from the k^{th} shehia who completed secondary school education. The final term in the model is an interaction between the indicator for having had 4 ANC visits, and the indicator for having completed secondary school. In our hypothetical study, estimation of the parameter associated with this interaction term, β_6, is of primary interest.

3.6.3 Cluster-Based ODS

Nearly complete information on the covariates in model (3.4) are available for the 28789 women considered for the analysis. An exception is the education variable, which we multiply impute and subsequently take to be the true values for the remainder of this section. For the purpose of illustrating the optimal allocation and adaptive sampling strategies, we assume that complete information is available for the shehia-specific count of the number of women delivering outside of a health facility (outcome), the number of women from each of the shehias, the island of residence, as well as the highest level of education each woman completed. Each woman's age, and information on whether the woman received 4 ANC visits, however, we take to be unknown.

In order to operationalize a cluster-based ODS design, one could proceed by stratifying the shehias according to $Y^{0.80*}$, the 80^{th} quantile of the number of women delivering outside of a health facility, and region, X_{loc}. Doing so yields the 2×2 stratification of the K=280 shehias:

	X_{loc}=0	X_{loc}=1
$Y^* < Y^*_{0.80}$	134	87
$Y^* \geq Y^*_{0.80}$	21	38

Suppose that resources allow only $K_s = 80$ shehias to be visited for data collection of the missing covariate values: mother's age and whether the mother had 4 ANC visits. In the following sections we illustrate how one could determine the optimal allocation of the sample size across strata using i) the complete data set (though not feasible in practice, this serves as the gold standard for the design approximations), ii) the adaptive strategy using MI for estimation of the variance components, and iii) the adaptive strategy using IPW for estimation of the variance components.

3.6.4 Optimal Allocation

The optimal allocation design would yield the following stratum-specific sample sizes: (k_{00}, k_{10}, k_{01}, k_{11}) = $(27.88683, 27.37851, 14.52733, 10.20732)$. We note that the optimal allocation formulae will often yield non-integer solutions, and must rounded after edge cases have been appropriately handled. Under this scheme there is one edge case, as k_{10} = 27.37851 >$K_{10} = 21$. We set k_{10}=21 and recalculate the remaining stratum-specific sample sizes using an updated constraint:

$$K_s^* = K_s - 21 = 59$$

This yields (k_{00}, k_{10}, k_{01}, k_{11})=(31.26713, 21, 16.28827, 11.44460). There are no longer any edge cases, and the only step that remains is to round the k_j, which yields (k_{00}^r, k_{10}^r, k_{01}^r, k_{11}^r) = $(31, 21, 16, 12)$ using a rounding threshold of 0.28828.

3.6.5 Adaptive + IPW

Under the IPW approach, we again $K_{s,1}$ =40, so that the Stage 1 sample sizes are ($k_{00,1}$, $k_{10,1}$, $k_{01,1}$, $k_{11,1}$) = $(10, 10, 10, 10)$. Computing the Stage 2 sample sizes then yields ($k_{00,2}$, $k_{10,2}$, $k_{01,2}$, $k_{11,2}$) = $(16.647703, 21.715965, 3.512166, -1.875834)$. We have two edge cases here: $k_{10,1} + k_{10,2} = 21.715965 > K_{10} = 21$, and $k_{11,2} < 0$. We must therefore fix these edge cases at the boundary (set $k_{10,2}$=11 and set $k_{11,2} = 0$), and recalculate $k_{00,2}$ and $k_{01,2}$ using the updated constraint:

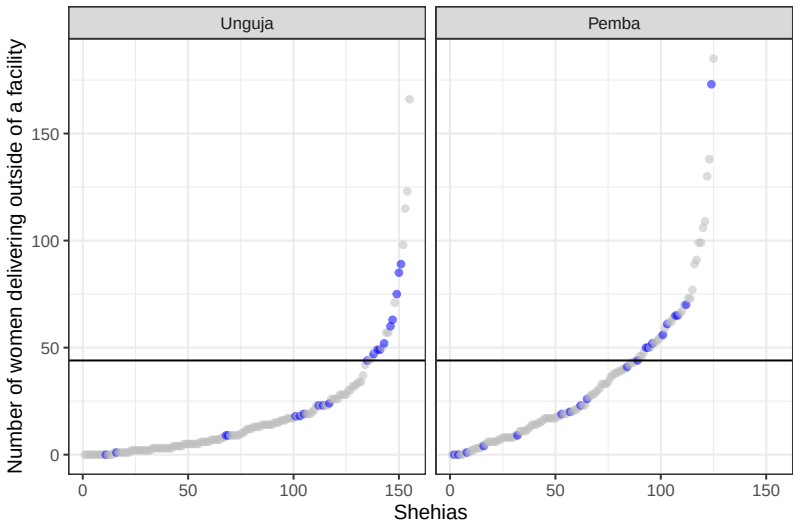

Figure 3.3: The number of women delivering outside of a health facility, in Unguja and Pemba. Each point represents one shehia; those shaded blue indicate the $K_{s,1} = 40$ shehias selected in the first wave.

$$K_s^* = K_s - 10 - 21 = 49$$

This results in $(k_{00,2}, k_{10,2}, k_{01,2}, k_{11,2}) = (22.513489, 11, 6.486511, 0)$ and yields overall stratum-specific sample sizes of $(k_{00}, k_{10}, k_{01}, k_{11}) = (32.51349, 21, 16.48651, 10)$. Finally, rounding with a rounding threshold of 0.5 gives us $(k^r_{00}, k^r_{10}, k^r_{01}, k^r_{11}) = (33, 21, 16, 10)$.

3.6.6 Adaptive + MI

Given a Stage 1 sample of size $K_{s,1} = 40$, we know, for the women sampled at Stage 1, the age of the mother and whether the woman had 4 ANC visits during pregnancy. For the women *not* sampled at Stage 1, we must impute these two covariates. First, we take a Stage 1 sample of size $K_{s,1} = 40$ clusters via balanced stratified sampling. Using the Stage 1 data, we impute M=5 complete datasets, and using these imputed datasets to compute the stratum-specific sample sizes, which gives $(k_{00,2}, k_{10,2}, k_{01,2}, k_{11,2}) = (13.9811422, 20.8349096, 4.6014676, 0.5824807)$. There is one edge case, as $k_{10,1} + k_{10,2} = 30.83491 > K_{10} = 21$. Setting $k_{10,2} = 11$, we recalculate the other Stage 2 sample

sizes using the updated constraint:

$$K_s^* = K_s - 21 = 59$$

The recalculation results in $(k_{00,2}, k_{10,2}, k_{01,2}, k_{11,2}) = (18.778293, 11, 7.522323, 2.699384)$. With a rounding threshold of 0.522324, the final stratum-specific sample sizes are $(k^r_{00}, k^r_{10}, k^r_{01}, k^r_{11}) = (29, 21, 17, 13)$.

3.6.7 Results

Table 3.3 shows results from the analysis of the data arising from the different sampling schemes. Looking at the complete data analysis (first column), we see that after adjusting for the other covariates in the model, women who had 4 ANC visits during pregnancy had a lower odds of delivering outside of a health facility compared to women who did not have 4 ANC visits, with a greater decrease among women who did not complete secondary school (OR=0.63, CI = 0.53, 0.74) than that among women who did complete secondary school (OR=0.94, CI=0.62, 1.43). Furthermore, the effect of having 4 ANC visits is statistically significant among women who did not complete secondary school, but is not statistically significant among women who did complete secondary school. On the other hand, older women had a higher odds of delivering outside of a health facility, as did women residing on the island of Pemba.

Figure B.6 shows the shehias plotted as a function of the island to which they belong and the number of women who delivered outside of a health facility; the blue points represent the shehias that were sampled at Stage 1. Table 3.4 summarizes the resulting designs under optimal allocation and the two adaptive sampling designs. In this case, both the IPW and the MI approach yield stratum-specific sample sizes that are quite close to the sample sizes one would get if complete information on all of the women in the dataset were available. In particular, displayed are the adjusted odds ratios and corresponding 95% confidence intervals using the complete dataset and samples

Table 3.3: Estimated odds ratios (OR) and 95% confidence intervals (CI) from IPW-GEE analysis, based on five samples drawn under five different sampling designs.

	Complete OR (CI)	**SRS** OR (CI)	**Balanced** OR (CI)
Intercept	0.18 (0.10, 0.33)	0.19 (0.07, 0.55)	0.16 (0.06, 0.42)
Region	2.02 (1.56, 2.60)	2.11 (1.24, 3.60)	2.16 (1.45, 3.23)
Age	1.10 (1.06, 1.13)	1.12 (1.06, 1.18)	1.12 (1.06, 1.18)
ANC4	0.63 (0.53, 0.74)	0.52 (0.39, 0.69)	0.54 (0.37, 0.77)
Ed	0.63 (0.55, 0.72)	0.58 (0.41, 0.81)	0.67 (0.52, 0.86)
Prop Ed	0.88 (0.76, 1.03)	0.90 (0.69, 1.16)	0.88 (0.69, 1.11)
ANC4 × Ed	1.51 (1.02, 2.23)	2.41 (0.85, 6.80)	1.56 (0.95, 2.56)

	Optimal OR (CI)	**2-stage/IPW** OR (CI)	**2-stage/MI** OR (CI)
Intercept	0.18 (0.08, 0.38)	0.18 (0.10, 0.32)	0.17 (0.09, 0.31)
Region	1.99 (1.31, 3.03)	1.80 (1.23, 2.65)	2.10 (1.39, 3.17)
Age	1.10 (1.04, 1.16)	1.11 (1.06, 1.17)	1.14 (1.08, 1.20)
ANC4	0.69 (0.53,0.89)	0.57 (0.45, 0.73)	0.61 (0.47, 0.80)
Ed	0.54 (0.43, 0.68)	0.65 (0.51, 0.84)	0.66 (0.51, 0.84)
Prop Ed	0.88 (0.73, 1.07)	0.86 (0.75, 1.00)	0.88 (0.76, 1.01)
ANC4 × Ed	1.52 (0.99, 2.34)	1.50 (0.96, 2.35)	1.49 (0.95, 2.35)

drawn via i) simple random sampling, ii) balanced stratified sampling, iii) the optimal allocation, iv) the adaptive strategy with IPW estimation, and v) the adaptive strategy with MI estimation.

We see that in general, the estimates of the adjusted odds ratios are consistent with the complete data analysis. Furthermore, the confidence intervals for the interaction term under the optimal and both adaptive sampling designs are substantially narrower than those under both the simple random sampling and balanced designs: compare (0.85, 6.80) under simple random sampling and (0.95, 2.56) under balanced sampling to (0.96, 2.35) under the adaptive sampling approach with IPW estimation, for example.

Table 3.4: Stratum specific sample sizes under optimal allocation, adaptive sampling with IPW estimation, and adaptive sampling with MI estimation.

	Strat00	Strat10	Strat01	Strat11
K_j	134	21	87	38
k_j				
Optimal	31	21	16	12
2-stage/IPW	33	21	16	10
2-stage/MI	29	21	17	13

3.7 Discussion

In this paper, we presented an adaptive sampling strategy that can be used to operationalize optimal allocation in single-stage stratified cluster-based ODS designs. In this context, optimal allocation was shown in Chapter 2 to yield efficiency gains for the parameter of interest. The adaptive sampling strategy presented in this paper allows for an optimal allocation design to be implemented in practice, by using the data collected at clusters sampled at Stage 1 to estimate the necessary components in the formulae used to compute the stratum-specific sample size, i.e. the k_j. Hence, the Stage 1 data can be thought of as an internal pilot study, which is more cost-effective than the alternative approach of using an external pilot study to estimate the k_j.

Results from our simulation study indicate that the adaptive sampling strategy works very well when multi-level multiple imputation is used to fill in the missing covariate values and subsequently estimate the k_js. When using the MI approach, a Stage 1 sample size of K_{s_1}=20 clusters, or 25% of the total number of K_s=80 clusters to be sampled, was sufficient to yield a design that is very close to the optimal allocation that would arise if complete knowledge of the data were available at the design stage. Necessary for good performance of the MI approach, however, is an imputation model that can correctly characterize the relationship between the covariate of interest and the stratification variables. In particular, if the parameter of interest is associated with a continuous covariate, and the variance of the covariate depends on the values of the stratification variable(s), we recom-

mend that the missing values be imputed within levels of the stratification variable separately.

The performance of the adaptive sampling strategy coupled with the IPW estimation approach was mixed. In contrast to the MI approach, a Stage 1 sample of 20 clusters resulted in substantial losses in efficiency compared to the optimal allocation design, most likely due to the high variability of IPW estimation given the small Stage 1 sample size. The performance improved by increasing the Stage 1 sample size to $K_{s,1} = 40$ or 60 clusters, and was better under the scenarios in which allocation was optimal with respect to a parameter associated with an individual-level covariate. Although there are instances in which a Stage 1 sample of 60 clusters yields a closer-to-optimal design than a Stage 1 sample of 40 clusters, we recommend sampling around 40 clusters at Stage 1, so as to have sufficient information for estimating the design components using IPW estimation, while avoiding sampling too many clusters via the non-optimal Stage 1 sampling strategy. Moreover, when the covariate associated with the parameter of interest is negatively associated with the stratification variable, we recommend sampling a smaller proportion of clusters from the strata corresponding to levels where the stratification variable is equal to 1.

In conclusion, optimal allocation in a stratified cluster-based ODS design can be a cost efficient strategy when there are finite resources for data collection. The obstacle to implementing optimal allocation in practice, however, is that the optimal allocation formulae depend on quantities that are unknown in practice. The adaptive sampling procedure we outline in this paper provides a practical strategy for implementing such a design. The specifics of the adaptive sampling strategy must be decided upon by researchers, taking into account the various factors described in this paper. While these methods correspond to a particular class of designs, i.e. single-stage stratified cluster-based ODS designs, it can be adapted to other settings, for example optimal allocation in two-stage stratified cluster-based ODS designs.

Acknowledgements

Ms. Sauer was funded by the Harvard T.H. Chan School of Public Health NIEHS-sponsored Environmental Training Grant T32ES007142. Dr. Haneuse was supported by NIH grant R01 HL094786. We are grateful to D-tree International and Zanzibar Ministry of Health for access to the Safer Deliveries (*Uzazi Salama*) program data. The data was collected as a byproduct of this global health program made possible through the generous support of the Saving Lives at Birth partners: the United States Agency for International Development (USAID), the Government of Norway, the Bill & Melinda Gates Foundation, Grand Challenges Canada, and the UK Government. It was prepared by the authors for illustrative purposes and does not reflect the views of the Saving Lives at Birth partners.

4

Conclusion

Chapters 1-3 focused on the design and analysis of single-stage cluster-based ODS designs. The results of Chapter 1 showed that given data that has been collected through a cluster-based ODS scheme, estimation and inference can be carried out using IPW-GEE. When the number of sampled clusters is small, researchers should apply small-sample bias corrections, particularly to the variance estimates. In Chapter 1, several small-sample corrections to the variance estimates were proposed, and in the settings considered, the Mancl and DeRouen-type[38,60] correction most consistently

yielded the closest to nominal coverage.

Chapter 2 of this dissertation focused on optimal allocation for single-stage stratified cluster-based ODS designs when the intended analysis method is IPW-GEE. Within this context, the optimal allocation design introduced in Chapter 2 was shown to yield efficiency gains for the parameter of interest over simple random sampling of clusters and balanced stratified sampling of clusters. In particular, optimal allocation for one parameter of interest was shown to be the most efficient design for the estimation of that parameter, but could result in losses for the other parameters in the model, depending on the relationship of the associated covariate with the covariate of interest. Optimal allocation for all parameters simultaneously using the A-optimality criterion generally resulted in more modest efficiency gains, but also smaller losses compared to the design optimizing for a single parameter.

The results of Chapter 2 showed that a wise selection of clusters can yield gains in statistical efficiency, even when the final analysis is at the level of the individual. Due to the fact that the optimal allocation formulae presented in Chapter 2 depend on quantities that are unknown in practice, the aim of Chapter 3 was to propose and evaluate an adaptive sampling strategy to to operationalize the optimal allocation. The adaptive sampling strategy involves sampling the clusters in two waves. The data collected from the individuals belonging to the clusters sampled in the first wave are treated as internal pilot data, that is used to estimate the components needed to optimally allocate the remaining resources. The adaptive sampling strategy was shown to yield a near-optimal design when using multi-level imputation in the estimation of the design components.

A natural next step is to consider optimal allocation in the context of *two-stage* stratified cluster-based ODS designs, again assuming that the intended analysis method is IPW-GEE. In the next few sections, we describe this setting in more detail, present some theory and initial work on this topic, and discuss gaps for future work.

4.1 Two-Stage Stratified Cluster-Based ODS Designs

Two-stage stratified cluster-based ODS can proceed as follows: at Stage I, cross-classify K clusters into J strata based on information contained in the previously defined $\mathcal{D}^*$. Then, k_j clusters are sampled from stratum j such that $\sum_j k_j = K_s$. At Stage II, the N_k individuals in cluster k are cross-classified into H strata based on information available on the individuals, for example the outcome Y. Then, n_{hk} individuals are sampled from stratum h_k, such that $\sum_{k \in s_I} \sum_{h_k \in k} n_{hk} = n$, where s_I is the set of clusters sampled at Stage I.

4.1.1 Estimation and Inference

When information is only available on a subset of individuals that have been selected through a two-stage stratified cluster-based ODS scheme, β can be estimated as the solution to the following weighted generalized estimating equations:

$$\mathcal{U}_w(\beta) = \sum_{k=1}^{K} \mathbf{D}_k^T \mathbf{V}_k^{-1} \mathbf{W}_{I,k} \text{diag}(\mathbf{R}_{I,k}) \mathbf{W}_{II,k} \text{diag}(\mathbf{R}_{II}) \boldsymbol{\varepsilon}_k = 0. \tag{4.1}$$

In the above, $\mathbf{W}_{I,k}$ is the $N_k \times N_k$ diagonal matrix with the Stage I weights for cluster k along the diagonal. Similarly, $\mathbf{W}_{II}$ is an $N_k \times N_k$ diagonal matrix, with entries along the diagonal equal to the Stage II weights for the individuals in cluster k. In both cases, the weights are equal to the inverse-probability of being sampled at the respective stages. Expression (4.1) can be rewritten as

$$\mathcal{U}_w(\beta) = \mathbf{U}^T \mathbf{W}_I \text{diag}(\mathbf{R}_I) \mathbf{W}_{II} \text{diag}(\mathbf{R}_{II}) 1_{N \times 1}, \tag{4.2}$$

where $\mathbf{W}_I$ and $\mathbf{W}_{II}$ are $N \times N$ diagonal matrices with the $\mathbf{W}_{I,k}$s and $\mathbf{W}_{II,k}$s, respectively, along the

diagonal. The variance of $\widehat{\beta}_w$ is given by the sandwich estimator

$$Var[\widehat{\beta}_w] = \mathbf{H}^{-1}(\mathbf{V}_0 + \mathbf{V}_I + \mathbf{V}_{II})\mathbf{H}^{-1}, \tag{4.3}$$

where $\mathbf{H}^{-1}\mathbf{V}_0\mathbf{H}^{-1}$ represents the variance that arises due to the observed population being a realization from the super-population, $\mathbf{H}^{-1}\mathbf{V}_I\mathbf{H}^{-1}$ represents the variance that arises due to sampling of clusters at the first stage, and $\mathbf{H}^{-1}\mathbf{V}_{II}\mathbf{H}^{-1}$ represents the variance that arises due to sampling of the individuals within the selected clusters.

4.1.2 Inference in Practice

To carry out inference in practice, $\text{Var}[\widehat{\beta}_w]$ can be replaced with $\widehat{\text{Var}}[\widehat{\beta}_w]$, where

$$\widehat{\text{Var}}[\widehat{\beta}_w]=\widehat{\mathbf{H}}^{-1}(\widehat{\mathbf{V}}_0 + \widehat{\mathbf{V}}_I + \widehat{\mathbf{V}}_{II})\widehat{\mathbf{H}}^{-1}.$$

In the expression above, $\widehat{\mathbf{H}} = \sum_{k=1}^{K} \mathbf{D}_k^T\mathbf{V}_k^{-1}\mathbf{W}_{I,k}\text{diag}(\mathbf{R}_{I,k})\mathbf{W}_{II,k}\text{diag}(\mathbf{R}_{II,k})\mathbf{D}_k$. Furthermore,

$$\widehat{\mathbf{V}_0} = \mathbf{U}^T\text{diag}(\mathbf{R}_t)\mathbf{W}_t\text{diag}(\mathbf{R}_t)\mathbf{U}$$

Where $\mathbf{W}_t$ is an $N \times N$ block diagonal matrix with entries in the k^{th} block equal to $1/\pi^k_{ii'}$, the joint probability of individuals i and i' in cluster k being sampled. This joint probability, $\pi^k_{ii'}$, is equal to $\frac{k_j}{K_j}\frac{n_{hk}}{N_{hk}}$ if $i = i'$; equal to $\frac{k_j}{K_j}\frac{n_{hk}}{N_{hk}}\frac{n_{hk}-1}{N_{hk}-1}$ if $i \neq i'$ are in the same stratum h_k; and, equal to $\frac{k_j}{K_j}\frac{n_{hk}}{N_{hk}}\frac{n_{h'k}}{N_{h'k}}$ if $i \neq i'$ are in different strata h_k and h'_k. The variance due to sampling clusters at Stage I can be estimated with

$$\widehat{\mathbf{V}}_I = \mathbf{U}^T\mathbf{W}_I\text{diag}(\mathbf{R}_I)\widetilde{\Delta}_I\text{diag}(\mathbf{R}_I)\mathbf{W}_I\mathbf{U}$$

where $\widetilde{\Delta}_I$ is an $N \times N$ matrix with entries all entries in the k^{th} $N_k \times N_k$ block along the diagonal equal to $\frac{\pi_k - \pi_k^2}{\pi_k}$, and entries in the off-diagonal blocks equal to $\frac{\pi_{kk'} - \pi_k\pi'_k}{\pi_{kk'}}$. If clusters k and k' are in

the same stratum j, then $\pi_{kk'} = \frac{k_j}{K_j}\frac{k_j-1}{K_j-1}$; if clusters k and k' are in different strata j and j', then $\pi_{kk'} = \frac{k_j}{K_j}\frac{k'_j}{K'_j}$. Finally, $\mathbf{V}_{II}$ can be estimated as

$$\widehat{\mathbf{V}}_{II} = \mathbf{U}^T\mathbf{W}_I^{1/2}\text{diag}(\mathbf{R}_I)\mathbf{W}_{II}\text{diag}(\mathbf{R}_{II})\widetilde{\mathbf{\Delta}}_{II}\text{diag}(\mathbf{R}_{II})\mathbf{W}_{II}\text{diag}(\mathbf{R}_I)\mathbf{W}_I^{1/2}\mathbf{U}$$

where $\widetilde{\mathbf{\Delta}}_{II}$ is an $N \times N$ block diagonal matrix with entries in the k^{th} block equal to $\frac{\pi_{ii'|k} - \pi_{i|k}\pi_{i'|k}}{\pi_k \pi_{ii'|k}}$.

4.1.3 Optimal Allocation for One Parameter of Interest

In this section we discuss optimal allocation for one parameter of interest under two different types of constraints. First, we consider the setting in which time and/or logistical constraints allow data collection on a total of $n <$ N individuals from $K_s <$ K clusters. In this context, we assume that an equal number of individuals, $n_k = n/K_s$, will be sampled from each of the selected clusters. Such a setting may arise if, for example, time constraints only permit one health center to be visited per day for data collection, and time constraints also require the determination of a sub-sample of individuals for detailed data collection within the selected health centers. The second constraint setting we consider arises when there is a cost associated with sampling clusters, and a different cost associated with collecting data on the individuals in these clusters. Given an overall budgetary constraint B, researchers must make a decision regarding *how many* clusters vs. *how many* individuals to sample, and subsequently *which* clusters and *which* individuals to sample.

As in Chapter 2, we suppose that primary interest lies in estimating the q^{th} parameter in a marginal mean model, β_q, with precision. The optimal allocation is therefore determined by minimizing the $[q,q]^{th}$ element along the diagonal of the variance-covariance matrix for $\widehat{\beta}$, Var$[\widehat{\beta}_w]$, given in expression (4.3). The first term of Var$[\widehat{\beta}_w]$, $\mathbf{H}^{-1}\mathbf{V}_0\mathbf{H}^{-1}$ does not depend on the sampling indicators, and therefore the optimal allocation problem becomes one of minimizing $\{\mathbf{H}^{-1}(\mathbf{V}_I + \mathbf{V}_{II})\mathbf{H}^{-1}\}_{[q,q]}$ subject to the relevant constraint, as described in the following two subsections. Before proceeding, note that the $[q,q]^{th}$ diagonal element of $\{\mathbf{H}^{-1}(\mathbf{V}_I + \mathbf{V}_{II})\mathbf{H}^{-1}\}_{[q,q]}$ can be expressed as:

$$\{\mathbf{H}^{-1}(\mathbf{V}_I + \mathbf{V}_{II})\mathbf{H}^{-1}\}_{[q,q]}$$
$$= \sum_{j=1}^{J} \frac{K_j - k_j}{k_j} C_{q,j}$$
$$+ \sum_{j=1}^{J} \frac{K_j}{k_j} \sum_{k \in j} \sum_{h \in k} \frac{(N_{hk} - n_{hk})}{n_{hk}} G_{q,hk}$$

where $C_{q,j}$ is as defined in Chapters 2 and 3, and $G_{q,hk} = [D_{q,hk} - \frac{F_{q,hk}}{N_{hk}-1}]$, where $D_{q,hk} = \sum_{i \in h_k} E[h_i^{[q]^2}]$ and $F_{q,hk} = \sum_{i \neq i' \in h_k} E[h_i^{[q]} h_{i'}^{[q]}]$.

Optimal Allocation: Fixed K_s and n, Equal n_k

In this setting, the constraints are $\sum_{j=1}^{J} k_j = K_s$ and $\sum_{hk=1}^{Hk} n_{hk} = n_k = \frac{n}{K_s}$ for each $k \in s_I$. We can define the Lagrangian

$$\mathcal{L}^1 := \sum_{j=1}^{J} \tfrac{K_j - k_j}{k_j} C_{q,j} + \sum_{j=1}^{J} \tfrac{K_j}{k_j} \sum_{k \in j} \sum_{h \in k} \tfrac{(N_{hk} - n_{hk})}{n_{hk}} G_{q,hk} + \lambda(\sum_{j=1}^{J} k_j - K_s) + \mu_k(\sum_{h_k \in k} n_{hk} - \tfrac{n}{K_s}).$$

Taking partial derivatives with respect to the k_j, n_{hk}, λ, and μ_k, setting these to 0 and solving the resulting system of equations yields

$$n_{hk} = \tfrac{n}{K_s} \frac{N_{hk}^{1/2} G_{q,hk}^{1/2}}{\sum_{h_k \in k} N_{hk}^{1/2} G_{q,hk}^{1/2}} \quad \text{and} \quad k_j = K_s \frac{K_j^{1/2} [C_{q,j} + M_{q,j}]^{1/2}}{\sum_{j=1}^{J} K_j^{1/2} [C_{q,j} + M_{q,j}]^{1/2}},$$

where $M_{q,j} = \sum_{k \in j} \sum_{h_k \in k} \frac{N_{hk} - n_{hk}}{n_{hk}} G_{q,hk}$. Note that the formula for k_j depends on the n_{hk} through $M_{q,j}$.

Optimal Allocation: Fixed Budgetary Constraint

With c_1 denoting the cost associated with sampling clusters, and c_2 denoting the cost associated with sampling individuals, the total cost of sampling can be expressed as:

$$B = \sum_{j=1}^{J} \left(c_1 k_j + c_2 \sum_{k \in s_{I_j}} \sum_{h_k \in k} n_{hk} \right) \tag{4.4}$$

The second term in 4.4 is random, as it depends on the set of sampled of clusters from stratum j, s_{I_j}. We therefore consider the *expected* cost:

$$B^* = \sum_{j=1}^{J} \left(c_1 k_j + c_2 \frac{k_j}{K_j} \sum_{k \in s_j} \sum_{h_k \in k} n_{hk} \right) \tag{4.5}$$

In this case, we define the Lagrangian as

$$\mathcal{L}^2 := \sum_{j=1}^{J} \frac{K_j - k_j}{k_j} C_{q,j} + \sum_{j=1}^{J} \frac{K_j}{k_j} \sum_{k \in s_{Ij}} \sum_{h_k \in k} \frac{(N_{hk} - n_{hk})}{n_{hk}} G_{q,hk} + \lambda \left[\sum_{j=1}^{J} \left(c_1 k_j + c_2 \frac{k_j}{K_j} \sum_{k \in s_{Ij}} \sum_{h_k \in k} n_{hk} \right) - B^* \right].$$

Taking partial derivatives with respect to the k_j, n_{hk}, and λ, setting these equal to 0, and solving the resulting system of equations, yields

$$k_j = B^* \frac{(K_j^{1/2} S_{q,j}^{1/2})/c_1^{1/2}}{\sum_{j=1}^{J} (c_1^{1/2} K_j^{1/2} S_{q,j}^{1/2} + c_2^{1/2} \sum_{k \in S_j} \sum_{h \in k} N_{hk}^{1/2} G_{hk}^{1/2})} \quad \text{and} \quad n_{hk} = \frac{K_j^{1/2} c_1^{1/2} N_{hk}^{1/2} G_{hk}^{1/2}}{S_{q,j}^{1/2} c_2^{1/2}},$$

where $S_{q,j} = C_{q,j} - \sum_{k \in s_{Ij}} \sum_{h_k \in k} G_{q,hk}$.

4.1.4 Practical Considerations

As in Chapter 2, the optimal allocation formulae presented in Sections **??** and 4.1.3, rely on quantities that are unknown in practice. The adaptive sampling strategy proposed in Chapter 3 can be

extended to this setting as well. If using multi-level imputation for estimation of the design components, care must be taken to use an imputation strategy that is valid when there is both systematically and sporadically missing data, as will be the case under two-stage stratified cluster-based ODS designs. Finally, we note that it is possible for $k_j > K_j$ and/or $n_{hk} > N_{hk}$ for $j = 1, \ldots, J$, $n_{hk} = 1, \ldots H_k$. Due to the potential large number of Stage I and Stage II strata, the approach to handling edge cases presented in Chapters 2 and 3 may not suffice here. Instead, the optimal Stage I and II stratum-specific sample sizes may be found using numerical methods. This is an area for future research.

4.2 Other Areas for Future Work

In addition to the extension of optimal allocation to two-stage stratified cluster-based ODS designs, there are a number of areas for future work in the context of single-stage cluster-based ODS, some of which have already been mentioned. First, it is known that careful stratification is also an important factor in determining the degree of statistical efficiency gains. Because only cluster-level summaries of the outcome are known at the design stage, a decision must be made regarding how to discretize the summary measures. More research is needed to help guide this decision. Second, we currently assume that the cluster-level summaries of the outcome and possibly other inherently individual-level covariates are measured without error, an assumption that is unlikely to hold in most public health research settings. For example, recent work looking at the quality of HMIS data for indicators related to maternal and newborn health showed that there is variability in the quality of HMIS data by health indicator[48]. An area for future research is therefore investigating the impact of the quality of the routinely collected data available at the design stage on efficiency gains. Third, we have assumed in this work that data is missing by design only. It will often be the case that there is data missing by happenstance as well. For example, when collecting detailed information on

individuals at selected health centers, there will likely be some data entries that are missing. An important area for future research is therefore how to adapt the optimal allocation formulae/strategy to accommodate the potential for missingness by happenstance as well.

A

Supplementary material to accompany Chapter 1

A.1 Asymptotic Theory for GEE

Before presenting the asympotic theory for WGEE, we first summarize the asymptotic theory results for GEE in the complete data setting, which was originally laid out by[81]. Given N individuals in

the population of interest, belonging to K clusters, let $N_k, k = 1, \ldots, K$ denote the number of individuals belonging to cluster k. In presenting the results for GEE and WGEE, we consider the case when $K \to \infty$ and m is bounded, where m= max$\{N_k; k = 1, \ldots, K\}$ is the maximum cluster size.

A.1.1 Notation

Let Y_{ki} denote the outcome for the i^{th} individual in the k^{th} cluster. We assume that the marginal density of Y_{ki} belongs to the exponential family, with density

$$f_{ki}(y_{ki}|X_{ki}, \beta, \varphi) = exp(\{y_{ki}\theta_{ki} - a(\theta_{ki}) + b(y_{ki})\}/\varphi),$$

with $\theta_{ki} = u(\eta_{ki})$, where u is an injective function and $\eta_{ki} = [X_{ki}]^T\beta$. β is a $(p \times 1)$ vector of unknown regression coefficients, φ is a nuisance scale paramater, $\mu_{ki} = E[Y_{ki}|\mathbf{X}_{ki}] = a'(\theta_{ki})$ and $\sigma^2_{ki} = Var(Y_{ki}|X_{ki}, \beta, \varphi) = a''(\theta_{ki})\varphi$. Let $\Theta = \{\theta|0 < \int exp(y\theta + b(y))dy < \infty\}$ be the natural parameter space of the exponential family. The interior of Θ is denoted by Θ^o. Letting $\mathbf{Y}_k$ be the $N_k \times 1$ vector of outcomes for all individuals in the k^{th} cluster, and denote $\mathbf{\Sigma}_k = Cov(\mathbf{Y}_k|\mathbf{X}_k, \beta, \varphi)$.[33] proposed estimating β by solving the following generalized estimating equations:

$$\mathcal{U}(\beta) = \sum_{k=1}^{K} D_k(\beta)^T\mathbf{V}_k(\beta)^{-1}(\mathbf{Y}_k - \mu_k(\beta)) = \sum_{k=1}^{K}\mathcal{U}_k(\beta), \tag{A.1}$$

where $\mathbf{D}_k = \partial\mu_k/\partial\beta$ is an $N_k \times p$ matrix of partial derivatives, $\mathbf{V}_k$ is an $N_k \times N_k$ working covariance matrix for $\mathbf{Y}_k$, and $\mu_k = (\mu_{k1}, \ldots, \mu_{kN_k})^T$. We define

$$\mathcal{U}_{Km}(\beta) = \frac{1}{K}\sum_{k=1}^{K}\mathbf{D}_k(\beta)^T\mathbf{V}_k(\beta)^{-1}(\mathbf{Y}_k - \mu_k(\beta)) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{U}_k(\beta), \tag{A.2}$$

where the sub-index Km is introduced to indicate that the estimating equations depend on the total

number of clusters and on the maximum cluster size. The following notation follows from above:

1. $\mathcal{D}_{Km} = -[\frac{\partial \mathcal{U}_{Km}(\beta)}{\partial \beta}]$
2. $\mathbf{M}_{Km} = Var(\mathcal{U}_{Km}) = \frac{1}{K^2}\sum_{k=1}^{K} \mathbf{D}_k(\beta)^T \mathbf{V}_k(\beta)^{-1} \mathbf{\Sigma}_k(\beta) \mathbf{V}_k(\beta)^{-1} \mathbf{D}_k(\beta)$
3. $\mathbf{H}_{Km} = -\frac{1}{K} E[\mathcal{D}_{Km}] = \frac{1}{K}\sum_{k=1}^{K} \mathbf{D}_k(\beta)^T \mathbf{V}_k(\beta)^{-1} bD_k(\beta)$
4. $\mathbf{F}_{Km} = \mathbf{H}_{Km}\mathbf{M}_{Km}^{-1}\mathbf{H}_{Km}$
5. Let $\bar{\mathbf{R}}_k$ denote the true correlation matrix, and $\mathbf{R}_k$ the working correlation matrix
6. Under working independence, $\mathbf{V}_k(\beta) = \mathbf{A}_k(\beta)$ is the diagonal matrix with the individual variances in the diagonal. With this, we have that
   $\mathbf{M}_{Km}(\beta) = \frac{1}{K^2}\sum_{k=1}^{K} \mathbf{D}_k(\beta)\mathbf{A}_k(\beta)^{-1/2}\bar{\mathbf{R}}_k\mathbf{A}_k(\beta)^{-1/2}\mathbf{D}_k(\beta)$
7. $\mathbf{V}_k = (\mathbf{A}_k)^{1/2}\mathbf{R}_k(\mathbf{A}_k)^{1/2}$, where $\mathbf{A}_k$ is the diagonal matrix with the individual variances on the diagonal
8. β_0 is the true regression parameter.

*Note that[81] use slightly different notation, with $g_{nm} = \sum_{i=1}^{n} \mathbf{D}_i(\beta)^T \mathbf{V}_i(\beta)^{-1}(\mathbf{Y}_i - \mu_i(\beta))$, with $n = K$ and $i = k$.

A.1.2 Regularity Assumptions

In the development of their results, Xie and Yang (2003) assume the following regularity conditions:

1. β is in an admissible set $\mathcal{B}$, where $\mathcal{B}$ is an open set of R^p
2. $\eta_{ki} = (\mathbf{X}_{ki})^T\beta \in g(\mathcal{M})$ for all $\beta \in \mathcal{B}$ and $\mathbf{X}_{ki} \in \mathcal{X}$, where $\mathcal{M}$ is the image of $a'(\Theta^0)$ and $\mathcal{X}$ is the set of all possible covariate variables.

3. $a'(\theta)$ is three times continuously differentiable and $a''(\theta) > 0$ in Θ^0. Also, $u(\eta)$ is three times continuously differentiable and $u'(\eta) > 0$ in $g(\mathcal{M})^0$.

4. $\mathbf{M}_{Km}$ and $\mathbf{H}_{Km}$ are positive definite when K or m are large.

A.1.3 Additional Conditions

In the development of their results,[81] also assume that the following conditions hold:

1. **Condition** I_w: The minimum eigenvalue of $\mathbf{F}_{Km}$, $\lambda_{min}(\mathbf{F}_{Km}) \to \infty$

2. **Condition** L_w: There exists a constant $c_0 > 0$, for any $r > 0$, such that

$$P(\mathcal{D}_{Km}^T \mathbf{M}_{Km}^{-1} \mathcal{D}_{Km} \geq c_0 \mathbf{F}_{Km} \text{ and } \mathcal{D}_{Km} \text{ is non-singular, for all } \beta \in B_{Km}(r)) \to 1,$$

   where $B_{Km}(r) = \{\beta : ||\mathbf{M}_{Km}^{-1/2} \mathbf{H}_{Km}(\beta - \beta_0)|| \leq r\}$

   The condition I_w, guarantees that $\mathbf{F}_{Km}$ diverges, which in turn implies that $\mathbf{F}_{Km}^{-1}$ converges. Condition L_w guarantees that if $\mathbf{F}_{Km}$ diverges, then so does $\mathcal{D}_{Km}^T \mathbf{M}_{Km}^{-1} \mathcal{D}_{Km}$. Note that I_w and L_w depend on $\mathbf{M}_{Km}$, a matrix that depends on the true correlation $\bar{\mathbf{R}}_k$, which is often unknown. The next two conditions I_w^* and L_w^* provide an alternative set of conditions that do not rely on $\bar{\mathbf{R}}_k$, only on $\mathbf{R}_k$.

3. **Condition** I_w^* : $(\tau_{Km})^{-1} \lambda_{min}(\mathbf{H}_{Km}) \to \infty$, where $\tau_{Km} = max_{k=1,\ldots,K}\{\lambda_{max}((\mathbf{R}_k)^{-1} \bar{\mathbf{R}}_k)\}$. This condition is stronger than I_w. Note that,

$$\mathbf{v}^T \mathbf{M}_{Km} \mathbf{v} =$$
$$\mathbf{v}^T (\mathbf{D}_k)^T (\mathbf{A}_k)^{-1/2} (\mathbf{R}_k)^{-1} (\mathbf{A}_k)^{-1/2} (\mathbf{A}_k)^{1/2} \bar{\mathbf{R}}_k (\mathbf{A}_k)^{1/2} (\mathbf{A}_k)^{-1/2} (\mathbf{R}_k)^{-1} (\mathbf{A}_k)^{-1/2} \mathbf{D}_k) \mathbf{v}$$

$$= \mathbf{v}^T (\mathbf{D}_k)^T (\mathbf{A}_k)^{-1/2} (\mathbf{R}_k)^{-1} \bar{\mathbf{R}}_k (\mathbf{R}_k)^{-1} (\mathbf{A}_k)^{-1/2} \mathbf{D}_k \mathbf{v}$$

$$\leq \tau_{Km}\mathbf{v}^T(\mathbf{D}_k)^T(\mathbf{S}_k)^{-1/2}\mathbf{R}_k^{-1}\mathbf{S}_k^{-1/2}\mathbf{D}_k)\mathbf{v}$$

$$= \tau_{Km}\mathbf{v}^T\mathbf{H}_{Km}v$$

*The eigenvalues of $\mathbf{M}_{Km}^{-1}\mathbf{H}_{Km}$ and $\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}\mathbf{M}_{Km}^{-1/2}$ are the same. The same applies to $\mathbf{H}_{Km}^{-1}\mathbf{M}_{Km}$ and $\mathbf{H}_{Km}^{-1/2}\mathbf{M}_{Km}\mathbf{H}_{Km}^{-1/2}$. Note that $\mathbf{M}_{Km} \leq \tau_{Km}\mathbf{H}_{Km}$ and $\mathbf{F}_{Km}^{-1} \leq \tau_{Km}\mathbf{H}_{Km}^{-1}$. This implies that $\lambda_{max}(\mathbf{F}_{Km}^{-1} - \tau_{Km}\mathbf{H}_{Km}^{-1}) < 0$, and that $\lambda_{max}(\mathbf{F}_{Km}^{-1}) \leq \tau_{Km}\lambda_{max}(\mathbf{H}_{Km}^{-1})$. Therefore $\lambda_{min}(\mathbf{F}_{Km}) > \lambda_{max}(\mathbf{H}_{Km}^{-1})/\tau_{Km} = \lambda_{min}(\mathbf{H}_{Km})/\tau_{Km}$. It follows that condition I_w^* implies I_w.

4. **Condition** L_w^*: There exists a constant $c_0 > 0$, for any $r > 0$, such that

$$P(\mathcal{D}_{Km}(\beta) \geq c_0\mathbf{H}_{Km} \text{ and } \mathcal{D}_{Km}(\beta) \text{ is nonsingular, for all } \beta \in B_{Km}^*(r)) \to 1$$

where $B_{Km}^*(r) = \{\beta : ||\mathbf{H}_{Km}^{1/2}(\beta - \beta_0)|| \leq (\tau_{Km})^{1/2}r\}$.

5. **Condition** CC: For any given $r > 0$ and $\delta > 0$,

$$P\left(\sup_{\beta \in B_{Km}^*(r)} ||\mathbf{H}_{Km}^{-1/2}\mathcal{D}_{Km}(\beta)\mathbf{H}_{Km}^{-1/2} - \mathbf{I}_{p\times p}|| < \delta\right) \to 1$$

where $B_{Km}^*(r) = \{\beta : ||\mathbf{H}_{Km}^{1/2}(\beta - \beta_0)|| \leq (\tau_{Km})^{1/2}r\}$ and the matrix norm is the Euclidean matrix norm. This condition is used to obtain the asymptotic distribution of the estimator $\widehat{\beta}$, by helping to establish the relationship between the asymptotic distributions of $\widehat{\beta}$ and $\mathcal{U}_{Km}$. This condition guarantees that for every $\beta \in B_{Km}^*(r)$, the matrix $\mathbf{H}_{Km}^{-1/2}\mathcal{D}_{Km}\mathbf{H}_{Km}^{-1/2}$ converges in probability to the identity matrix $\mathbf{I}_{p\times p}$.

6. **Condition** N_δ: Let $\mathbf{y}_k^* = (\mathbf{y}_{k1}^*, \ldots, \mathbf{y}_{kN^k}^*)^T = \mathbf{A}_k^{-1/2}(\mathbf{Y}_k - \boldsymbol{\mu}_k)$, $c_{Km} = \lambda_{max}(\mathbf{M}_{Km}^{-1}\mathbf{H}_{Km})$ and $\gamma_{Km}^{(D)} = max_{1\leq k\leq K}\lambda_{max}(\mathbf{H}_{Km}^{-1/2}(\mathbf{D}_k)^T(\mathbf{V}_k)^{-1}\mathbf{D}_k\mathbf{H}_{Km}^{-1/2})$. Then there exists a $\delta > 0$, such that $E[(\mathbf{y}_{kj_i}^*)^{2+(2/\delta)}]$ is uniformly bounded above, and

$$(c_{Km}m)^{1+\delta}\gamma_{Km}^{(D)} \to 0$$

A.1.4 Asympotic Results for GEE

Theorem 1 (Theorem 1,[81]). *Under conditions I_w and L_w, there exist a sequence of random variables $\widehat{\beta}_{Km} = \widehat{\beta}$, such that*

$$\mathcal{U}_{Km}(\widehat{\beta}) = 0 \text{ and } \widehat{\beta} \xrightarrow{\text{P}} \beta_0$$

[81] also proved that Theorem 1 holds when conditions I_w and L_w are replaced by I_w^* and L_w^*:

Theorem 2 (Theorem 2,[81]). *Under conditions I_w^* and L_w^*, there exist a sequence of random variables $\widehat{\beta}_{Km} = \widehat{\beta}$, such that*

$$\mathcal{U}_{Km}(\widehat{\beta}) = 0 \text{ and } \widehat{\beta} \xrightarrow{\text{P}} \beta_0$$

Towards proving the asymptotic normality of the estimator $\widehat{\beta}$,[81] first show that the asymptotic distribution of $\widehat{\beta}$ and $\mathcal{U}_{Km}$ are closely related:

Theorem 3 (Theorem 3,[81]) *Suppose that conditions I_w, L_w, and CC hold, or the conditions I_w and CC hold. Then, there exists a sequence of solutions $\widehat{\beta}$ to the GEE equation in $B_{Km}^*(r)$ such that $\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}(\widehat{\beta} - \beta_0)$ and $\mathbf{M}_{Km}^{-1/2}\mathcal{U}_{Km}$ are asymptotically identically distributed*

When the cluster size varies, this convergence is not guaranteed. However, it would be guaranteed when the Lindeberg condition holds:

Lemma 6 Lindeberg theorem,[7]). *Let $S_k = X_{m(K)1} + \ldots + X_{m(K)K}$, where $X_{m(K)1}, \ldots, X_{m(K)K}$ are independent variables with mean zero. Let $\sigma_{Km}^2 = E(X_{m(K)k}^2)$ and assume that $\sigma_K^2 = \sum_{k=1}^K \sigma_{Km}^2 > 0$. Additionally, assume that the following (Lindeberg condition) is satisfied:*

$$\lim_{K\to\infty} \sum_{k=1}^K \frac{1}{\sigma_K^2} \int_{|X_m(K)k| > \varepsilon\sigma_K} X_{m(K)k}^2 dP = 0 \text{ for all } \varepsilon > 0.$$

Then $S_K/\sigma_K \xrightarrow{d} N(0,1)$.

If we demonstrate that the Lindeberg theorem holds for every linear combination $\mathbf{v}^T\mathbf{M}_{Km}^{-1/2}\mathcal{U}_{Km}$, where $\mathbf{v}$ is a vector such that $||\mathbf{v}|| = 1$, this will imply, by Cramer-Wold Theorem, that $\mathbf{M}_{Km}^{-1/2}\mathcal{U}_{Km}$

converges to a multivariate normal random variable. However, as noted by Xie and Yang (2003), direct verification of this condition requires knowledge of the true correlation matrix $\bar{\mathbf{R}}_k$.

Lemma 7: (Lemma 2,[81]). *For $t > 0$, let $\psi(t)$ be a positive non-decreasing function such that $\lim_{t\to\infty} \psi(t) = \infty$ and $t\psi(t)$ is a convex funcion. Recall that $\mathbf{y}_k^* = (y_{k1}^*, \ldots, y_{kN^k}^*)^T = (\mathbf{A}_k)^{-1/2}(\mathbf{Y}_k - \boldsymbol{\mu}_k)$, $c_{km} = \lambda_{max}(\mathbf{M}_{Km}^{-1}\mathbf{H}_{Km})$ and $\gamma_{Km}^{(D)} =$*

$max_{1\le k\le K}\lambda_{max}(\mathbf{H}_{Km}^{-1/2}(\mathbf{D}_k)^T(\mathbf{V}_k)^{-1}\mathbf{D}_k\mathbf{H}_{Km}^{-1/2})$. Under the GEE setting, suppose there exist a constant K_0 (independent of K) and an integer m_0 such that, for $j = 1, \ldots, N_k$ and $k = 1, \ldots, K$, when $K > m_0$,

$$E[(y_{ki}^*)^2\psi((y_{ki}^*)^2)] \le K$$

In addition, for any $\varepsilon > 0$,

$$c_{Km}m\left[\psi\left(\frac{\varepsilon}{c_{Km}m\gamma_{Km}^{(D)}}\right)\right]^{-1} \to 0$$

Then when $K \to \infty$, we have

$$\mathbf{M}_{Km}^{-1/2}\mathcal{U}_{Km} \to N(0, \mathbf{I}_{p\times p}).$$

The following theorem establishes the asymptotic normality of $\widehat{\boldsymbol{\beta}}$:

Theorem 4 (Theorem 4,[81]) *Recall that $\mathbf{y}_k^* = (y_{k1}^*, \ldots, y_{kN_k}^*)^T = (\mathbf{A}_k)^{-1/2}(\mathbf{Y}_k - \boldsymbol{\mu}_k)$, $c_{Km} = \lambda_{max}(\mathbf{M}_{Km}^{-1}\mathbf{H}_{Km})$ and $\gamma_{Km}^{(D)} = max_{1\le k\le K}\lambda_{max}(\mathbf{H}_{Km}^{-1/2}(\mathbf{D}_k)^T(\mathbf{V}_k)^{-1}\mathbf{D}_k\mathbf{H}_{Km}^{-1/2})$. Suppose that the marginal distribution of each individual observation has a density of the form specified in Section 1.1. If condition (N_δ) is satisfied, then, when $K \to \infty$, we have,*

$$\mathbf{M}_{Km}^{-1/2}\mathcal{U}_{Km} \to N(0, \mathbf{I}_{p\times p}).$$

Further, under the conditions in Theorem 3, there exists a sequence of weakly consistent GEE estimators $\widehat{\boldsymbol{\beta}}$, and

$$\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(0, \mathbf{I}_{p\times p}).$$

A.2 Asymptotic Theory for WGEE

Moving away from the complete data setting, we now consider the setting where we do not have data on all of the N individuals from the K clusters in the population of interest. In some settings analysts may not have access to complete data on all elements of (Y, X) for all N individuals in the K clusters. They may, however, have access to resources that permit ascertainment of this information in a sub-sample of, say, $n < N$ individuals. Furthermore, they may have access to select components of (Y, X), as well as other variables/information that are not of direct relevance to the scientific question, denoted here by Z, that can, in principle, be used to make decisions regarding the sub-sampling. Moving forward we refer to this information as being available at the *design stage*.

Suppose the readily-available information is of the form $\mathcal{D}_k^* = (N_k, \mathbf{Y}_k^*, \mathbf{X}_k^*, \mathbf{Z}_k^*)$ where $\mathbf{Y}_k^*$ is a cluster-level summary of the outcomes (i.e. across the N_k individuals in the k^{th} cluster), $\mathbf{X}_k^*$ is a cluster-level feature or a cluster-level summary of elements of X that are readily-available at the outset, and $\mathbf{Z}_k^*$ is a cluster-level summary of Z. For example, if Y is binary, then $\mathbf{Y}_k^*$ may be the prevalence in the last six months at the health center. Furthermore, $\mathbf{X}_k^*$ may be a feature of the cluster, such as whether the health center is in a rural or urban setting or if it is private or publicly-funded, and/or it may be an aggregated summary of individual-level data, such as the percentage of mothers that are less than 18 years of age. Finally, $\mathbf{Z}_k^*$ may be the prevalence of some other outcome or comorbidity that is routinely collected.

The information represented by $\mathcal{D}_k^*$ can, in principle, be used to inform a cluster-based outcome-dependent sampling design. For this type of design, rather than selecting individuals directly, some sub-sample of $K_s < K$ clusters is initially selected. Then, the otherwise unavailable elements of X are ascertained on all individuals within the sampled clusters.

Let R_k be a binary indicator of whether the k^{th} cluster is selected and $\pi_k = \Pr(R_k = 1|\mathcal{D}^*)$ the corresponding probability of being selected, where $\mathcal{D}^* = \{\mathcal{D}_1^*, \ldots, \mathcal{D}_K^*\}$ is the totality of the

information available at the design stage. Towards operationalizing the sampling scheme, researchers may opt to use stratification or to use Poisson sampling. For the first of these, the clusters are cross-classified on the basis of some variable S that is defined on the basis of one or more of the variables contained in $\mathcal{D}^*$ and is assumed to take on one of J levels. From this, suppose K_j clusters are classified as belonging to the j^{th} stratum. We assume that the stratification scheme is specified such that $K_j > 0$ for all $j = 1, \ldots, J$. Then $k_j \leq K_j$ clusters are randomly selected from those in the j^{th} stratum, such that $\sum_{j=1}^{J} k_j = K_s$. Note, for each of the clusters in the j^{th} stratum, we have that $\pi_k = k_j/K_j$. For those that are selected in this way we set R_k=1. Under the second option of Poisson sampling, one first pre-specifies each of the π_k as a function of elements of $\mathcal{D}^*$. For example, one could specify a logistic regression model for R_k as a function of the clusters' outcome prevalence. Whether a cluster is selected by the design is then determined by an independent Bernoulli trial with probability π_k.

A.2.1 Notation

We have

$$\mathcal{U}_{k,w} = (\mathbf{D}_k)^T(\mathbf{V}_k)^{-1}\mathbf{W}_k diag(\mathbf{R}_k)(\mathbf{Y}_k - \boldsymbol{\mu}_k) \tag{A.3}$$

and define

$$\mathcal{U}_w = \mathcal{U}_{w,Km} = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{D}_k)^T(\mathbf{V}_k)^{-1}\mathbf{W}_k diag(\mathbf{R}_k)(\mathbf{Y}_k - \boldsymbol{\mu}_k) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{U}_{k,w} \tag{A.4}$$

where $\mathbf{R}_k$ is an $N_k \times 1$ vector with all entries equal to R_k and $\mathbf{W}_k$ is an $N_k \times N_k$ diagonal matrix with the all entries on the diagonal equal to $W_k = \frac{1}{\pi_k}$. Note that while stratification and sampling occur at the level of the cluster, so that R_k and $W_k = \frac{1}{\pi_k}$ are cluster-level quantities, every individual in cluster k inherits the values of R_k and W_k, as well as the stratum membership j of cluster k.

1. Let $U_{ji}^{k}, k = 1, \ldots, K, i = 1, \ldots, N_k$ be random variables. We use the sub-index k to indicate that the subject i from cluster k also belongs to a stratum j.

2. Let $\mathcal{F}_K = \{\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{S}\}$, where $\mathbf{S}$ is a $K \times 1$ vector with the k^{th} entry equal to the stratum j, $j = 1, \ldots, J$ that cluster k belongs to. We use the subscript K to indicate that the sample can be cluster-correlated.

3. $\mathcal{D}_{w,Km} = -[\frac{\partial \mathcal{U}_{w,Km}}{\partial \beta}]$

A.2.2 Conditions and Assumptions

1. We assume that $P(R|X, Y, Z) = P(R|V, Y)$ (MAR assumption).

2. Recall that we are considering the case where $K \to \infty$ and $m = max\{N_k; k = 1, \ldots, K\}$ is bounded, where N_k is the number of individuals in cluster k.

3. We assume that $K_s/K > c > 0$ and that $\lim_{N\to\infty} n = \infty$.

4. $\pi_k > c > 0, \quad \forall k = 1, .., K.$

5. Under a cluster-stratified design, let n_j be the number of individuals sampled in stratum j, let N_j denote the number of individuals in stratum j, k_j the number of clusters sampled from stratum j, and K_j the number of clusters in stratum j. We assume that $k_j/K_j > c > 0$, and that $\lim_{N\to\infty} n_j = \infty$, $\lim_{N\to\infty} n_j/N_j = f_{j,\infty}$, where $0 < f_{j,\infty} \leq 1$. Furthermore, we assume that $\lim_{N\to\infty} N_j/N = W_{j,\infty}$, where $0 < W_{j,\infty} < 1$, and that $\lim_{N\to\infty} K_j/N_j \to k_{j,\infty}$, where K_j is the number of clusters in stratum j. These conditions imply that when K increases, N_j increases but N_k does not.

6. We assume that each map $\mathcal{U}_{w,Km}$ is continuous.

7. $\mathcal{D}_{w,Km}$ is positive definite and non-singular.

8. $M_{w,Km} = Var(\mathcal{U}_{w,Km})$ is continuously differentiable.

9. Under assumptions in Xie and Yang (2003), we know that $\widehat{\beta}$ is consistent for β_0.

We will appeal to the following lemmas in developing the asymptotic theory for WGEE.

Lemma 1. (Lemma A,[14]) *Let H be a smooth injection from R^p to R^p with $H(x_0) = y_0$. Define $B_r(x_0) = \{x \in R^p, ||x - x_0|| < r\}$ and $S_r(x_0) = \partial B_r(x_0) = \{x \in R^p, ||x - x_0|| = r\}$. Then, $\inf_{x \in S_r(x_0)} ||H(x) - y_0|| \geq a$ implies*:

1. $B_a(y_0) = \{y \in R^p, ||y - y_0|| \leq a\} \subseteq H(B_r(x_0))$;

2. $H^{-1}(B_a(y_0)) \subseteq B_r(x_0)$.

Lemma 2. (Lemma 1,[81]) *Suppose $\mathbf{C}$ is a $p \times p$ matrix. For any $p \times 1$ vector $\mathbf{v}$, $||\mathbf{v}|| = 1$, we have $\mathbf{v}^T C^T \mathbf{C}\mathbf{v} \geq (\mathbf{v}^T \mathbf{C}\mathbf{v})^2$*

Definition 1. (Definition 13.1,[20]) *Given a sequence of finite populations, $\{\mathcal{F}_K\}$, and an associated sequence of sample designs, the estimator $\widehat{\theta}$ is design consistent for the finite population parameter θ_N, if for every $\varepsilon > 0$,*

$$\lim_{N \to \infty} P(|\widehat{\theta} - \theta_N| > \varepsilon | \mathcal{F}_K) \to 0,$$

where the notation indicates that for the sequence of finite populations, the probability is that determined by the sample design.

Lemma 3. (Lemma 5.10,[73]) *Let Θ be a subset of the real line and let Υ_K be random functions and Υ a fixed function of θ such that $\Upsilon_K(\theta) \xrightarrow{p} \Upsilon(\theta)$ for every θ. Assume that each map $\theta \to \Upsilon_K(\theta)$ is continuous and has exactly one zero $\widehat{\theta}$, or is non-decreasing with $\Upsilon_K(\widehat{\theta}) = 0_p(1)$. Let θ_0 be a point such that $\Upsilon(\theta_0 - \varepsilon) < 0 < (\Upsilon + \varepsilon)$ for every $\varepsilon > 0$. Then $\widehat{\theta} \xrightarrow{p} \theta_0$.*

Lemma 4a. (Theorem 1.3.3 (Fuller, 2009)). Let $u_1, u_2, \ldots$ be a sequence of real numbers and let $\pi_1, \pi_2, \ldots$ be a sequence of probabilities, with $0 < \pi_i < 1$. Let a Poisson sample be selected from the

population $\mathcal{F}_N = \{u_1, \ldots, u_N\}$, and let $\mathbf{g}_i = (1, u_i, \alpha_N \pi_i^{-1}, \alpha_N \pi_i^{-1} u_i)'$, where $\alpha_N = N^{-1} n_B$ and $n_B = E(n_N|N)$, where n_N is the final sample size, which is a function of N and is random. Assume that

$$\lim_{N\to\infty} n_B^{-1} \sum_{i=1}^{N} \mathbf{g}_i \pi_i = \mathbf{g}_\infty$$

$\lim_{N\to\inf} n_B^{-1} \sum_{i=1}^{N} \pi_i(1-\pi_i)\mathbf{g}_i\mathbf{g}_i' = \mathbf{g}_\infty = \Sigma_\mathbf{g}$, the submatrices of $\Sigma_\mathbf{g}$ associated with $(1, u_i, \alpha_N \pi_i^{-1})$ and $(\alpha_N \pi_i^{-1} u_i)$ are positive definite. Also assume that

$$\lim_{N\to\infty} \sup_{1\le k\le N} \left(\sum_{i=1}^{N} \pi_i(1-\pi_i)(\gamma'\mathbf{g}_i)^2 \right)^{-1} (\gamma'\mathbf{g}_i)^2 = 0$$

for every fixed vector γ' such that $\gamma\Sigma_{gg}\gamma > 0$. Let $\widehat{\boldsymbol{\mu}}_g = n_B^{-1} \sum_{i=1}^{N} R_i \mathbf{g}_i$ and $\boldsymbol{\mu}_{gN} = n_B^{-1} \sum_{i=1}^{N} \mathbf{g}\pi_i$ Then

$$n_B^{1/2}(\widehat{\boldsymbol{\mu}}_g - \boldsymbol{\mu}_{gN})|\mathcal{F}_N \xrightarrow{D} N(0, \Sigma_{gg}),$$

If, in addition, $\lim_{N\to\infty} n_B^{-1} \sum_{i=1}^{N} \pi_i |\mathbf{g}|^4 = M_g$, for some finite M_g, then

$$\widehat{V}(\widehat{t}_{\pi u}|\mathcal{F}_N)^{-1/2} (\widehat{t}_{\pi u} - t_u)|\mathcal{F}_N \xrightarrow{D} N(0, I),$$

where $\widehat{t}_{\pi u}$ is the Horvitz-Thompson estimator, $|\mathbf{g}_i| = (\mathbf{g}_i'\mathbf{g}_i)^{1/2}$ and $\widehat{V}(\widehat{t}_{\pi u}|\mathcal{F}_N) = \sum_s (1 - \pi_i)\pi_i^2 u_i \mathbf{g}_i'$.

Lemma 4b. (Corollary 1.3.4.1, [20]). *Let $\mathcal{F}_N$ be a sequence of populations, where the Nth population is composed of J strata with $\mathcal{F}_{jN} = \{u_{j1}, \ldots, u_{jN_j}\}; h = 1, \ldots, J$. Assume that $u_{j_i}, j = 1, \ldots, J; i = 1, \ldots, N_j$ are sequences of real numbers satisfying*

$$\lim_{N\to\infty} \tfrac{1}{N_j} \sum_{i=1}^{N_j} [u_{j_i}, (u_{j_i} - \bar{u}_j)^2, u_{j_i}^4] = [M_{1i,\infty}, S_{j,\infty}^2, M_{4j}, \infty],$$

where $M_{2j,\infty}, M_{4j,\infty}$ and $S_{i,\infty}^2$ are finite and positive. Then

$$Var(\widehat{u} - \bar{u}|\mathcal{F}_N)^{-1/2}(\widehat{u} - \bar{u}) \xrightarrow{\text{d}} N(0,1)$$

where $\widehat{u} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j}\frac{u_{ji}R_{ji}}{\pi_{ij}}$, $\bar{u} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j}u_{ji}$, *and*

$$Var(\widehat{u} - \bar{u}|\mathcal{F}_N) = \frac{1}{N^2}\sum_{j=1}^{J}\frac{N_j}{n_j}(N_j - n_j)S_j^2,$$

where $S_j^2 = \sum_{i=1}^{N_j}(u_{ji} - \bar{u}_j)^2/(N_j - 1)$ *and* $\bar{u}_j = \sum_{i=1}^{N_j}u_{ji}/N_j$.

*Below we give the details for showing that

$$Var(\widehat{u} - \bar{u}|\mathcal{F}_N) = \frac{1}{N^2}\sum_{j=1}^{J}\frac{N_j}{n_j}(N_j - n_j)S_j^2.$$

Note that

$$\begin{aligned} &Var(\widehat{u} - \bar{u}|\mathcal{F}_N) \\ &\qquad = Var(\widehat{u}|\mathcal{F}_N) = Var(\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j}\frac{u_{ji}R_{ji}}{\pi_{ij}}) \\ &\qquad = \frac{1}{N^2}\sum_{j=1}^{J}Var(\sum_{i=1}^{N_j}\frac{u_{ji}R_{ji}}{\pi_{ij}}) \end{aligned}$$

since the observations from different strata are independent. Now,

$$\begin{aligned} &\frac{1}{N^2}\sum_{j=1}^{J}Var(\sum_{i=1}^{N_j}\frac{u_{ji}R_{ji}}{\pi_{ij}}) \\ &\qquad = \frac{1}{N^2}\sum_{j=1}^{J}[\sum_{i=1}^{N_j}\frac{u_{ji}^2 Var(R_{ji})}{\pi_{ij}^2} + \sum_{i\neq i'}\frac{u_{ji}u_{ji'}Cov(R_{ji}, R_{ji'})}{\pi_{ij}\pi_{i'j}}] \end{aligned}$$

Note that $\pi_{ij} = \frac{n_j}{N_j}$ for all $i = 1, \ldots, N_j$. Furthermore, $Var(R_{ji}) = \pi_{ij}(1 - \pi_{ij}) = \frac{n_j}{N_j}(1 - \frac{n_j}{N_j})$ and that for individuals i and i' in stratum j, $Cov(R_{ji}, R_{ji'}) = \pi_{ii'j} - \pi_{ij}\pi_{i'j} = [\frac{n_j(n_j-1)}{N_j(N_j-1)} - (\frac{n_i}{N_j})^2]$. It then follows that

$$\frac{1}{N^2}\sum_{j=1}^{J} Var(\sum_{i=1}^{N_j}\frac{u_{j_i}R_{j_i}}{\pi_{ij}})$$

$$= \frac{1}{N^2}\sum_{j=1}^{J}[\sum_{i=1}^{N_j}\frac{u_{j_i}^2 Var(R_{j_i})}{\pi_{ij}^2} + 2\sum_{i<i'}\frac{u_{j_i}u_{j_{i'}}Cov(R_{j_i},R_{j_{i'}})}{\pi_{ij}\pi_{i'j}}]$$

$$= \frac{1}{N^2}\sum_{j=1}^{J}[\sum_{i=1}^{N_j}u_{j_i}^2\frac{N_j}{n_j}(1-\frac{n_j}{N_j}) + 2\sum_{i<i'}u_{j_i}u_{j_{i'}}(\frac{N_j(n_j-1)}{n_j(N_j-1)}-1)]$$

$$= \frac{1}{N^2}\sum_{j=1}^{J}\frac{N_j}{n_j}(N_j-n_j)[\sum_{i=1}^{N_j}\frac{u_{j_i}^2}{N_j} - 2\sum_{i<i'}\frac{u_{j_i}u_{j_{i'}}}{(N_j-1)N_j}]$$

$$= \frac{1}{N^2}\sum_{j=1}^{J}\frac{N_j}{n_j}(N_j-n_j)\frac{1}{N_j-1}[\sum_{i=1}^{N_j}\frac{u_{j_i}^2(N_j-1)}{N_j} - 2\sum_{i<i'}\frac{u_{j_i}u_{j'_i}}{N_j}]$$

$$= \frac{1}{N^2}\sum_{j=1}^{J}\frac{N_j}{n_j}(N_j-n_j)\frac{1}{N_j-1}[\sum_{i=1}^{N_j}u_{j_i}^2 - \frac{1}{N_j}\sum_{i=1}^{N_j}u_{j_i}^2 - \frac{2}{N_j}\sum_{i<i'}u_{j_i}u_{j'_i}]$$

$$= \frac{1}{N^2}\sum_{j=1}^{J}\frac{N_j}{n_j}(N_j-n_j)\frac{1}{N_j-1}[\sum_{i=1}^{N_j}u_{j_i}^2 - N_j\bar{u}_j^2]$$

$$= \frac{1}{N^2}\sum_{j=1}^{J}\frac{N_j}{n_j}(N_j-n_j)\frac{1}{N_j-1}\sum_{i=1}^{N_j}(u_{j_i}-\bar{u}_j)^2$$

$$= \frac{1}{N^2}\sum_{j=1}^{J}\frac{N_j}{n_j}(N_j-n_j)S_j^2$$

Lemma 5 (Theorem 1.3.6[20]). *Let $\{\mathcal{F}_N\}$ be a sequence of finite populations, let θ_N be a function of the elements of $\mathcal{F}_N$ and let the sequence of samples be selected from $\{\mathcal{F}_N\}$ by a design such that*

$$(\theta_N - \theta_N^0) \xrightarrow{\mathrm{d}} N(0, V_1)$$

Additionally, assume that

$$(\widehat{\theta} - \theta_N)|\mathcal{F}_N \xrightarrow{\mathrm{d}} N(0, V_2)$$

for a fixed sequence $\{\theta_N^0\}$ *and an estimator,* $\widehat{\theta}$*, where* $V_1 + V_2 > 0$*. Then,*

$$(V_1 + V_2)^{-1/2}(\widehat{\theta} - \theta_N^0) \xrightarrow{\mathrm{d}} N(0, I).$$

Remark 1a. Let $(U_1, U_1^*), \ldots, (U_N, U_N^*)$ be a sample of N variables. Under a Poisson sampling design, the design-based covariance of $\widehat{U}_w - \bar{U} = \frac{1}{K}\sum_{i=1}^N \frac{R_i U_i}{\pi_i} - \frac{1}{K}\sum_{i=1}^N U_i$ and $\widehat{U}_w^* - \bar{U}^* = \frac{1}{K}\sum_{i=1}^N \frac{R_i U_i^*}{\pi_i} - \frac{1}{K}\sum_{i=1}^N U_i^*$ is

$$Cov(\widehat{U}_w, \widehat{U}_w^*|(U_1, U_1^*), \ldots, (U_N, U_N^*)) = \frac{1}{K^2}\sum_{i=1}^N \frac{(1-\pi_i)}{\pi_i} U_i U_i^*$$

Similarly, the design-based variance of $\widehat{U}_W - \bar{U} = \frac{1}{K}\sum_{i=1}^N \frac{R_i U_i}{\pi_j} - \frac{1}{K}\sum_{i=1}^N U_i$, is

$$Var(\widehat{U}_w|U_1, \ldots, U_N) = \frac{1}{K^2}\sum_{i=1}^N \frac{(1-\pi_i)}{\pi_i} U_i^2.$$

Remark 1b. Let $(U_1, U_1^*), \ldots, (U_N, U_N^*)$ be a sample of N variables. Under a stratified sampling design, the design-based covariance of $\widehat{U}_w - \bar{U} = \frac{1}{K}\sum_{j=1}^J \sum_{i=1}^{N_j} \frac{R_{j_i} U_{j_i}}{\pi_{j_i}} - \frac{1}{K}\sum_{j=1}^J \sum_{i=1}^{N_j} U_{j_i}$ and $\widehat{U}_w^* - \bar{U}^* = \frac{1}{K}\sum_{j=1}^J \sum_{i=1}^{N_j} \frac{R_{j_i} U_{j_i}^*}{\pi_{j_i}} - \frac{1}{K}\sum_{j=1}^J \sum_{i=1}^{N_j} U_{j_i}^*$ is

$$Cov(\widehat{U}_w, \widehat{U}_w^*|(U_1, U_1^*), \ldots, (U_N, U_N^*)) = \frac{1}{K^2}\sum_{j=1}^J \frac{N_j(N_j - n_j)}{n_j} S_{uu^*j}$$

where $S_{uu^*j} = \sum_{i=1}^{N_j}(U_{j_i} - \bar{U}_j)(U_{j_i}^* - \bar{U}_j^*)/(N_j - 1)$ and $\bar{U}_j = \sum_{i=1}^{N_j} U_{j_i}/N_j$, $\bar{U}_j^* = \sum_{i=1}^{N_j} U_{j_i}^*/N_j$. Similarly, the design-based variance of $\widehat{U}_w - \bar{U} = \frac{1}{K}\sum_{j=1}^J \sum_{i=1}^{N_j} \frac{R_{j_i} U_{j_i}}{\pi_{j_i}} - \frac{1}{K}\sum_{j=1}^J \sum_{i=1}^{N_j} U_{j_i}$, is

$$Var(\widehat{U}_w|U_1, \ldots, U_N) = \frac{1}{K^2}\sum_{j=1}^J \frac{N_j}{n_j}(N_j - n_j)S_j^2,$$

where $S_j^2 = \sum_{i=1}^{N_j}(U_{j_i} - \bar{U}_j)^2/(N_j - 1)$ and $\bar{U}_j = \sum_{i=1}^{N_j} U_{j_i}/N_j$.

Remark 2a. : Let $U_1, \ldots, U_N$ be a sample of N independent variables with variance σ^2. If $\frac{1}{N}\sum_{i=1}^N \frac{(1-\pi_i)}{\pi_i} U_i^2 \to v_\infty$, then under a Poisson sampling design and using Chebychev inequality,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{R_i U_i}{\pi_i} - \frac{1}{N}\sum_{i=1}^{N} U_i | U_1, \ldots, U_N \xrightarrow{\text{p}} 0.$$

Chebychev inequality: $P(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2}$

Remark 2b. : Let $U_1, \ldots, U_N$ be a sample of N independent variables with variance σ^2. If $S_j^2 \rightarrow S_{j,\infty}^2$, then under a stratified sampling design and using Chebychev inequality,

$$\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j}\frac{R_{j_i} U_{j_i}}{\pi_{j_i}} - \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N_j} U_{j_i} | U_1, \ldots, U_N \xrightarrow{\text{p}} 0.$$

Chebychev inequality: $P(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2}$

Remark 3a. : Let $U_1, \ldots, U_N$ be a sample of N cluster-correlated independent variables. Under a Poisson sampling design, if $\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N_k}\frac{(1-\pi_k)}{\pi_k}(U_i^k)^2 \rightarrow v_\infty$, then using Chebyshev inequality,

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N_k}\frac{R_k U_i^k}{\pi_k} - \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N^k} U_i^k | \{U_i^k, \forall i, k\} \xrightarrow{\text{p}} 0.$$

Remark 3b. : Let $U_1, \ldots, U_N$ be a sample of N cluster-correlated independent variables. If $S_j^2(N_j - 1)/K \rightarrow S_{j,\infty}^2$, then using Chebyshev inequality,

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{N_j^k}\frac{R_{j_i}^k U_{j_i}^k}{\pi_{j_i}^k} - \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{N_j^k} U_{j_i}^k | \{U_{j_i}^k, \forall i, j, k\} \xrightarrow{\text{p}} 0.$$

A.2.3 Existence

We show the existence of the weighted estimator $\widehat{\beta}_w$ following a similar approach to that taken by[81] (see Theorem 1 of[81]).

Theorem 5. *Under the regularity assumptions, there exist a sequence of random variables $\widehat{\beta}_w$, such that*

$$P(\mathcal{U}_{w,Km}(\widehat{\beta}_w) = 0) \rightarrow 1.$$

Proof. Let $\mathbf{h}_{w,Km}(\beta)$=$\mathbf{M}_{Km}^{-1/2}(\beta)\mathcal{U}_{w,Km}(\beta)$ and $\mathbf{h}_{Km}(\beta) = \mathbf{M}_{Km}^{-1/2}(\beta)\mathcal{U}_{Km}(\beta)$, where $\mathcal{U}_{w,Km}(\beta) = \mathcal{U}_w(\beta)$ and $\mathcal{U}_{Km}(\beta) = \mathcal{U}(\beta)$. Let

$$\mathbf{V}_T = \mathbf{V}_1 + E[\mathbf{V}_2],$$

where $\mathbf{V}_2 = Var[\mathbf{h}_{w,Km}|\mathcal{F}_K]$ and $\mathbf{V}_1 = Var[\mathbf{h}_{Km}]$. For any $r > 0$, define

$$B_{w,Km}(r) = \{\beta : ||\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}(\beta - \beta_0)|| \leq r\},$$

and let

$$E_{w,Km} = \left\{\omega : r_{Km} \leq \inf_{\beta \in \partial B_{w,Km(r)}} ||\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}(T_{Km}(\beta) - T_{Km}(\beta_0))||\right\},$$

where $T_{Km}(\beta) = \mathbf{H}_{Km}^{-1}\mathcal{U}_{w,Km}(\beta), r_{Km} = ||\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}T_{Km}(\beta_0)||\}$ and $\partial B_{w,Km}(r)$ is the boundary of the sphere $B_{w,Km}(r)$. Under the regularity assumptions stated earlier, the mapping T_{Km} is continuously differentiable. Because $\mathcal{D}_{w,Km}$ is non-singular for $\beta \in B_{w,Km(r)}$, then T_{Km} is an injection from $B_{w,Km}(r)$ to $T_{Km}(B_{w,Km}(r))$. According to Lemma 1, on the set $E_{w,Km} \cap \{\mathcal{D}_{W,Km}(\beta)$ is nonsingular $\}$, there exist $\widehat{\beta}_w \in B_{w,Km}(r)$ such that $\mathcal{U}_{w,Km}(\widehat{\beta}_w) = 0$.

A.2.4 Consistency

Lemma 8. *Let $U_{ji}^k, k = 1, \ldots, K, i = 1, \ldots, N^k$ be bounded variables. Let $A = \{k : k = 1, ..., K\}$ the set of clusters at first phase. Let $a^k = \sum_{i=1}^{N_k} \frac{(1-\pi_k)}{\pi_k}(U_i^k)^2$. Assume that the first and second moments of a^k exist and are bounded. Assume that $\lim_{K\to\infty} \sum_{k\in A} E[a^k]/K \equiv L_{a,\infty}$ exists. Additionally, assume that $\mathbf{U}^k = [U_1^k, \ldots, U_{N^k}^k]^T, k = 1, \ldots, K$ are independent random vector-variables. Then, under a correlated setting under the conditions in the previous section,*

$$\sum_{k\in A} a^k/K \overset{a.s.}{\to} L_{a,\infty} \tag{A.5}$$

We need to prove that $\sum_{k\in A} a^k/K - \sum_{k\in A} E[a^k]/K \xrightarrow{\text{a.s.}} 0$ as $K \to \infty$. Using Kronecker's Lemma, it suffices to prove that $\lim_{K\to\infty} \sum_{k\in A}(a^k - E[a^k])/k < \infty$. Note that $Var[(a^k -$

$E[a^k])/k] \leq K_0 m^4/k^2$, where K_0 is a constant (for U_i^k bounded). Using the fact that the series $\sum_{k=1}^{K} 1/k^2$ converges, we have that $\sum_{k\in A_j} Var[(a_j^k - E[a_j^k])/k] < \infty$, and by the Khintchine-Kolmogorov Convergence Theorem, we have that $\sum_{k\in A_j}(a_j^k - E[a_j^k])/k < \infty$. Then, it follows that $\lim_{K\to\infty} \sum_{k\in A_j} a_j^k/K_j - \sum_{k\in A_j} E[a_j^k]/K_j = 0$. Recall that $\lim_{K\to\infty} \frac{1}{K}\sum_{k\in A} E[a^k]$ exists, then this implies that implies that equation (A.5) holds.

*

Lemma 8. *Let $U_{j_i}^k, k = 1,\ldots,K, i = 1,\ldots,N_k$ be bounded variables, and let $A_j = \{k|$ cluster k is in stratum $j\}$. Let $a_j^k = \sum_{i=1}^{N_k}(U_{j_i}^k)^2$ if $k \in A_j$, and $b_j^k = \sum_{i=1}^{N_k} U_{j_i}^k$ if $k \in A_j$. Define a_j^k and b_j^k as zero if stratum j does not contain cluster k, in other words if $k \notin A_j$. Assume that the first and second moments of a_j^k and b_j^k exist and are bounded. Let K_j be the number of clusters in stratum j and assume that $\lim_{K\to\infty} \sum_{k\in A_j} E[a_j^k]/K_j \equiv L_{a_j,\infty}$ and $\lim_{K\to\infty} \sum_{k\in A_j} E[b_j^k]/K_j \equiv L_{b_j,\infty}$ exist. Define $S_j^2 = \sum_{i=1}^{N_j}(U_{j_i} - \bar{U}_j)^2/(N_j - 1)$ and $\bar{U}_j = \sum_{i=1}^{N_j} U_{j_i}/N_j$. Additionally, assume that $U^k = [U_1^k,\ldots,U_{N^k}^k]^T, k = 1,\ldots,K$ are independent random vector-variables. Then, under a correlated setting and under the conditions in the previous section,*

$$S_j^2(N_j - 1)/K_j \xrightarrow{\text{a.s.}} L_{a_j,\infty} - k_{j,\infty} L_{b_j,\infty}^2.$$

where $k_{j,\infty} = \lim_{K\to\infty} K/N_j$.

Let a_j^k and b_j^k be defined as above. Note that $(N_j - 1)S_j^2 = \sum_{i=1}^{N_j} U_{j_i}^2 - 2\bar{U}_j \sum_{i=1}^{N_j} U_{j_i} + N_j\bar{U}_j^2 = \sum_{i=1}^{N_j} U_{j_i}^2 - N_j\bar{U}_j^2 = \sum_{k\in A_j}\sum_{i=1}^{N^k}(U_{j_i}^k)^2 - \frac{1}{N_j}(\sum_{k\in A_j}\sum_{i=1}^{N^k} U_{j_i}^k)^2$, so

$$(N_j - 1)S_j^2 = \sum_{k\in A_j} a_j^k - \frac{1}{N_j}\left(\sum_{k\in A_j} b_j^k\right)^2.$$

We therefore need to prove that $\sum_{k\in A_j} a_j^k/K_j - \sum_{k\in A_j} E[a_j^k]/K_j \xrightarrow{\text{a.s.}} 0$ as $K \to \infty$ and that $\sum_{k\in A_j} b_j^k/K_j - \sum_{k\in A_j}^{K_j} E[b_j^k]/K_j \xrightarrow{a.s} 0$ as $K \to \infty$. Using Kronecker's Lemma, it suffices to prove

* *Khintchine-Kolmogorov Convergence Theorem: Suppose $X_1, X_2, \ldots$, are independent with mean o such that $\sum_n Var(X_n) < \infty$. Then $\sum_n X_n < \infty$ a.s.* By Kronecker's Lemma, to prove $\lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K} x_k = 0$ a.s., it suffices to show that the random series $\sum_{k=1}^{\infty} x_k/k$ converges a.s. Proving that a series converges is usually easier than proving that a sequence converges to a specific limit.

that $lim_{K\to\infty}\sum_{k\in A_j}(a_j^k - E[a_j^k])/k < \infty$. Note that $Var[(a_j^k - E[a_j^k])/k] \leq K_0 m^4/k^2$, where K_0 is a constant. Using the fact that the series $\sum_{k=1}^{K} 1/k^2$ converges, we have that $\sum_{k\in A_j} Var[(a_j^k - E[a_j^k])/k] < \infty$, and by the Khintchine-Kolmogorov Convergence Theorem, we have that $\sum_{k\in A_j}(a_j^k - E[a_j^k])/k < \infty$. Then by Kronecker's Lemma, it follows that $\lim_{K\to\infty}\sum_{k\in A_j} a_j^k/K_j - \sum_{k\in A_j} E[a_j^k]/K_j = 0$. Similarly, note that $Var[(b_j^k - E[b_j^k])/k] \leq K_0 m^4/k^2$. The series, $\sum_{k=1}^{\infty} 1/k^2$ converges and by the Khintchine-Kolmogorov Convergence Theorem, $\sum_{k\in A_j}(b_j^k - E[b_j^k])/k < \infty$, so by Kronecker's Lemma it follows that $\lim_{K\to\infty}\sum_{k\in A_j} b_j^k/K_j - \sum_{k\in A_j} E[b_j^k]/K_j = 0$. Recall that we assume that $\lim_{K\to\infty} K/N_j = k_{j,\infty}$ and that $\lim_{K\to\infty}\frac{1}{K_j}\sum_{k\in A_j} E[a_j^k]$ and $\lim_{K\to\infty}\frac{1}{K_j}\sum_{k\in A_j} E[b_j^k]$ exist, then this implies that

$$S_j^2(N_j-1)/K_j \xrightarrow{\text{a.s.}} \lim_{K\to\infty}\frac{1}{K_j}\sum_{k\in A_j} E[a_j^k] - \lim_{K\to\infty}\frac{K_j}{N_j}\left(\frac{1}{K_j}\sum_{k\in A_j} E[b_j^k]\right)^2.$$

i.e.

$$S_j^2(N_j-1)/K_j \xrightarrow{\text{a.s.}} L_{a_j,\infty} - k_{j,\infty}(L_{b_j,\infty})^2.$$

Theorem 6a. Let $\mathbf{L} = \mathbf{D}^T\mathbf{V}^{-1}$, where $\mathbf{L}$ is a $p \times N$ matrix and each vector-row l of $\mathbf{L}$ is of the form $\mathbf{L}_l = [D_{1l}V_{11}^{-1}, \ldots, D_{Nl}V_{NN}^{-1}], l = 1, \ldots, p$, where V_{ii}^{-1} is the $ii-th$ element of the matrix $\mathbf{V}^{-1}$. Let $A = \{k : k = 1, \ldots K\}$. The term $\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l$ can be written as $\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l = \frac{1}{K}\sum_{k\in A}\sum_{i=1}^{N_k} R_k w_k U_{i,l}^k - \mathcal{U}_{Km}(\beta)_l$, where $U_{i,l}^k(\beta) = L_i^k(Y_i^k - \mu_i^k(\beta))$. Let $a^k = \sum_{i=1}^{N_k}(U_{i,l}^k)^2$. Assume that first and second moments of a^k and b^k exist and are bounded for all β, and that $\lim_{K\to\infty}\sum_{k\in A} E[a^k]/K$ exists a.s. Then,

$$\widehat{\beta}_w \xrightarrow{\text{p}} \beta_0.$$

We first prove that the WGEE consistently estimates $\widehat{\beta}$ under a design-based approach (Fuller, 2009). Based on this, we then prove that $\widehat{\beta}_w$ is consistent for β_0. The existence of the limits above

is relevant because they are sufficient to prove that the design-based variance of the estimating equations converges to 0. Each entry of $\mathcal{U}_{w,Km}(\beta)$ can be written as $\mathcal{U}_{w,Km}(\beta)_l = \frac{1}{K}\sum_{j=1}^{J}\sum_{k\in A}\sum_{i=1}^{N_k} R_k w_k U^k_{i,l}$, where $U^k_{i,l} = L^k_{i,l}(Y^k_i - \mu^k_i)$. Then,

$$Var(\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l|\mathcal{F}_K) = \frac{1}{K^2}\sum_{k\in A}\sum_{i=1}^{N_k}\frac{1-\pi_k}{\pi_k}\left(U^k_{i,l}\right)^2.$$

Under the assumptions in Lemma 8, $Var(\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l|\mathcal{F}_K) \xrightarrow{\text{a.s.}} 0$ and $E[Var[\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l|\mathcal{F}_K]] \xrightarrow{\text{a.s.}} 0$. By Chebychev's inequality,

$$\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l|\mathcal{F}_K \xrightarrow{\text{p}} 0$$

and

$$\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l \xrightarrow{\text{p}} 0$$

As stated before, we assume that $\mathcal{U}_{w,Km}(\beta)_l$ is continuous and has exactly one zero at $\widehat{\beta}_w$. We also know that $\mathcal{U}_{Km}(\beta)_l = E[\mathcal{U}_{w,Km}(\beta)_l|\mathcal{F}_K]$ and because it is continuous and has one zero at $\widehat{\beta}$, then $\mathcal{U}_{Km}(\widehat{\beta}_{-\varepsilon_l}) < 0 < \mathcal{U}_{Km}(\widehat{\beta}_{+\varepsilon_l})$ for every $\varepsilon > 0$,where $\widehat{\beta}_{-\varepsilon_l} = [\widehat{\beta}_1, \ldots, \widehat{\beta}_l - \varepsilon, \ldots, \widehat{\beta}_p]$ and $\widehat{\beta}_{+\varepsilon_l} = [\widehat{\beta}_1, \ldots, \widehat{\beta}_l + \varepsilon, \ldots, \widehat{\beta}_p]$. Therefore

$$\widehat{\beta}_w - \widehat{\beta}|\mathcal{F}_K \xrightarrow{\text{p}} 0 \tag{A.6}$$

Now, note that

$$||\widehat{\beta}_w - \beta_0|| = ||\widehat{\beta}_w - \widehat{\beta} + \widehat{\beta} - \beta_0|| \leq ||\widehat{\beta}_w - \widehat{\beta}|| + ||\widehat{\beta} - \beta_0||$$

Then,

$$\begin{aligned} P\left(||\widehat{\beta} - \beta_0|| > \varepsilon\right) &\leq P\left(||\widehat{\beta}_w - \widehat{\beta}|| + ||\widehat{\beta} - \beta_0|| > \varepsilon\right) = E\left[P\left(||\widehat{\beta}_w - \widehat{\beta}|| + ||\widehat{\beta} - \beta_0|| > \varepsilon \middle| \mathcal{F}_K\right)\right] \\ &= E\left[P\left(||\widehat{\beta}_w - \widehat{\beta}|| > \varepsilon - ||\widehat{\beta} - \beta_0|| \middle| \mathcal{F}_K\right)\right] = E\left[P\left(||\widehat{\beta}_w - \widehat{\beta}|| > \varepsilon_1 \middle| \mathcal{F}_K\right)\right] \end{aligned}$$

where $\varepsilon_1 = \varepsilon - ||\widehat{\beta} - \beta_0||$. Since $\widehat{\beta}_w - \widehat{\beta}|\mathcal{F}_K \xrightarrow{p} 0$, it follows that $\lim_{K,n\to\infty} P\left(||\widehat{\beta}_w - \widehat{\beta}|| > \varepsilon \middle| \mathcal{F}\right) = 0, \forall \varepsilon > 0$. Then,

$$\begin{aligned}\lim_{K,n\to\infty} P\left(||\widehat{\beta}_w - \beta_0|| > \varepsilon\right) &\leq \lim_{K,n\to\infty} E\left[P\left(||\widehat{\beta}_w - \widehat{\beta}|| > \varepsilon_1 \middle| \mathcal{F}_K\right)\right] \\ &= E\left[\lim_{K,n\to\infty} P\left(||\widehat{\beta}_w - \widehat{\beta}|| > \varepsilon_1 \middle| \mathcal{F}_K\right)\right] = 0.\end{aligned}$$

Therefore, $\widehat{\beta}_w \xrightarrow{p} \beta_0$.

Theorem 6b. Let $\mathbf{L} = \mathbf{D}^T\mathbf{V}^{-1}$, where $\mathbf{L}$ is a $p \times N$ matrix and each vector-row l of $\mathbf{L}$ is of the form $\mathbf{L}_l = [D_{1l}V_{11}^{-1}, \ldots, D_{Nl}V_{NN}^{-1}], l = 1, \ldots, p$, where V_{ii}^{-1} is the $ii-th$ element of the matrix $\mathbf{V}^{-1}$. Let $A_j = \{k : \text{cluster } k \text{ in stratum } j\}$. The term $\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l$ can be written as $\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l = \frac{1}{K}\sum_{j=1}^{J}\sum_{k\in A_j}\sum_{i=1}^{N_k} R_k w_k U_{ji,l}^k - \mathcal{U}_{Km}(\beta)_l$, where $U_{ji,l}^k(\beta) = L_{ji}^k(Y_{ji}^k - \mu_{ji}^k(\beta))$. Let $a_j^k = \sum_{i=1}^{N_k}(U_{ji,l}^k)^2$ and $b_j^k = \sum_{i=1}^{N_k} U_{ji,l}^k$ if cluster k is in stratum j and o if stratum cluster k is not in stratum j. Assume that first and second moments of a_j^k and b_j^k exist and are bounded for all β, and that $\lim_{K\to\infty}\sum_{k\in A_j} E[a_j^k]/K_j$ and $\lim_{K\to\infty}\sum_{k\in A_j} E[b_j^k]/K_j$ exist a.s., where K_j is the number of clusters in stratum j. Then,

$$\widehat{\beta}_w \xrightarrow{p} \beta_0.$$

We first prove that the WGEE consistently estimates $\widehat{\beta}$ under a design-based approach [20]. Based on this, we then prove that $\widehat{\beta}_w$ is consistent for β_0. The existence of the limits above is relevant because they are sufficient to prove that the design-based variance of the estimating equations converges to 0. Each entry of $\mathcal{U}_{w,Km}(\beta)$ can be written as $\mathcal{U}_{w,Km}(\beta)_l = \frac{1}{K}\sum_{j=1}^{J}\sum_{k\in A_j}\sum_{i=1}^{N_k} R^k w^k U_{ji,l}^k$, where $U_{ji,l}^k = L_{ji,l}^k(Y_{ji}^k - \mu_{ji}^k)$. Then,

$$Var(\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l|\mathcal{F}_K) = \frac{1}{K^2}\sum_{j=1}^{J}\frac{N_j}{n_j}(N_j - n_j)S_{jl}^2,$$

where $S_{jl}^2 = \sum_{k\in A_j}\sum_{i=1}^{N_k}(U_{j_i,l}^k - \bar{U}_{j,l})^2/(N_j-1)$, and $\bar{U}_{j,l} = \sum_{k\in A_j}\sum_{i=1}^{N_k} U_{j_i,l}/N_j$. Under the assumptions in Lemma 8,

$$S_{jl}^2(N_j-1)/K_j \xrightarrow{\text{a.s.}} \lim_{K\to\infty}\tfrac{1}{K_j}\sum_{k\in A_j}E[a_j^k] - k_{j,\infty}\left(\lim_{K\to\infty}\tfrac{1}{K_j}\sum_{k\in A_j}E[b^k]\right)^2.$$

Therefore $Var(\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l|\mathcal{F}_K) \xrightarrow{\text{a.s.}} 0$ and $E[Var[\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l|\mathcal{F}_K]] \xrightarrow{\text{a.s.}} 0$. By Chebychev's inequality,

$$\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l|\mathcal{F}_K \xrightarrow{\text{p}} 0$$

and

$$\mathcal{U}_{w,Km}(\beta)_l - \mathcal{U}_{Km}(\beta)_l \xrightarrow{\text{p}} 0$$

As stated before, we assume that $\mathcal{U}_{w,Km}(\beta)_l$ is continuous and has exactly one zero at $\widehat{\beta}_w$. We also know that $\mathcal{U}_{Km}(\beta)_l = E[\mathcal{U}_{w,Km}(\beta)_l|\mathcal{F}_K]$ and because it is continuous and has one zero at $\widehat{\beta}$, then $\mathcal{U}_{Km}(\widehat{\beta}_{-\varepsilon_l}) < 0 < \mathcal{U}_{Km}(\widehat{\beta}_{+\varepsilon_l})$ for every $\varepsilon > 0$,where $\widehat{\beta}_{-\varepsilon_l} = [\widehat{\beta}_1,\ldots,\widehat{\beta}_l - \varepsilon,\ldots,\widehat{\beta}_p]$ and $\widehat{\beta}_{+\varepsilon_l} = [\widehat{\beta}_1,\ldots,\widehat{\beta}_l + \varepsilon,\ldots,\widehat{\beta}_p]$. Therefore

$$\widehat{\beta}_w - \widehat{\beta}|\mathcal{F}_K \xrightarrow{\text{p}} 0 \tag{A.7}$$

Now, note that

$$||\widehat{\beta}_w - \beta_0|| = ||\widehat{\beta}_w - \widehat{\beta} + \widehat{\beta} - \beta_0|| \leq ||\widehat{\beta}_w - \widehat{\beta}|| + ||\widehat{\beta} - \beta_0||$$

Then,

$$
\begin{aligned}
&P\Big(||\widehat{\beta} - \beta_0|| > \varepsilon\Big) \\
&\qquad \leq P\Big(||\widehat{\beta}_w - \widehat{\beta}|| + ||\widehat{\beta} - \beta_0|| > \varepsilon\Big) \\
&\qquad = E\Big[P\Big(||\widehat{\beta}_w - \widehat{\beta}|| + ||\widehat{\beta} - \beta_0|| > \varepsilon \Big| \mathcal{F}_K\Big)\Big] \\
&\qquad = E\Big[P\Big(||\widehat{\beta}_w - \widehat{\beta}|| > \varepsilon - ||\widehat{\beta} - \beta_0|| \Big| \mathcal{F}_K\Big)\Big] \\
&\qquad = E\Big[P\Big(||\widehat{\beta}_w - \widehat{\beta}|| > \varepsilon_1 \Big| \mathcal{F}_K\Big)\Big]
\end{aligned}
$$

where $\varepsilon_1 = \varepsilon - ||\widehat{\beta} - \beta_0||$. Since $\widehat{\beta}_w - \widehat{\beta}|\mathcal{F}_K \xrightarrow{\text{p}} 0$, it follows that $\lim_{K,n\to\infty} P\Big(||\widehat{\beta}_w - \widehat{\beta}|| > \varepsilon \Big| \mathcal{F}\Big) = 0$ $\forall \varepsilon > 0$. Then,

$$
\begin{aligned}
&\lim_{K,n\to\infty} P\Big(||\widehat{\beta}_w - \beta_0|| > \varepsilon\Big) \\
&\qquad \leq \lim_{K,n\to\infty} E\Big[P\Big(||\widehat{\beta}_w - \widehat{\beta}|| > \varepsilon_1 \Big| \mathcal{F}_K\Big)\Big] \\
&\qquad = E\Big[\lim_{K,n\to\infty} P\Big(||\widehat{\beta}_w - \widehat{\beta}|| > \varepsilon_1 \Big| \mathcal{F}_K\Big)\Big] = 0.
\end{aligned}
$$

Therefore, $\widehat{\beta}_w \xrightarrow{\text{p}} \beta_0$.

A.2.5 Asymptotic Distribution

In this section, we demonstrate the asymptotic distribution of $\mathcal{U}_{w,Km}(\beta_0)$, which implies the asymptotic normality of $\widehat{\beta}_w$, given that the conditions outlined below hold.

A.2.6 Additional Notation

Let

$$\mathbf{h}_{w,Km}(\beta) = \mathbf{M}_{Km}^{-1/2}\mathcal{U}_{w,Km}(\beta)$$

$$\mathbf{h}_{Km}(\beta) = \mathbf{M}_{Km}^{-1/2}\mathcal{U}_{Km}(\beta)$$

where $\mathcal{U}_{Km}(\beta) = \mathcal{U}(\beta)$ and $\mathcal{U}_{w,Km}(\beta) = \mathcal{U}_w(\beta)$. Let $\mathbf{v}$ be a vector of size p such that $||\mathbf{v}|| = 1$ and let

$$t_{Km}(\beta) = \mathbf{v}'\mathbf{V}_T^{-1/2}\mathbf{h}_{Km}(\beta)$$

$$t_{w,Km}(\beta) = \mathbf{v}'\mathbf{V}_T^{-1/2}(\mathbf{h}_{w,Km}(\beta) - \mathbf{h}_{Km}(\beta)),$$

where

$$\mathbf{V}_T(\beta) = \mathbf{V}_1(\beta) + E[\mathbf{V}_2(\beta)]$$

$$\mathbf{V}_2(\beta) = Var[\mathbf{h}_{w,Km}(\beta)|\mathcal{F}_K]$$

$$\mathbf{V}_1(\beta) = Var[\mathbf{h}_{Km}(\beta)]$$

Furthermore, let $\sigma^2_{1Km}(\beta) = Var[t_{Km}(\beta)] = \mathbf{v}'\mathbf{B}_1(\beta)\mathbf{v}$ and $\sigma^2_{2Km}(\beta) = Var[t_{w,Km}(\beta)|\mathcal{F}_K] = \mathbf{v}'\mathbf{B}_2(\beta)\mathbf{v}$, where

$$\mathbf{B}_1(\beta) = \mathbf{V}_T^{-1/2}\mathbf{V}_1\mathbf{V}_T^{-1/2} = \mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{M}_{Km}\mathbf{M}_{Km}^{-1/2}\mathbf{V}_T^{-1/2} = \mathbf{V}_T^{-1},$$

and

$$\mathbf{B}_2(\beta) = \mathbf{V}_T^{-1/2} Var(\mathbf{h}_{w,Km} - \mathbf{h}_{Km}|\mathcal{F}_k)\mathbf{V}_T^{-1/2} = \mathbf{V}_T^{-1/2} Var(\mathbf{h}_{w,Km}|\mathcal{F}_K)\mathbf{V}_T^{-1/2} = \mathbf{V}_T^{-1/2}\mathbf{V}_2\mathbf{V}_T^{-1/2} = \mathbf{V}_T^{-1}.$$

Note that $\mathbf{B}_1 = \{\mathbf{V}_1 + E[\mathbf{V}_2]\}^{-1} = \{\mathbf{I} + E[\mathbf{V}_2]\}^{-1}$ and $\mathbf{B}_2 = \{\mathbf{I} + E[\mathbf{V}_2]\}^{-1/2}\mathbf{V}_2\{\mathbf{I} + E[\mathbf{V}_2]\}^{-1/2}$.

A.2.7 Additional conditions

The following condition is adapted from condition *CC* for the complete data scenario in [81].

Condition CC_{w_0}: Note that $E[\mathcal{D}_{w,Km}(\beta)] = E[\mathcal{D}_{Km}(\beta)]$. For any given $r > 0$ and $\delta > 0$,

$$P\left(sup_{\beta \in B_{w,Km}(r)} ||\mathbf{H}_{Km}^{-1/2}\mathcal{D}_{w,Km}(\beta)\mathbf{H}_{Km}^{-1/2} - \mathbf{I}_{p\times p}|| < \delta\right) \to 1$$

where $B_{w,Km(r)} = \{\beta : ||\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}(\beta - \beta_0)|| \leq r$, and the matrix norm is the Euclidean matrix norm.

Condition CC_w: For any given $r > 0$ and $\delta > 0$, there exist matrices $\mathbf{B}_{1,\infty} = \lim_{K\to\infty} \mathbf{V}_T^{-1}$ and $\mathbf{B}_{2,\infty} = \lim_{K\to\infty} \mathbf{V}_T^{-1/2} E[\mathbf{V}_2]\mathbf{V}_T^{-1/2}$ such that

$$||\mathbf{B}_2(\beta_0) - \mathbf{B}_{2,\infty}|| \xrightarrow{\text{a.s.}} 0$$

$$||\mathbf{B}_1(\beta_0) - \mathbf{B}_{1,\infty}|| \xrightarrow{\text{a.s.}} 0$$

where the matrix norm is the Eucildean matrix norm. Note that $\mathbf{B}_{1,\infty} + \mathbf{B}_{2,\infty} = \mathbf{I}_{p\times p}$, since $\mathbf{V}_T^{-1} + \mathbf{V}_T^{-1/2} E[\mathbf{V}_2]\mathbf{V}_T^{-1/2} = \mathbf{V}_T^{-1/2}(\mathbf{V}_T^{-1/2} + E[\mathbf{V}_2]\mathbf{V}_T^{-1/2}) = \mathbf{V}_T^{-1/2}((\mathbf{I} + E[\mathbf{V}_2])\mathbf{V}_T^{-1/2}) = \mathbf{V}_T^{-1/2}\mathbf{V}_T\mathbf{V}_T^{-1/2} = \mathbf{I}_{p\times p}$.

The following theorem, which mirrors Theorem 3 in [81], shows that the asymptotic distributions of $\widehat{\beta}_w$ and $\mathcal{U}_{w,Km}(\beta)$ are closely related.

Theorem 7 *Suppose that conditions I_w, L_w, and CC_{w_0} hold. Then, there exists a sequence of solutions $\widehat{\beta}_w$ to the WGEE equation in $B_{w,Km}(r)$ such that $\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}(\widehat{\beta}_w - \beta_0)$ and $\mathbf{V}_T^{-1/2}\mathbf{h}_{w,Km}$ are asymptotically distributed, where $\mathbf{h}_{w,Km}(\beta) = \mathbf{M}_{Km}^{-1/2}\mathcal{U}_{w,Km}(\beta)$.*

Proof. Let $\bar{\beta} \in B_{w,Km}(r)$ such that $\bar{\beta}$ is between β_0 and $\widehat{\beta}_w$. Using Taylor's theorem, we have that:

$$\mathcal{U}_{w,Km}(\beta_0) = \mathcal{D}_{Km}(\bar{\beta})(\widehat{\beta} - \beta_0)$$

since

$$0 = \mathcal{U}_{w,Km}(\widehat{\beta}) = \mathcal{U}_{w,Km}(\beta_0) + \frac{\partial \mathcal{U}_{w,Km}}{\partial \beta^T}\bigg|_{\beta=\bar{\beta}}(\widehat{\beta} - \beta_0) = \mathcal{U}_{w,Km}(\beta_0) - \mathcal{D}_{w,Km}(\bar{\beta})(\widehat{\beta} - \beta_0)$$

It then follows that

$$\begin{aligned}
&\mathbf{H}_{Km}^{-1/2}\mathcal{U}_{w,Km} \\
&\qquad = \mathbf{H}_{Km}^{-1/2}\mathcal{D}_{w,Km}(\bar{\beta})(\widehat{\beta}_w - \beta_0) \\
&\qquad = \mathbf{H}_{Km}^{-1/2}\mathcal{D}_{w,Km}(\bar{\beta})\mathbf{H}_{Km}^{-1/2}\mathbf{H}_{Km}^{1/2}(\widehat{\beta}_w - \beta_0)
\end{aligned}$$

By condition CC_{w_0} and Slutsky's theorem, $\mathbf{H}_{Km}^{-1/2}\mathcal{U}_{w,Km}$ and $\mathbf{H}_{Km}^{1/2}(\widehat{\beta}_w - \beta_0)$ are asymptotically identically distributed. Therefore, $\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathcal{U}_{w,Km}$ and $\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}(\beta - \beta_0)$ are also asymptotically identically distributed.

We now prove the asymptotic normality of $\widehat{\beta}_w$ by first showing the asymptotic normality of $\mathcal{U}_{w,Km}$.

Theorem 8a For each $k = 1, \ldots, K; i = 1, \ldots, N_k$, let $U_i^{k*}(\beta) = L_i^{k*}(\beta)(Y_i - \mu_i(\beta))$, where $L_i^{k*}(\beta)$ denotes the i^k-th entry of the vector $\mathbf{L}^*(\beta) = \mathbf{v}^T\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}D^T\mathbf{V}^{-1}$, where $\mathbf{v}$ is a $p \times 1$ vector with $||\mathbf{v}|| = 1$. Suppose that all assumptions in Lemma 4 hold for the variable U. Then, $\sigma_{2Km}^{-1}t_{w,Km}|\mathcal{F}_K \xrightarrow{\text{d}} N(0,1), \sigma_{1Km}^{-1}t_{Km} \xrightarrow{\text{d}} N(0,1), \mathbf{V}_T^{-1/2}\mathbf{h}_{w,Km} \xrightarrow{\text{d}} N(0, \mathbf{B}_{1,\infty} + \mathbf{B}_{2,\infty})) = N(0, \mathbf{I}_{p\times p})$, and

$$\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}(\widehat{\beta}-\beta_0)\xrightarrow{\mathrm{d}}N(0,\mathbf{I}_{p\times p})$$

If, additionally, we assume that $lim_{K\to\infty}K\mathbf{M}_{Km}^{-1/2}=\mathbf{M}_\infty$, $lim_{K\to\infty}\mathbf{H}_{Km}=\mathbf{H}_\infty$, so that $K\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}\to\mathbf{B}_{1,\infty}^{1/2}\mathbf{M}_\infty\mathbf{H}_\infty$, then

$$\sqrt{K}(\widehat{\beta}_w-\beta_0)\xrightarrow{\mathrm{d}}N\Big(0,[\mathbf{B}_{1,\infty}^{1/2}\mathbf{M}_\infty\mathbf{H}_\infty]^{-2}\Big)$$

Proof. Note that the vector $\mathbf{L}^*(\beta)$, which can be reexpressed as $\mathbf{L}^*(\beta)=\mathbf{v}^*T\mathbf{D}^T\mathbf{V}^{-1}$, where $\mathbf{v}^*=\mathbf{M}_{Km}^{-1/2}\mathbf{V}_T^{-1/2}\mathbf{v}$, is of the form

$$\mathbf{L}^*(\beta)=\Big[\textstyle\sum_{l=1}^p v_l^*D_l^1(V_{11}^1)^{-1},\ldots,\sum_{l=1}^p v_l^*D_{N^K,l}^K(V_{N^KN^K}^K)^{-1}\Big],$$

where $(V_{ii}^{kk})^{-1}$ denotes the i^ki^k-th element of the matrix $\mathbf{V}^{-1}(\beta)$. Then $t_{w,Km}$ can be written as $t_{w,Km}=\sum_{k=1}^K\sum_{j=1}^J\sum_{i=1}^{N_j^k}R^kw^kU_{j_i}^{k*}(\beta_0)-\sum_{k=1}^K\sum_{j=1}^J\sum_{i=1}^{N_j^k}U_{j_i}^{k*}(\beta_0)$, where $U_{j_i}^{k*}(\beta)=L_{j_i}^{k*}(\beta)(Y_{j_i}^k-\mu_{j_i}^k(\beta))$. It follows that $Var(t_{w,Km}|\mathcal{F}_K)=\sum_{k\in A}\sum_{i=1}^{N^k}\frac{1-\pi_i^k}{\pi_i^k}U_i^{k*2}$. By assumption (5) and Lemma 4, we conclude that

$$t_{w,Km}/\sigma_{2Km}=\frac{1}{\sigma_{2Km}}\Bigg(\sum_{k=1}^K\sum_{k=1}^{N_k}R_kw_kU_i^{k*}(\beta)-\sum_{k=1}^K\sum_{i=1}^{N_k}U_i^{k*}(\beta_0)\Bigg)\Bigg|\mathcal{F}_K\xrightarrow{\mathrm{d}}N(0,1).$$

This together with condition CC_w implies that

$$t_{W,Km}=\Bigg(\sum_{k=1}^K\sum_{i=1}^{N^k}R_kw_kU_i^{k*}(\beta)-\sum_{k=1}^K\sum_{i=1}^{N^k}U_i^{k*}(\beta_0)\Bigg)\Bigg|\mathcal{F}_K\xrightarrow{\mathrm{d}}N(0,\sigma_{2,\infty}^2).$$

So $t_{w,Km}=\mathbf{v}^T\mathbf{V}_T^{-1/2}(\mathbf{h}_{w,Km}-\mathbf{h}_{Km})|\mathcal{F}_K\xrightarrow{\mathrm{d}}N(0,\sigma_{2,\infty}^2)$, where $\sigma_{2,\infty}^2=\mathbf{v}^T\mathbf{B}_{2,\infty}\mathbf{v}$. By Theorem 4 in Xie and Yang (2003), $\mathbf{h}_{Km}\xrightarrow{\mathrm{d}}N(0,\mathbf{I}_{p\times p})$. Then, by condition CC_w and because $\mathbf{V}_T^{-1/2}=\mathbf{B}_1^{1/2}$, we have $\mathbf{v}^T\mathbf{V}_T^{-1/2}\mathbf{h}_{Km}\xrightarrow{\mathrm{d}}N(0,\sigma_{1,\infty})$, where $\sigma_{1,\infty}^2=\mathbf{v}^T\mathbf{B}_{1,\infty}\mathbf{v}$. In summary, we have that

$$\mathbf{v}^T\mathbf{V}^{-1/2}(\mathbf{h}_{Km})\xrightarrow{\mathrm{d}}N(0,\sigma_{1,\infty}^2)$$

$$\mathbf{v}^T\mathbf{V}^{-1/2}(\mathbf{h}_{w,Km} - \mathbf{h}_{Km})|\mathcal{F}_K \xrightarrow{\text{d}} N(0, \sigma^2_{2,\infty})$$

Then by Lemma 5, we obtain $v^T V^{-1/2} h_{w,Km} \xrightarrow{\text{d}} N(0, \sigma^2_{1,\infty} + \sigma^2_{2,\infty})$, for any v. Therefore,

$$\mathbf{V}_T^{-1/2}\mathbf{h}_{w,Km} \xrightarrow{\text{d}} N(0, \mathbf{B}_{1,\infty} + \mathbf{B}_{2,\infty}) = N(0, \mathbf{I}_{p\times p})$$

This result and Theorem 7 imply that

$$\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}(\widehat{\beta} - \beta_0) \xrightarrow{\text{d}} N(0, \mathbf{I}_{p\times p}).$$

and that

$$\sqrt{K}(\widehat{\beta}_w - \beta_0) \xrightarrow{\text{d}} N(0, [\mathbf{B}_{1,\infty}^{1/2}\mathbf{M}_\infty\mathbf{H}_\infty]^{-2})$$

Theorem 8b. For each $k = 1, \ldots, K; j = 1, \ldots, J; i = 1, \ldots, N_j^k$, let $U_{j_i}^{k*}(\beta) = L_{j_i}^{k*}(\beta)(Y_{j_i} - \mu_{j_i}(\beta))$, where $L_{j_i}^{k*}(\beta)$ denotes the j_i^k-th entry of the vector $\mathbf{L}^*(\beta) = \mathbf{v}^T\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{D}^T\mathbf{V}^{-1}$, where $\mathbf{v}$ is a $p \times 1$ vector with $||\mathbf{v}|| = 1$. Let $A_j = \{\text{k} : \text{cluster } k \text{ is in stratum } j\}$. Assume that

$$\frac{1}{N_j}\sum_{k\in A_j}\sum_{i=1}^{N^k}\left[U_{j_i}^{k*}(\beta_0), (U_{j_i}^{k*}(\beta_0) - \bar{U}_j^*(\beta_0))^2, (U_{j_i}^{k*}(\beta_0))^4\right] \xrightarrow{\text{a.s.}} [M_{1j}, M_{2j}, M_{4j}], \tag{A.8}$$

where $\bar{U}_j^*(\beta_0) = \frac{1}{N_j}\sum_{k\in A_j}\sum_{i=1}^{N_k} U_{j_i}^{k*}(\beta_0)$ and M_{1j}, M_{2j}, M_{4j} are constants. Then, under the assumptions above, $\sigma_{2Km}^{-1}t_{w,Km}|\mathcal{F}_K \xrightarrow{\text{d}} N(0,1)$, $\sigma_{1Km}^{-1}t_{Km} \xrightarrow{\text{d}} N(0,1)$, $\mathbf{V}_T^{-1/2}\mathbf{h}_{w,Km} \xrightarrow{\text{d}} N(0, \mathbf{B}_{1,\infty} + \mathbf{B}_{2,\infty})) = N(0, \mathbf{I}_{p\times p})$, and

$$\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}(\widehat{\beta} - \beta_0) \xrightarrow{\text{d}} N(0, \mathbf{I}_{p\times p})$$

If we additionally assume that $lim_{K\to\infty}K\mathbf{M}_{Km}^{-1/2} = \mathbf{M}_\infty$, $lim_{K\to\infty}\mathbf{H}_{Km} = \mathbf{H}_\infty$, so that $K\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km} \to \mathbf{B}_{1,\infty}^{1/2}\mathbf{M}_\infty\mathbf{H}_\infty$, then

$$\sqrt{K}(\widehat{\beta}_w - \beta_0) \xrightarrow{\mathrm{d}} N\Big(0, [\mathbf{B}_{1,\infty}^{1/2}\mathbf{M}_\infty\mathbf{H}_\infty]^{-2}\Big)$$

Proof. Note that the vector $\mathbf{L}^*(\beta)$, which can be reexpressed as $\mathbf{L}^*(\beta) = \mathbf{v}^{*T}\mathbf{D}^T\mathbf{V}^{-1}$, where $\mathbf{v}^* = \mathbf{M}_{Km}^{-1/2}\mathbf{V}_T^{-1/2}\mathbf{v}$, is of the form

$$\mathbf{L}^*(\beta) = \left[\sum_{l=1}^p v_l^* D_{1_1 l}^1 (V_{1_1 1_1})^{-1}, \ldots, \sum_{l=1}^p v_l^* D_{J_{N_j^K}, l}^K (V_{J_{N_j} N_j}^{KK})^{-1}\right],$$

where $(V_{ii}^{kk})^{-1}$ denotes the $i^k i^k - th$ element of the matrix $\mathbf{V}^{-1}(\beta)$. Then $t_{w,Km}$ can be written as $t_{w,Km} = \sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^{N_j^k} R^k w^k U_{j_i}^{k*}(\beta_0) - \sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^{N_j^k} U_{j_i}^{k*}(\beta_0)$, where $U_{j_i}^{k*}(\beta) = L_{j_i}^k(\beta)(Y_{j_i}^k - \mu_{j_i}^k(\beta))$. It follows that $Var(t_{w,Km}|\mathcal{F}_K) = \sum_{j=1}^J \frac{N_j}{n_j}(N_j - n_j)S_j^{*2}$, where $S_j^{*2} = \sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^{N_j^k}(U_{j_i}^{k*}(\beta_0) - \bar{U}_{j_i}^{k*}(\beta_0))^2/(N_j - 1)$ and $\bar{U}_j^*(\beta_0) = \sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^{N_j^k} U_{j_i}^{k*}(\beta)/N_j$. By assumption (5) and Lemma 4, we conclude that

$$t_{w,Km}/\sigma_{2Km} = \frac{1}{\sigma_{2Km}}\left(\sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^{N_j^k} R^k w^k U_{j_i}^{k*}(\beta) - \sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^{N_j^k} U_{j_i}^{k*}(\beta_0)\right)\Big|\mathcal{F}_K \xrightarrow{\mathrm{d}} N(0,1).$$

This together with condition CC_w implies that

$$t_{w,Km} = \left(\sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^{N_j^k} R_k w_k U_{j_i}^{k*}(\beta) - \sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^{N_j^k} U_{j_i}^{k*}(\beta_0)\right)\Big|\mathcal{F}_K \xrightarrow{\mathrm{d}} N(0, \sigma_{2,\infty}^2).$$

So $t_{w,Km} = \mathbf{v}^T\mathbf{V}_T^{-1/2}(\mathbf{h}_{w,Km} - \mathbf{h}_{Km})|\mathcal{F}_K \xrightarrow{\mathrm{d}} N(0, \sigma_{2,\infty}^2)$, where $\sigma_{2,\infty}^2 = \mathbf{v}^T\mathbf{B}_{2,\infty}\mathbf{v}$. By Theorem 4 in [81], $\mathbf{h}_{Km} \xrightarrow{\mathrm{d}} N(0, \mathbf{I}_{p\times p})$. Then, by condition CC_w and because $\mathbf{V}_T^{-1/2} = \mathbf{B}_1^{1/2}$, we have $\mathbf{v}^T\mathbf{V}_T^{-1/2}\mathbf{h}_{Km} \xrightarrow{\mathrm{d}} N(0, \sigma_{1,\infty})$, where $\sigma_{1,\infty}^2 = \mathbf{v}^T\mathbf{B}_{1,\infty}\mathbf{v}$. In summary, we have that

$$\mathbf{v}^T\mathbf{V}^{-1/2}(\mathbf{h}_{Km}) \xrightarrow{\mathrm{d}} N(0, \sigma_{1,\infty}^2)$$

$$\mathbf{v}^T\mathbf{V}^{-1/2}(\mathbf{h}_{w,Km} - \mathbf{h}_{Km})|\mathcal{F}_K \xrightarrow{\mathrm{d}} N(0, \sigma_{2,\infty}^2)$$

Then by Lemma 5, we obtain $\mathbf{v}^T\mathbf{V}^{-1/2}\mathbf{h}_{w,Km} \xrightarrow{\text{d}} N(0, \sigma^2_{1,\infty} + \sigma^2_{2,\infty})$, for any $\mathbf{v}$. Therefore,

$$\mathbf{V}_T^{-1/2}\mathbf{h}_{w,Km} \xrightarrow{\text{d}} N(0, \mathbf{B}_{1,\infty} + \mathbf{B}_{2,\infty}) = N(0, \mathbf{I}_{p\times p})$$

This result and Theorem 7 imply that

$$\mathbf{V}_T^{-1/2}\mathbf{M}_{Km}^{-1/2}\mathbf{H}_{Km}(\widehat{\beta} - \beta_0) \xrightarrow{\text{d}} N(0, \mathbf{I}_{p\times p}).$$

and that

$$\sqrt{K}(\widehat{\beta}_w - \beta_0) \xrightarrow{\text{d}} N(0, [\mathbf{B}_{1,\infty}^{1/2}\mathbf{M}_\infty\mathbf{H}_\infty]^{-2})$$

A.3 Bias Correction for Point Estimates under Cluster-Based Outcome-Dependent Sampling Design

Building upon the work of[36], we derive the bias formula that incorporates the weights needed to account for the non-random sampling design. Under a cluster-based outcome-dependent sampling design, the bias formula $B_w(\beta_0)$, taking into account the weights, can be expressed as

$$\hat{\iota}_w^{l_1 l_2}\hat{\iota}_w^{l_3 l_4}\hat{\kappa}_w^{l_2 l_3, l_4} - \frac{1}{2}\hat{\iota}_w^{l_1 l_2}\hat{\iota}_w^{l_3 l_4}\hat{\iota}_w^{l_5 l_6}\hat{\kappa}_w^{l_2 l_3 l_5}\hat{\kappa}_w^{l_4, l_6}, \tag{A.9}$$

where letting $\bar{V}_k = V_k^{-1}$,

$$\mathcal{U}_w^{l_1} = \sum_{k=1}^{K} \frac{R_k}{\pi_k} D_k^{t_1 l_1} \bar{V}_k^{t_1 t_2} (Y_k^{t_2} - \mu_k^{t_2}),$$

$$\mathcal{U}_w^{l_1 l_2} = \sum_{k=1}^{K} \{(\partial/\partial\beta^{l_2}) \frac{R_k}{\pi_k} D_k^{t_1 l_1} \bar{V}_k^{t_1 t_2}\}(Y_k^{t_2} - \mu_k^{t_2}) - \frac{R_k}{\pi_k} D_k^{t_1 l_1} \bar{V}_k^{t_1 t_2} D^{t_2 l_2}$$

and

$$\mathcal{U}_W^{l_1 l_2 l_3} = \sum_{k=1}^{K} \{(\partial^2/\partial\beta^{l_2}\beta^{l_3}) \frac{R_k}{\pi_k} D_k^{t_1 l_1} \bar{V}_k^{t_1 t_2}\}(Y_k^{t_2} - \mu_k^{t_2}) - \{(\partial/\partial\beta^{l_2}) \frac{R_k}{\pi_k} D_k^{t_1 l_1} \bar{V}_k^{t_1 t_2}\} D_k^{t_2 l_3}$$

$$-\{(\partial/\partial\beta^{l_3}) \frac{R_k}{\pi_k} D^{t_1 l_1} \bar{V}_k^{t_1 t_2}\} D_k^{t_2 l_2} - \frac{R_k}{\pi_k} D_k^{t_1 l_1} \bar{V}_k^{t_1 t_2} \{(\partial/\partial\beta^{l_3}) D_k^{t_2 l_2}\}.$$

The moments needed to obtain (10) are

$$\kappa_w^{l_1 l_2} = -\sum_{k=1}^{K} D_k^{t_1 l_1} \bar{V}_k^{t_1 t_2} D_k^{t_2 l_2},$$

which can be estimated by

$$\widehat{\kappa}_w^{l_1 l_2} = -\sum_{k=1}^{K} \frac{R_k}{\pi_k} D_k^{t_1 l_1} \bar{V}_k^{t_1 t_2} D_k^{t_2 l_2},$$

and $\iota_w^{l_1 l_2}$ is then defined to be the inverse of $\kappa_w^{l_1 l_2}$. $\kappa_w^{l_1, l_2}$ is the $l_1 l_2{}^{th}$ component of

$$Var(\mathcal{U}_w) = Var[\mathbf{U}^T 1_{N\times 1}] + E[\mathbf{U}^T \mathbf{W} Var(\mathbf{R}|\mathcal{F}_k)\mathbf{W}\mathbf{U}],$$

which, as described in section 5.2 of the main paper, can be estimated by

$$V.I + V.II = \mathbf{U}^T diag(\mathbf{R})\mathbf{W}_2^k diag(\mathbf{R})\mathbf{U} + \mathbf{U}^T \mathbf{W} diag(\mathbf{R})\widetilde{\Delta} diag(\mathbf{R})\mathbf{W}\mathbf{U}.$$

$$\kappa_w^{l_1 l_2 l_3} = -\sum_{k=1}^{K} \{(\partial/\partial \beta^{l_2}) D_k^{t_1 l_1} \bar{V}_k^{t_1 t_2}\} D_k^{t_2 l_3} + \{(\partial/\partial \beta^{l_3}) D_k^{t_1 l_1} \bar{V}_k^{t_1 t_2}\} D_k^{t_2 l_2} + D_k^{t_1 l_2} \bar{V}_k^{t_1 t_2} \{\partial/\partial \beta^{l_3}) D_k^{t_2 l_2}\},$$

which can be estimated by

$$\widehat{\kappa}_w^{l_1 l_2 l_3} = -\sum_{k=1}^{K} \frac{R_k}{\pi_k} \{(\partial/\partial \beta^{l_2}) D_k^{t_1 l_1} \bar{V}_k^{t_1 t_2}\} D_k^{t_2 l_3} + \frac{R_k}{\pi_k} \{(\partial/\partial \beta^{l_3}) D_k^{t_1 l_1} \bar{V}_k^{t_1 t_2}\} D_k^{t_2 l_2} + \frac{R_k}{\pi_k} D_k^{t_1 l_2} \bar{V}_k^{t_1 t_2} \{\partial/\partial \beta^{l_3}) D_k^{t_2 l_2}\}.$$

Finally, $\kappa_w^{l_1 l_2, l_3}$ is the $l_1 l_3{}^{rd}$ component of

$$
\begin{aligned}
&E[\{\partial/\partial\beta^{l_2}\mathcal{U}_w\}\mathcal{U}_w^T|X] \\
&\qquad = E[\{\partial/\partial\beta^{l_2}\mathbf{U}^T\mathbf{W}\mathbf{R}\}\mathbf{R}^T\mathbf{W}\mathbf{U}|X] \\
&\qquad = E[\{\partial/\partial\beta^{l_2}\mathbf{U}^T\}\mathbf{W}\mathbf{R}\mathbf{R}^T\mathbf{W}\mathbf{U}|X] \\
&\qquad = E[E[\{\partial/\partial\beta^{l_2}\mathbf{U}^T\}\mathbf{W}\mathbf{R}\mathbf{R}^T\mathbf{W}\mathbf{U}|X, Y]|X] \\
&\qquad = E[\{\partial/\partial\beta^{l_2}\mathbf{U}^T\}\mathbf{W}E[\mathbf{R}\mathbf{R}^T|X, Y]\mathbf{W}\mathbf{U}|X]
\end{aligned}
$$

E[$\mathbf{R}\mathbf{R}^T$] is an $N \times N$ matrix with entries equal to the joint probability of selection for pairs of individuals. For two individuals belonging to cluster k, the joint probability of selection is π_k; for two individuals from clusters k and k', the joint probability of selection is $\pi_{kk'}$, which under a cluster-stratified design is equal to $\frac{k_j}{K_j}\frac{k_j-1}{K_j-1}$ if clusters k and k' belong to the same stratum j, and is equal to $\frac{k_j}{K_j}\frac{k_{j'}}{K_{j'}}$ if clusters k and k' belong to different strata j and j'.

The expression above can be estimated by $\{\partial/\partial\beta^{l_2}\mathbf{U}^T\}\mathbf{W}\mathbf{R}\mathbf{R}^T\mathbf{W}\mathbf{U}$.

A.4 Mancl and DeRouen-type Variance Bias Correction

For data collected through a cluster-based outcome-dependent sampling design, parameter estimates $\widehat{\beta}_w$, are obtained by solving

$$\mathcal{U}_w = \sum_{k=1}^{K} \frac{R_k}{\pi_k} \mathbf{D}_k^T \mathbf{V}_k^{-1} (\mathbf{Y}_k - \boldsymbol{\mu}_k) = 0$$

Furthermore,

$$\widehat{Var}[\widehat{\beta}_w] \;=\; \widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1} \left\{ \widehat{\mathbf{V}}_I(\widehat{\beta}_w) + \widehat{\mathbf{V}}_{II}(\widehat{\beta}_w) \right\} \widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}, \tag{A.10}$$

where $\widehat{\mathbf{H}}_w(\beta) \;=\; -\sum_{k=1}^{K} \widehat{\mathbf{H}}_{w,k}(\beta) = -\sum_{k=1}^{K} \frac{R_k}{\pi_k} \mathbf{D}_k^T \mathbf{V}_k^{-1} \mathbf{D}_k$, and

$$\widehat{\mathbf{V}}_I(\widehat{\beta}_w) = \sum_{k=1}^{K} \frac{R_k}{\pi_k} \mathbf{D}_k^T \mathbf{V}_k^{-1} \widehat{\varepsilon}_{w,k} \widehat{\varepsilon}_{w,k}^T \mathbf{V}_k^{-1} \mathbf{D}_k$$

and

$$\begin{aligned}
&\widehat{\mathbf{V}}_{II}(\widehat{\beta}_w) \\
&\quad = \sum_{k=1}^{K} B_{kk} \mathbf{D}_k^T \mathbf{V}_k^{-1} \widehat{\varepsilon}_{w,k} \widehat{\varepsilon}_{w,k}^T \mathbf{V}_k^{-1} \mathbf{D}_k + \sum_{k=1}^{K} \sum_{k' \neq k} \mathbf{D}_k^T \mathbf{V}_k^{-1} diag(\widehat{\varepsilon}_{w,k}) \mathbf{B}_{kk'} diag(\widehat{\varepsilon}_{w,k'}) \mathbf{V}_{k'}^{-1} \mathbf{D}_{k'} \\
&\quad = \widehat{\mathbf{V}}_{II}^{*}(\widehat{\beta}_w) + \widehat{\mathbf{V}}_{II}^{e}(\widehat{\beta}_w)
\end{aligned}$$

with $\widehat{\varepsilon}_{w,k} = (\mathbf{Y}_k - \widehat{\boldsymbol{\mu}}_{w,k})$, where $\widehat{\boldsymbol{\mu}}_{w,k}$ is an N_k-vector with elements $\widehat{\mu}_{w,ki} \;=\; g^{-1}(X_{ki}^T \widehat{\beta}_w)$, $B_{kk} \;=\; \pi_k^{-3}(\pi_k - \pi_k^2) R_k$, and $\mathbf{B}_{kk'}$ an $N_k \times N_{k'}$ matrix with all entries equal to $\frac{(\pi_{kk'} - \pi_k \pi_{k'}) R_k R_{k'}}{\pi_{kk'} \pi_k \pi_{k'}}$. The residual estimator, $\widehat{\varepsilon}_{w,k} \widehat{\varepsilon}_{w,k}^T$, may be downward biased in small samples, and we adapt the approach of[38] to correct $\widehat{\mathbf{V}}_I(\widehat{\beta}_w)$ and $\widehat{\mathbf{V}}_{II}^{*}(\widehat{\beta}_w)$ for this bias. First note that

$$
\begin{aligned}
&E[\widehat{\mathbf{V}}_I(\widehat{\boldsymbol{\beta}}_w)] \\
&\qquad = E[E[\widehat{\mathbf{V}}_I(\widehat{\boldsymbol{\beta}}_w)|\mathbf{R}]] \\
&\qquad = \sum_{k=1}^{K} E[\frac{R_k}{\pi_k} E[\mathbf{D}_k^T \mathbf{V}_k^{-1} \widehat{\boldsymbol{\varepsilon}}_{w,k} \widehat{\boldsymbol{\varepsilon}}_{w,k}^T \mathbf{V}_k^{-1} \mathbf{D}_k|\mathbf{R}]] \\
&\qquad = \sum_{k=1}^{K} E[\frac{R_k}{\pi_k} \mathbf{D}_k^T \mathbf{V}_k^{-1} E[\widehat{\boldsymbol{\varepsilon}}_{w,k} \widehat{\boldsymbol{\varepsilon}}_{w,k}^T|\mathbf{R}] \mathbf{V}_k^{-1} \mathbf{D}_k]
\end{aligned}
$$

and

$$
\begin{aligned}
&E[\widehat{\mathbf{V}}_{II}^*(\widehat{\boldsymbol{\beta}}_w)] \\
&\qquad = E[E[\widehat{\mathbf{V}}_{II}^*(\widehat{\boldsymbol{\beta}}_w)|\mathbf{R}]] \\
&\qquad = \sum_{k=1}^{K} E[B_{kk} E[\mathbf{D}_k^T \mathbf{V}_k^{-1} \widehat{\boldsymbol{\varepsilon}}_{w,k} \widehat{\boldsymbol{\varepsilon}}_{w,k}^T \mathbf{V}_k^{-1} \mathbf{D}_k|\mathbf{R}]] \\
&\qquad = \sum_{k=1}^{K} E[B_{kk} \mathbf{D}_k^T \mathbf{V}_k^{-1} E[\widehat{\boldsymbol{\varepsilon}}_{w,k} \widehat{\boldsymbol{\varepsilon}}_{w,k}^T|\mathbf{R}] \mathbf{V}_k^{-1} \mathbf{D}_k]
\end{aligned}
$$

Now, carrying out a Taylor series expansion of $\widehat{U}_k = U_k(\widehat{\beta}_w) = \mathbf{D}_k^T(\widehat{\beta}_w)\mathbf{V}_k^{-1}(\widehat{\beta}_w)(\mathbf{Y}_k(\widehat{\beta}_w) - \boldsymbol{\mu}_k(\widehat{\beta}_w))$ about β yields:

$$
\widehat{U}_k \approx U_k(\beta) + \left.\frac{\partial U_k}{\partial \beta}\right|_{\beta=\beta} (\widehat{\beta}_w - \beta) = U_k(\beta) - H_k(\beta)(\widehat{\beta}_w - \beta) \tag{A.11}
$$

where $\mathbf{H}_k(\beta) = -[\frac{\partial U_k}{\partial \boldsymbol{\beta}}] = \mathbf{D}_k^T \mathbf{V}_k^{-1} \mathbf{D}_k$. Taking the sum over the clusters, $k = 1, \ldots, K$, yields:

$$
\sum_{k=1}^{K} \widehat{U}_k \approx \sum_{k=1}^{K} U_k - \sum_{k=1}^{K} H_k(\beta)(\widehat{\beta}_w - \beta)
$$

$$
\sum_{k=1}^{K} U_k - \sum_{k=1}^{K} \widehat{U}_k \approx \sum_{k=1}^{K} H_k(\beta)(\widehat{\beta}_w - \beta)
$$

From which it follows that

$$(\widehat{\beta}_w - \beta) \approx H(\beta)^{-1} \sum_{k=1}^{K} U_k \tag{A.12}$$

where $\mathbf{H}(\beta) = \sum_{k=1}^{K} H_k(\beta)$. Now, in the same vein as Mancl and DeRouen [38], carrying out a first-order Taylor-series expansion of the fitted residuals, $\widehat{\varepsilon}_{w,k}$ about β yields

$$\widehat{\varepsilon}_{w,k} = \varepsilon_k + \frac{\partial \varepsilon_k}{\partial \beta}(\widehat{\beta}_w - \beta).$$

Squaring this and taking the expectation conditional on $\mathbf{R}$, it follows that

$$E[\widehat{\varepsilon}_{w,k}\widehat{\varepsilon}_{w,k}^T|\mathbf{R}] = E[\varepsilon_k\varepsilon_k^T] + E[\varepsilon_k(\widehat{\beta}_w - \beta)^T\frac{\partial \varepsilon_k^T}{\partial \beta^T}] + E[\frac{\partial \varepsilon_k}{\partial \beta}(\widehat{\beta}_w - \beta)\varepsilon_k^T] + E[\frac{\partial \varepsilon_k}{\partial \beta}(\widehat{\beta}_w - \beta)(\widehat{\beta}_w - \beta)^T\frac{\partial \varepsilon_k^T}{\partial \beta^T}]$$

where the notation indicating that the expectation is conditional on $\mathbf{R}$ is suppressed on the right-hand side of th above equation. Then, using the approximation $(\widehat{\beta}_w - \beta) \approx \mathbf{H}(\beta)^{-1}\sum_{k=1}^{K} U_k$, the conditional expectation $E[\widehat{\varepsilon}_{w,k}\widehat{\varepsilon}_{w,k}^T|\mathbf{R}]$ can be approximated by,

$$\text{Cov}[\mathbf{Y}_k] - \text{Cov}[\mathbf{Y}_k]\mathbf{A}_{kk,w}^T - \mathbf{A}_{kk,w}\text{Cov}[\mathbf{Y}_k] + \mathbf{A}_{kk,w}\text{Cov}[\mathbf{Y}_k]\mathbf{A}_{kk,w}^T + \sum_{k' \neq k} \mathbf{A}_{kk',w}\text{Cov}[\mathbf{Y}_{k'}]\mathbf{A}_{kk',w}^T. \tag{A.13}$$

where $\mathbf{A}_{kk',w} =$

$$\mathbf{D}_k(\textstyle\sum_{m=1}^{K} \mathbf{D}_m^T\mathbf{V}_m^{-1}\mathbf{D}_m)^{-1}\mathbf{D}_{k'}^T\mathbf{V}_{k'}^{-1} = \mathbf{D}_k\mathbf{H}(\beta)^{-1}\mathbf{D}_{k'}^T\mathbf{V}_{k'}^{-1}.$$

Assuming that the last term in (A.13) is negligible, rearranging yields the following approximation:

$$E[\widehat{\varepsilon}_{w,k}\widehat{\varepsilon}_{w,k}^T|\mathbf{R}] \approx (\mathbf{I}_{N_k} - \mathbf{A}_{kk,w})\text{Cov}[\mathbf{Y}_k](\mathbf{I}_{N_k} - \mathbf{A}_{kk,w})^T. \tag{A.14}$$

where $\mathbf{I}_{N_k}$ is the $N_k \times N_k$ identity matrix. The approximation in (A.14) and leads to the following bias-corrected variance estimator:

$$\widehat{\widehat{Var}}[\widehat{\beta}_w] \;=\; \widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}\left\{\widehat{\widehat{\mathbf{V}}}_I(\widehat{\beta}_w) + \widehat{\widehat{\mathbf{V}}}_{II}(\widehat{\beta}_w)\right\}\widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}, \tag{A.15}$$

where $\widehat{\mathbf{H}}_w(\beta) \;=\; -\sum_{k=1}^K \mathbf{D}_k^T\mathbf{V}_k^{-1}\mathbf{W}_k diag(\mathbf{R}_k)\mathbf{D}_k$ and

$$\widehat{\widehat{\mathbf{V}}}_I(\widehat{\beta}_w) \;=\; \sum_{k=1}^K \frac{R_k}{\pi_k}\mathbf{D}_k^T\mathbf{V}_k^{-1}(\mathbf{I}_{N_k} - \widehat{\mathbf{A}}_{kk,w})^{-1}\widehat{\varepsilon}_{w,k}\widehat{\varepsilon}_{w,k}^T(\mathbf{I}_{N_k} - \widehat{\mathbf{A}}_{kk,w}^T)^{-1}\mathbf{V}_k^{-1}\mathbf{D}_k$$

and

$$\begin{aligned}\widehat{\widehat{\mathbf{V}}}_{II}(\widehat{\beta}_w) \quad &= \; \sum_{k=1}^K B_{kk}\mathbf{D}_k^T\mathbf{V}_k^{-1}(\mathbf{I}_{N_k} - \widehat{\mathbf{A}}_{kk,w})^{-1}\widehat{\varepsilon}_{w,k}\widehat{\varepsilon}_{w,k}^T(\mathbf{I}_{N_k} - \widehat{\mathbf{A}}_{kk,w}^T)^{-1}\mathbf{V}_k^{-1}\mathbf{D}_k \\ &\quad + \; \sum_{k=1}^K \sum_{k'\neq k} \mathbf{D}_k^T\mathbf{V}_k^{-1} diag(\widehat{\varepsilon}_{w,k})\mathbf{B}_{kk'} diag(\widehat{\varepsilon}_{w,k'})\mathbf{V}_{k'}^{-1}\mathbf{D}_{k'}\end{aligned}$$

where $B_{kk} = \pi_k^{-3}(\pi_k - \pi_k^2)R_k$, $\mathbf{B}_{kk'}$ is an $N_k \times N_{k'}$ matrix with all entries equal to $\frac{(\pi_{kk'} - \pi_k\pi_{k'})R_kR_{k'}}{\pi_{kk'}\pi_k\pi_{k'}}$ and $\widehat{\mathbf{A}}_{kk',w}(\beta) =$

$$\mathbf{D}_k(\textstyle\sum_{m=1}^K \mathbf{D}_m^T\mathbf{V}_m^{-1}\mathbf{W}_m diag(\mathbf{R}_m)\mathbf{D}_m)^{-1}\mathbf{D}_{k'}^T\mathbf{V}_{k'}^{-1}\mathbf{W}_{k'} diag(\mathbf{R}_{k'}) =$$
$$\mathbf{D}_k\widehat{\mathbf{H}}_w^{-1}(\beta)\mathbf{D}_{k'}^T\mathbf{V}_{k'}^{-1}\mathbf{W}_{k'} diag(\mathbf{R}_{k'}).$$

If the data arise from a cluster-stratified design, the correction for the naïve estimator of the variance that ignores the negative correlation in the sampling indicators is:

$$\widehat{\widehat{Var}}^*[\widehat{\beta}_w] \;=\; \widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}\left\{\widehat{\widehat{\mathbf{V}}}_I(\widehat{\beta}_w) + \widehat{\widehat{\mathbf{V}}}^*_{II}(\widehat{\beta}_w)\right\}\widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1},$$

where

$$\widehat{\widehat{\mathbf{V}}}^*_{II}(\widehat{\beta}_w) \;=\; \sum_{k=1}^K B_{kk}\mathbf{D}_k^T\mathbf{V}_k^{-1}(\mathbf{I}_k - \widehat{\mathbf{A}}_{kk,w})^{-1}\widehat{\varepsilon}_{w,k}\widehat{\varepsilon}_{w,k}^T(\mathbf{I}_k - \widehat{\mathbf{A}}_{kk,w}^T)^{-1}\mathbf{V}_k^{-1}\mathbf{D}_k.$$

Note that this is also the corrected variance estimator for data arising from a Poisson sampling design.

A.5 Kauermann and Carroll-type Variance Bias Correction

The Kauermann and Carroll[30]-type bias correction to the variance estimator is similar to the bias correction derived in the previous section. We therefore simply state the form of the correction below:

$$\widehat{Var}_{KC}[\widehat{\beta}_w] = \widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}\left\{\widehat{\mathbf{V}}_{I,KC}(\widehat{\beta}_w) + \widehat{\mathbf{V}}_{II,KC}(\widehat{\beta}_w)\right\}\widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}, \tag{A.16}$$

where $\widehat{\mathbf{H}}_w(\beta) = -\sum_{k=1}^{K}\mathbf{D}_k^T\mathbf{V}_k^{-1}\mathbf{W}_k diag(\mathbf{R}_k)\mathbf{D}_k$ and

$$\widehat{\mathbf{V}}_{I,KC}(\widehat{\beta}_w) = \sum_{k=1}^{K}\frac{R_k}{\pi_k}\mathbf{D}_k^T\mathbf{V}_k^{-1}(\mathbf{I}_{N_k} - \widehat{\mathbf{A}}_{kk,w})^{-\frac{1}{2}}\widehat{\varepsilon}_{w,k}\widehat{\varepsilon}_{w,k}^T(\mathbf{I}_{N_k} - \widehat{\mathbf{A}}_{kk,w}^T)^{-\frac{1}{2}}\mathbf{V}_k^{-1}\mathbf{D}_k$$

and

$$\begin{aligned}\widehat{\mathbf{V}}_{II,KC}(\widehat{\beta}_w) &= \textstyle\sum_{k=1}^{K} B_{kk}\mathbf{D}_k^T\mathbf{V}_k^{-1}(\mathbf{I}_{N_k} - \widehat{\mathbf{A}}_{kk,w})^{-\frac{1}{2}}\widehat{\varepsilon}_{w,k}\widehat{\varepsilon}_{w,k}^T(\mathbf{I}_{N_k} - \widehat{\mathbf{A}}_{kk,w}^T)^{-\frac{1}{2}}\mathbf{V}_k^{-1}\mathbf{D}_k \\ &\quad + \textstyle\sum_{k=1}^{K}\sum_{k'\neq k}\mathbf{D}_k^T\mathbf{V}_k^{-1}diag(\widehat{\varepsilon}_{w,k})\mathbf{B}_{kk'}diag(\widehat{\varepsilon}_{w,k'})\mathbf{V}_{k'}^{-1}\mathbf{D}_{k'}\end{aligned}$$

where $B_{kk} = \pi_k^{-3}(\pi_k - \pi_k^2)R_k$, $\mathbf{B}_{kk'}$ is an $N_k \times N_{k'}$ matrix with all entries equal to $\frac{(\pi_{kk'} - \pi_k\pi_{k'})R_kR_{k'}}{\pi_{kk'}\pi_k\pi_{k'}}$ and $\widehat{\mathbf{A}}_{kk',w}(\beta) =$

$$\mathbf{D}_k(\textstyle\sum_{m=1}^{K}\mathbf{D}_m^T\mathbf{V}_m^{-1}\mathbf{W}_m diag(\mathbf{R}_m)\mathbf{D}_m)^{-1}\mathbf{D}_{k'}^T\mathbf{V}_{k'}^{-1}\mathbf{W}_{k'}diag(\mathbf{R}_{k'}) =$$
$$\mathbf{D}_k\widehat{\mathbf{H}}_w^{-1}(\beta)\mathbf{D}_{k'}^T\mathbf{V}_{k'}^{-1}\mathbf{W}_{k'}diag(\mathbf{R}_{k'}).$$

If the data arise from a cluster-stratified design, the correction for the naïve estimator of the variance that ignores the negative correlation in the sampling indicators is:

$$\widehat{Var}^*_{KC}[\widehat{\beta}_w] = \widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}\left\{\widehat{\mathbf{V}}_{I,KC}(\widehat{\beta}_w) + \widehat{\mathbf{V}}^*_{II,KC}(\widehat{\beta}_w)\right\}\widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1},$$

where

$$\widehat{\mathbf{V}}^{*}_{II,KC}(\widehat{\boldsymbol{\beta}}_{w}) \;=\; \sum_{k=1}^{K} B_{kk}\mathbf{D}_{k}^{T}\mathbf{V}_{k}^{-1}(\mathbf{I}_{k}-\widehat{\mathbf{A}}_{kk,w})^{-\frac{1}{2}}\widehat{\boldsymbol{\varepsilon}}_{w,k}\widehat{\boldsymbol{\varepsilon}}_{w,k}^{T}(\mathbf{I}_{k}-\widehat{\mathbf{A}}_{kk,w}^{T})^{-\frac{1}{2}}\mathbf{V}_{k}^{-1}\mathbf{D}_{k}.$$

This is also the corrected variance estimator for data arising from a Poisson sampling design.

A.6 Fay and Graubard-type Variance Bias Correction

Do a Taylor series expansion of U_k about $\widehat{\beta}_w$:

$$U_k \approx U_k(\widehat{\beta}_w) + \left.\frac{\partial U_k}{\partial \beta}\right|_{\beta=\widehat{\beta}_w} (\beta - \widehat{\beta}_w) = U_k(\widehat{\beta}_w) - H_k(\widehat{\beta}_w)(\beta - \widehat{\beta}_w) \tag{A.17}$$

where $H_k = -[\frac{\partial U_k}{\partial \beta}] = \mathbf{D}_k^T \mathbf{V}_k^{-1} \mathbf{D}_k$. Now, summing over the clusters, $k = 1, \ldots, K$, yields:

$$\sum_{k=1}^K U_k \approx \sum_{k=1}^K \widehat{U}_k - \sum_{k=1}^K H_k(\widehat{\beta}_w)(\beta - \widehat{\beta}_w)$$

$$\sum_{k=1}^K \widehat{U}_k - \sum_{k=1}^K U_k \approx \sum_{k=1}^K H_k(\widehat{\beta}_w)(\beta - \widehat{\beta}_w)$$

$$\sum_{k=1}^K U_k - \sum_{k=1}^K \widehat{U}_k \approx \sum_{k=1}^K H_k(\widehat{\beta}_w)(\widehat{\beta}_w - \beta)$$

From which it follows that

$$(\widehat{\beta}_w - \beta) \approx H(\widehat{\beta}_w)^{-1} \sum_{k=1}^K U_k = H(\widehat{\beta}_w)^{-1} \mathcal{U} \tag{A.18}$$

where $\mathbf{H}(\beta) = \sum_{k=1}^K H_k(\beta)$. We follow the approach of Fay and Graubard (2001) and use approximations (A.17) and (A.18) to obtain an approximation for $E[\widehat{U}_k \widehat{U}_k^T | \mathbf{R}]$:

$$
\begin{aligned}
&\widehat{U}_k\widehat{U}_k^T \\
&\quad \approx (U_k + \mathbf{H}_k(\widehat{\beta}_w)(\beta - \widehat{\beta}_w))(U_k^T + (\beta - \widehat{\beta}_w)^T\mathbf{H}_k^T(\widehat{\beta}_w)) \\
&\quad = U_kU_k^T + U_k(\beta - \widehat{\beta}_w)^T\mathbf{H}_k^T(\widehat{\beta}_w) + \mathbf{H}_k(\widehat{\beta}_w)(\beta - \widehat{\beta}_w)U_k^T + \mathbf{H}_k(\widehat{\beta}_w)(\beta - \widehat{\beta}_w)(\beta - \widehat{\beta}_w)^T\mathbf{H}_k(\widehat{\beta}_w)^T \\
&\quad = U_kU_k^T - U_k(\sum_{k=1}^{K} U_k)^T(\mathbf{H}(\widehat{\beta}_w)^{-1})^T\mathbf{H}_k(\widehat{\beta}_w)^T - \mathbf{H}_k(\widehat{\beta}_w)\mathbf{H}^{-1}(\widehat{\beta}_w)(\sum_{k=1}^{K} U_k)U_k^T \\
&\quad + \mathbf{H}_k(\widehat{\beta}_w)\mathbf{H}(\widehat{\beta}_w)^{-1}(\sum_{k=1}^{K} U_k)(\sum_{k=1}^{K} U_k)^T\mathbf{H}^{-1}(\widehat{\beta}_w)\mathbf{H}_k(\widehat{\beta}_w)^T \\
&\quad = U_kU_k^T - [U_kU_k^T(\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w)^T + U_k(\sum_{j\neq k} U_k)^T(\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w)^T] \\
&\quad - [\mathbf{H}_k(\widehat{\beta}_w)(\mathbf{H}^{-1}(\widehat{\beta}_w))^TU_kU_k^T + \mathbf{H}_k(\widehat{\beta}_w)(\mathbf{H}(\widehat{\beta}_w)^{-1})^T(\sum_{j\neq k} U_j)U_k^T] \\
&\quad + \mathbf{H}_k(\widehat{\beta}_w)(\mathbf{H}^{-1}(\widehat{\beta}_w))U_kU_k^T(\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w)^T \\
&\quad + \mathbf{H}_k(\widehat{\beta}_w)\mathbf{H}^{-1}(\widehat{\beta}_w)(\sum_{j\neq k} U_jU_j^T)(\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w) \\
&\quad + \mathbf{H}_k(\widehat{\beta}_w)\mathbf{H}(\widehat{\beta}_w)^{-1}(\sum_{k=1}^{K} U_k\sum_{j\neq k} U_j^T)(\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w)
\end{aligned}
$$

Now, taking expectations conditional on **R** (supressing the conditional notation on the right-hand side of the expression below) yields:

$$
\begin{aligned}
&E[\widehat{U}_k\widehat{U}_k^T|\mathbf{R}] \\
&\quad\approx E[U_kU_k^T] - E[U_kU_k^T](\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w)^T - E[U_k(\sum_{j\neq k}U_j)^T](\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w)^T \\
&\quad- [\mathbf{H}_k(\widehat{\beta}_w)(\mathbf{H}^{-1}(\widehat{\beta}_w))^TE[U_kU_k^T] - \mathbf{H}_k(\widehat{\beta}_w)(\mathbf{H}^{-1}(\widehat{\beta}_w))^TE[(\sum_{j\neq k}U_j)U_k^T] \\
&\quad+ \mathbf{H}_k(\widehat{\beta}_w)(\mathbf{H}(\widehat{\beta}_w)^{-1})E[U_kU_k^T](\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w)^T \\
&\quad+ \mathbf{H}_k(\widehat{\beta}_w)\mathbf{H}^{-1}(\widehat{\beta}_w)(\sum_{j\neq k}E[U_jU_j^T)](\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w) \\
&\quad+ \mathbf{H}_k(\widehat{\beta}_w)\mathbf{H}(\widehat{\beta}_w)^{-1})E[(\sum_{k=1}^{K}\sum_{j\neq k}U_kU_j^T)](\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w) \\
&\quad= \Psi_k - \Psi_k(\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w)^T - \mathbf{H}_k(\widehat{\beta}_w)(\mathbf{H}^{-1}(\widehat{\beta}_w))^T\Psi_k \\
&\quad+ \mathbf{H}_k(\widehat{\beta}_w)(\mathbf{H}^{-1}(\widehat{\beta}_w))\Psi_k(\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w)^T \\
&\quad+ \mathbf{H}_k(\widehat{\beta}_w)(\mathbf{H}^{-1}(\widehat{\beta}_w))(\sum_{j\neq k}\Psi_j)(\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w) \\
&\quad= (I_p - \mathbf{H}_k(\widehat{\beta}_w)(\mathbf{H}^{-1}(\widehat{\beta}_w)))\Psi_k(I_p - \mathbf{H}_k(\widehat{\beta}_w)(\mathbf{H}^{-1})(\widehat{\beta}_w))^T \\
&\quad+ \mathbf{H}_k(\widehat{\beta}_w)(\mathbf{H}^{-1}(\widehat{\beta}_w))(\sum_{j\neq k}\Psi_j)(\mathbf{H}^{-1}(\widehat{\beta}_w))^T\mathbf{H}_k(\widehat{\beta}_w)^T
\end{aligned}
$$

where $\Psi_k = E[U_kU_k^T]$. Suppose that $\Psi_k \approx c\mathbf{H}_k(\widehat{\beta}_w)$, with c some constant and let $\widehat{\mathbf{H}}_k = \mathbf{H}_k(\widehat{\beta}_w)$:

$$
\begin{aligned}
&E[\widehat{U}_k\widehat{U}_k^T|\mathbf{R}] \\
&\quad\approx c\widehat{\mathbf{H}}_k - c\widehat{\mathbf{H}}_k(\widehat{\mathbf{H}}^{-1})^T\widehat{\mathbf{H}}_k^T - c\widehat{\mathbf{H}}_k(\widehat{\mathbf{H}}_k^{-1})\widehat{\mathbf{H}}_k + c\widehat{\mathbf{H}}_k(\widehat{\mathbf{H}}^{-1})\widehat{\mathbf{H}}_k(\widehat{\mathbf{H}}^{-1})^T\widehat{\mathbf{H}}_k^T \\
&\quad+ c\widehat{\mathbf{H}}_k(\widehat{\mathbf{H}}^{-1})(\sum_{j\neq k}\widehat{\mathbf{H}}_j)(\widehat{\mathbf{H}}^{-1})^T\widehat{\mathbf{H}}_k^T \\
&\quad= c\widehat{\mathbf{H}}_k - c\widehat{\mathbf{H}}_k(\widehat{\mathbf{H}}^{-1})^T\widehat{\mathbf{H}}_k^T - c\widehat{\mathbf{H}}_k(\widehat{\mathbf{H}}^{-1})\widehat{\mathbf{H}}_k + c\widehat{\mathbf{H}}_k(\widehat{\mathbf{H}}^{-1})(\sum_{k=1}^{K}\widehat{\mathbf{H}}_j)(\widehat{\mathbf{H}}^{-1})^T\widehat{\mathbf{H}}_k^T \\
&\quad= c\widehat{\mathbf{H}}_k - c\widehat{\mathbf{H}}_k(\widehat{\mathbf{H}}^{-1})^T\widehat{\mathbf{H}}_k^T - c\widehat{\mathbf{H}}_k(\widehat{\mathbf{H}}^{-1})\widehat{\mathbf{H}}_k + c\widehat{\mathbf{H}}_k(\widehat{\mathbf{H}}^{-1})\widehat{\mathbf{H}}_k^T \\
&\quad= c\widehat{\mathbf{H}}_k - c\widehat{\mathbf{H}}_k(\widehat{\mathbf{H}}^{-1})\widehat{\mathbf{H}}_k \\
&\quad= c\widehat{\mathbf{H}}_k(I_p - (\widehat{\mathbf{H}}^{-1})\widehat{\mathbf{H}}_k) \\
&\quad\approx \Psi_k(I_p - (\widehat{\mathbf{H}}^{-1})\widehat{\mathbf{H}}_k) \approx (I_p - \widehat{\mathbf{H}}_k(\widehat{\mathbf{H}}^{-1})))\Psi_k
\end{aligned}
$$

Therefore, estimate Ψ_k with $\widehat{\Psi}_k$=$\mathbf{F}_{k,w}\widehat{U}_k\widehat{U}_k^T\mathbf{F}_{k,w}^T$, where $\mathbf{F}_{k,w}$ is a $p \times p$ diagonal matrix with the jj^{th} element equal to $(1 - min(b, (\widehat{\mathbf{H}}_k\widehat{\mathbf{H}}^{-1})_{jj})^{-\frac{1}{2}}$. We estimate $\mathbf{F}_{k,w}$ with $\widehat{\mathbf{F}}_{k,w}$, a $p \times p$ diagonal matrix with the jj^{th} element equal to $(1 - min(b, (\widehat{\mathbf{H}}_{w,k}\widehat{\mathbf{H}}_w^{-1})_{jj})^{-\frac{1}{2}}$, where
$\widehat{\mathbf{H}}_w(\beta) = -\sum_{k=1}^{K}\widehat{\mathbf{H}}_{w,k}(\beta) = -\sum_{k=1}^{K}\mathbf{D}_k^T\mathbf{V}_k^{-1}\mathbf{W}_k diag(\mathbf{R}_k)\mathbf{D}_k$, and b is set to 0.75.
The bias-corrected sandwich estimator is then

$$
\widehat{Var}_{FG}[\widehat{\beta}_w] = \widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}\left\{\widehat{\mathbf{V}}_{I,FG}(\widehat{\beta}_w) + \widehat{\mathbf{V}}_{II,FG}(\widehat{\beta}_w)\right\}\widehat{\mathbf{H}}_w(\widehat{\beta}_w)^{-1}, \tag{A.19}
$$

where

$$
\widehat{\mathbf{V}}_{I,FG}(\widehat{\beta}_w) = \sum_{k=1}^{K}\frac{R_k}{\pi_k}\widehat{\mathbf{F}}_{k,w}\mathbf{D}_k^T\mathbf{V}_k^{-1}\widehat{\varepsilon}_{w,k}\widehat{\varepsilon}_{w,k}^T\mathbf{V}_k^{-1}\mathbf{D}_k(\widehat{\mathbf{F}}_{k,w})^T
$$

$$\widehat{\mathbf{V}}_{II,FG}(\widehat{\boldsymbol{\beta}}_w) = \sum_{k=1}^{K} B_{kk}\widehat{\mathbf{F}}_{k,w}\mathbf{D}_k^T\mathbf{V}_k^{-1}\widehat{\boldsymbol{\varepsilon}}_{w,k}\widehat{\boldsymbol{\varepsilon}}_{w,k}^T\mathbf{V}_k^{-1}\mathbf{D}_k(\widehat{\mathbf{F}}_{k,w})^T + \sum_{k=1}^{K}\sum_{k'\neq k}\mathbf{D}_k^T\mathbf{V}_k^{-1}diag(\widehat{\boldsymbol{\varepsilon}}_{w,k})\mathbf{B}_{kk'}diag(\widehat{\boldsymbol{\varepsilon}}_{w,k'})\mathbf{V}_{k'}^{-1}\mathbf{D}_{k'}$$

where $\mathbf{B}_{kk'}$ is an $N_k \times N_{k'}$ matrix with all entries equal to $\frac{(\pi_{kk'}-\pi_k\pi_{k'})R_kR_{k'}}{\pi_{kk'}\pi_k\pi_{k'}}$.

If the data arise from a cluster-stratified design, the correction for the naïve estimator of the variance that ignores the negative correlation in the sampling indicators is:

$$\widehat{Var}^*_{FG}[\widehat{\boldsymbol{\beta}}_w] = \widehat{\mathbf{H}}_w(\widehat{\boldsymbol{\beta}}_w)^{-1}\left\{\widehat{\mathbf{V}}_{I,FG}(\widehat{\boldsymbol{\beta}}_w) + \widehat{\mathbf{V}}^*_{II,FG}(\widehat{\boldsymbol{\beta}}_w)\right\}\widehat{\mathbf{H}}_w(\widehat{\boldsymbol{\beta}}_w)^{-1},$$

where

$$\widehat{\mathbf{V}}^*_{II,FG}(\widehat{\boldsymbol{\beta}}_w) = \sum_{k=1}^{K} B_{kk}\widehat{\mathbf{F}}_{k,w}\mathbf{D}_k^T\mathbf{V}_k^{-1}\widehat{\boldsymbol{\varepsilon}}_{w,k}\widehat{\boldsymbol{\varepsilon}}_{w,k}^T\mathbf{V}_k^{-1}\mathbf{D}_k(\widehat{\mathbf{F}}_{k,w})^T.$$

A.7 Smoothed Bootstrap

We adapt the smoothed bootstrap approach proposed by [32] for the complete data setting. When data has been collected through a cluster-based outcome-dependent sampling scheme, the estimates, $\widehat{\beta}_w$, are obtained by solving

$$\mathcal{U}_w = \sum_{k=1}^{K} \frac{R_k}{\pi_k} \mathbf{D}_k^T \mathbf{V}_k^{-1} (\mathbf{Y}_k - \boldsymbol{\mu}_k) \tag{A.20}$$

One can obtain smoothed bootstrap replicates, $\widetilde{\beta}_w$, by solving the following perturbed inverse-probability-weighted generalized estimating equations:

$$\widetilde{\mathcal{U}}_w = \sum_{k=1}^{K} w_k \frac{R_k}{\pi_k} \mathbf{D}_k^T(\widehat{\beta}_w) \mathbf{V}_k^{-1}(\widehat{\beta}_w)(\mathbf{Y}_k(\beta) - \boldsymbol{\mu}_k(\beta)) \tag{A.21}$$

where the w_k are taken to be independent realizations from a distribution with unit mean and unit variance. If the clusters have been sampled through Poisson sampling, the known π_k are used in (A.21). The bootstrap replicates are then obtained by solving $\widetilde{\mathcal{U}}_w = 0$:

$$\widetilde{\beta}_w^{(0)} = \widehat{\beta}_w$$

$$\widetilde{\beta}_w^{(1)} = \widetilde{\beta}_w^{(0)} + (\textstyle\sum_{k=1}^{K} w_k \mathbf{H}_{w,k}(\widehat{\beta}_w))^{-1} (\sum_k \frac{R_k}{\pi_k} w_k \mathbf{D}_k^T(\widetilde{\beta}_w^{(0)}) \mathbf{V}_k^{-1}(\widetilde{\beta}_w^{(0)})(\mathbf{Y}_k(\widetilde{\beta}_w^{(0)}) - \boldsymbol{\mu}_k(\widetilde{\beta}_w^{(0)}))$$

$$\widetilde{\beta}_w^{(2)} = \widetilde{\beta}_w^{(1)} + (\textstyle\sum_{k=1}^{K} w_k \mathbf{H}_{w,k}(\widehat{\beta}_w))^{-1} (\sum_k \frac{R_k}{\pi_k} w_k \mathbf{D}_k^T(\widetilde{\beta}_w^{(1)}) \mathbf{V}_k^{-1}(\widetilde{\beta}_w^{(1)})(\mathbf{Y}_k(\widetilde{\beta}_w^{(1)}) - \boldsymbol{\mu}_k(\widetilde{\beta}_w^{(1)}))$$

where $H_{w,k}(\beta) = \mathbf{D}_k^T(\beta) \mathbf{V}_k^{-1}(\beta) \mathbf{W}_k diag(\mathbf{R}_k) \mathbf{D}_k(\beta)$.

If the clusters are sampled through a cluster-stratified design, the selection indicators R_k, $R_{k'}$ are negatively correlated for clusters k and k' belonging to the same stratum. To account for this, we take the approach of Cai and Zheng (2013) [13]: namely, perturbing the selection indicators as though they were independent, and using estimated selection probabilities rather than the known probabilities to induce a correlation among the selection indicators. In this case, this would involve treating

the $R'_k s$ as though the clusters were sampled via Poisson sampling: cluster k has a probability π_k of being sampled, where π_k is determined by the stratum j to which the cluster belongs. Selection of cluster k into the sample is determined by a Bernoulli trial with probability π_k. In such a setting, π_k can be estimated empirically as k_j/K_j, where k_j is the number of observed clusters sampled from stratum j. Furthermore, the joint probability of selection, $\pi_{kk'}$, for $k \neq k'$ can be empirically estimated as follows: $\widehat{\pi}_{kk'} = \frac{k_j}{K_j}\frac{k_j-1}{K_j-1}$ if clusters k and k' belong to the same stratum, and $\widehat{\pi}_{kk'} = (\frac{k_j}{K_j})^2$ if clusters k and k' belong to different strata. Note that the asymptotic variance of $\widehat{\beta}_w$ under Poisson sampling with *estimated* weights is the equal to the asymptotic variance of $\widehat{\beta}_w$ under cluster-stratified sampling with *known* weights. Below are the steps for the smoothed bootstrap procedure when the clusters are selected through a cluster-stratified sampling scheme.

1. Generate K_s realizations from a distribution with unit mean and variance: $w_k, k = 1, \ldots, K_s$
2. Use the w_k to define perturbed weights, $\frac{R_k w_k}{\widehat{\pi}_k}$, where $\widehat{\pi}_k = k_j/K_j$
3. Solve the perturbed estimating equations for $\widetilde{\beta}_w$ using the iterative procedure described above
4. repeat (1-3) B times

If the clusters are instead selected through a poisson sampling scheme, the above steps may be followed, replacing $\widehat{\pi}_k$ with the known probabilities of selection π_k.
In the simulation results presented in the next section, we used the Gam(4, 2) - 1 distribution to generate the w_k, and set $B = 1000$. Confidence intervals were then constructed as

$$(\widehat{\beta}_w - 1.96 * sd(\widetilde{\beta}_{w,1:B}), \widehat{\beta}_w + 1.96 * sd(\widetilde{\beta}_{w,1:B}))$$

A.8 Simulation Results: Absolute Bias in Point Estimates

A.8.1 Simulation 1

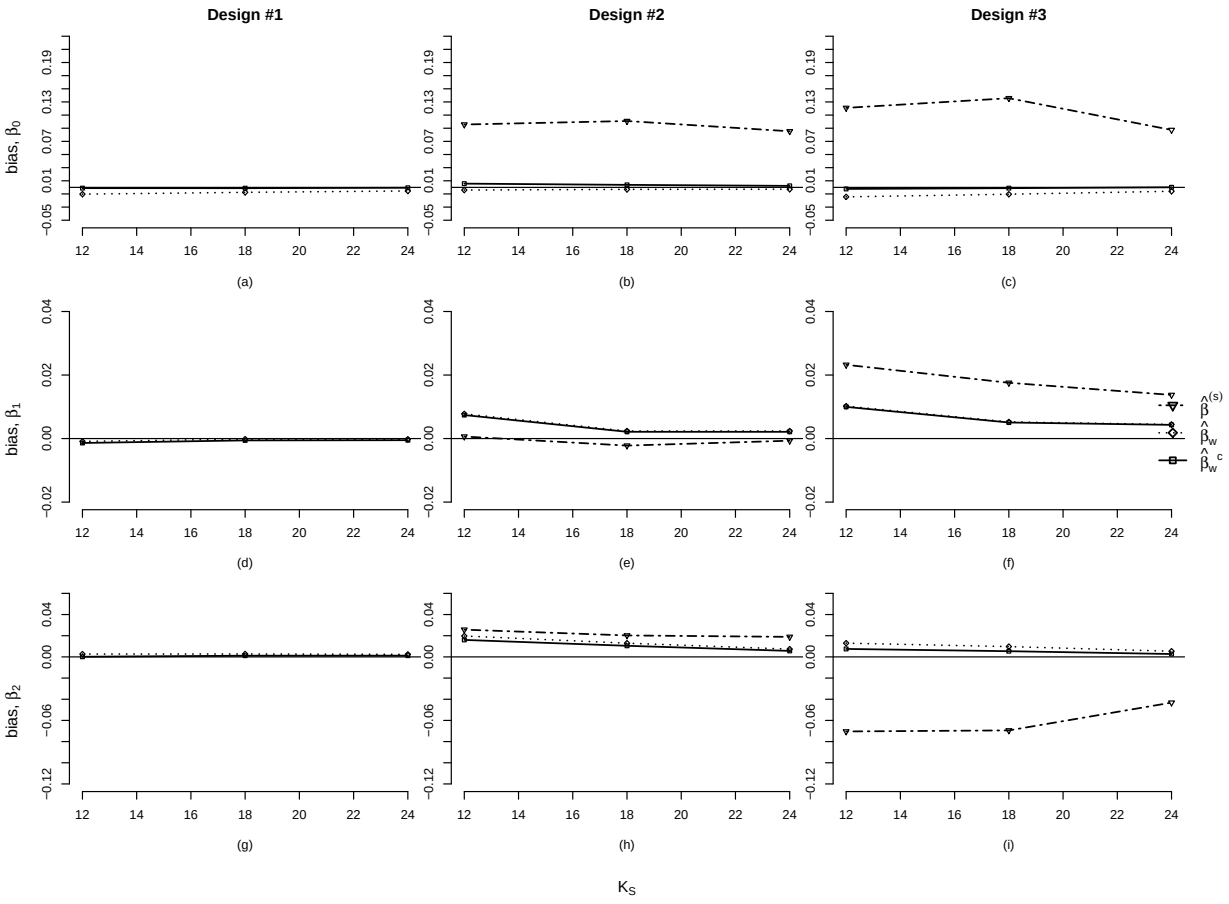

Figure A.1: Absolute bias in point estimates, σ_V=0.25, working independence

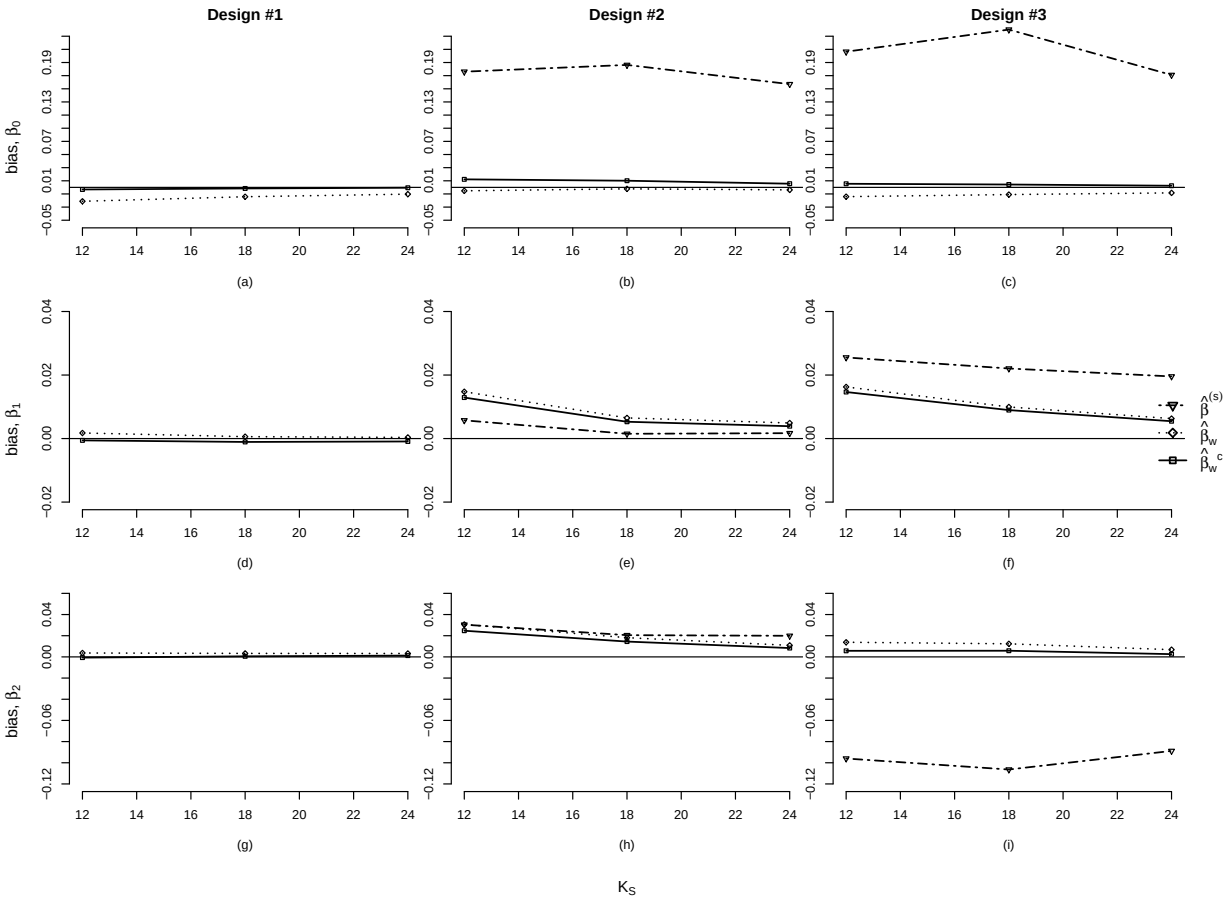

Figure A.2: Absolute bias in mean point estimates, σ_V=0.5, working independence

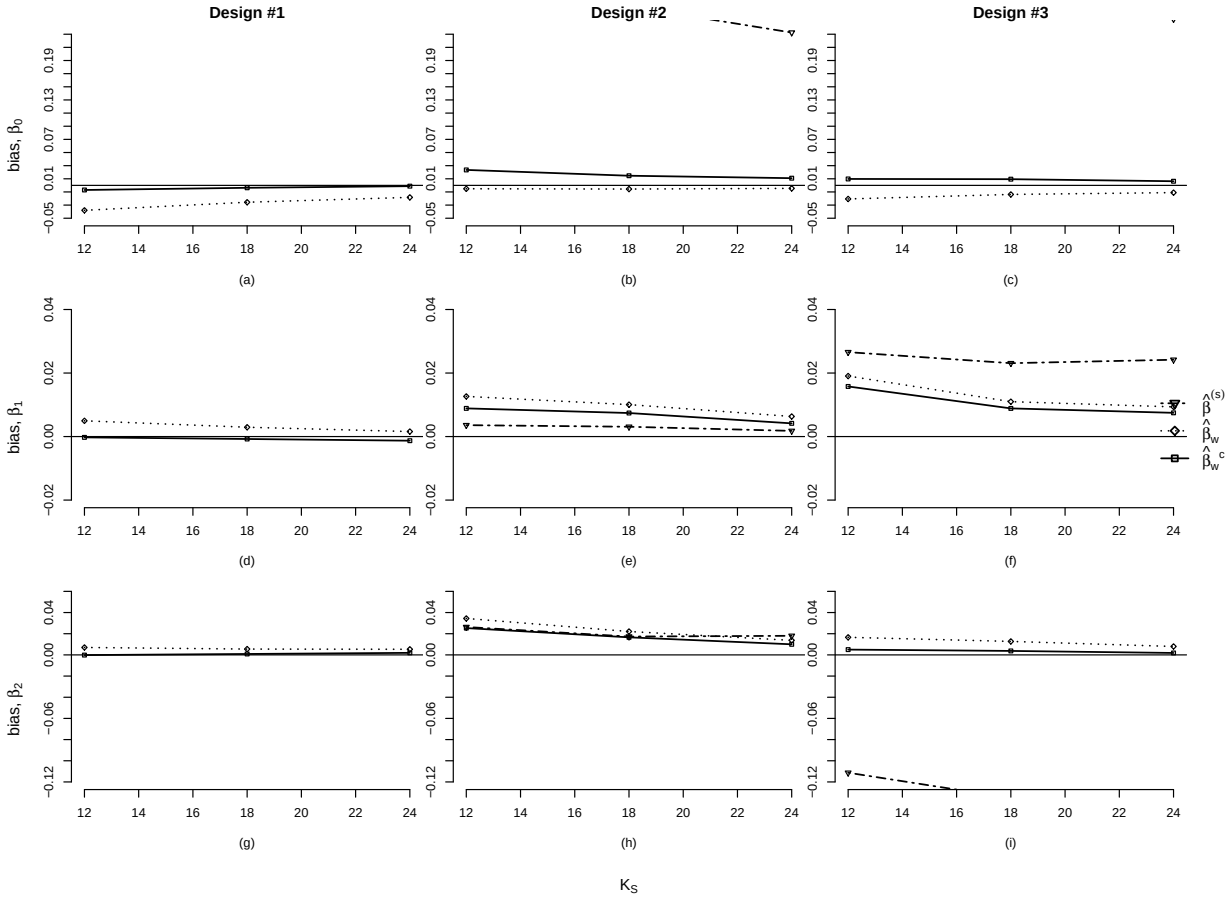

Figure A.3: Absolute bias in mean point estimates, σ_V=0.75, working independence

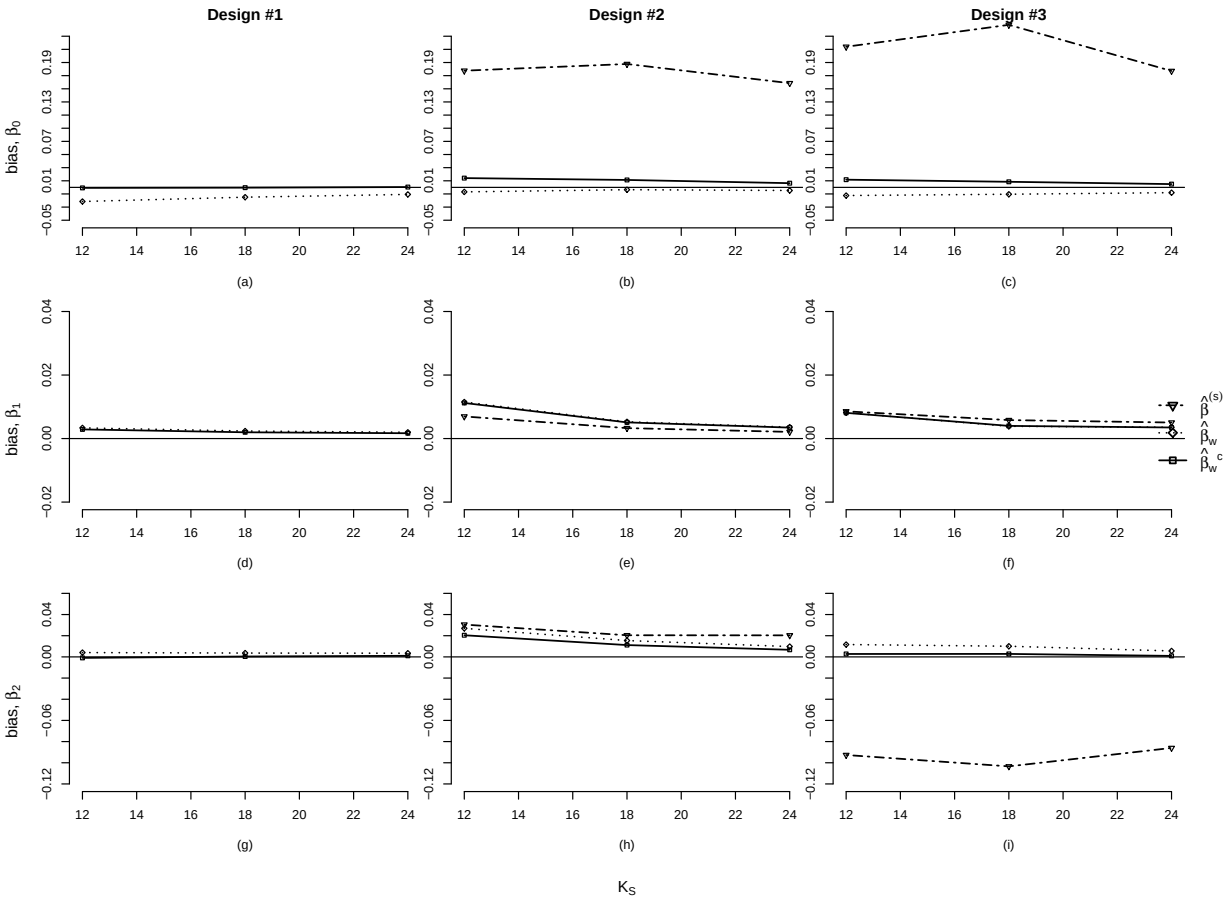

Figure A.4: Absolute bias in mean point estimates, σ_V=0.5, working exchangeable

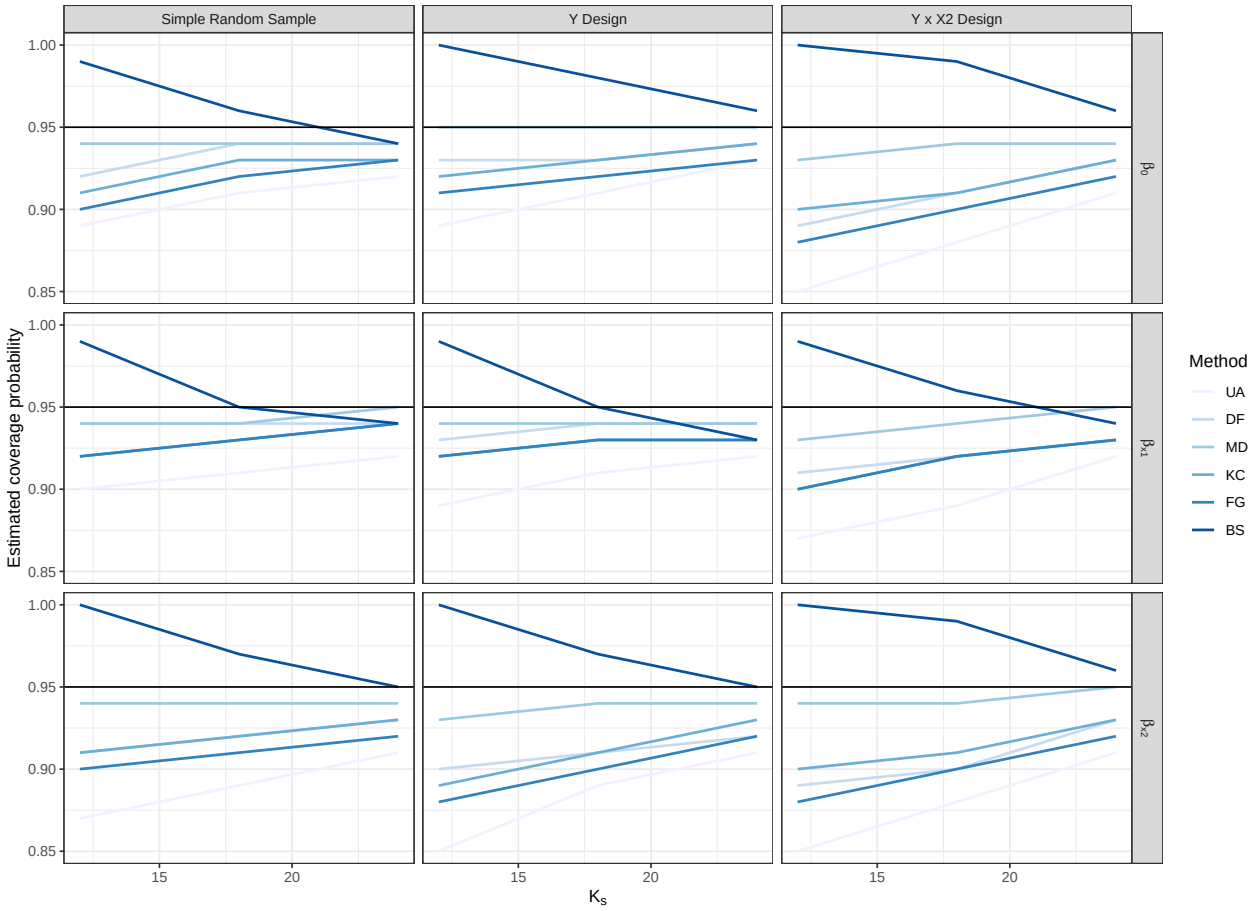

Figure A.5: Estimated coverage probabilities with $\widehat{\beta}_w$, using normal distribution for confidence interval construction, σ_Y=0.5

A.9 Simulation Results: Coverage Probabilities

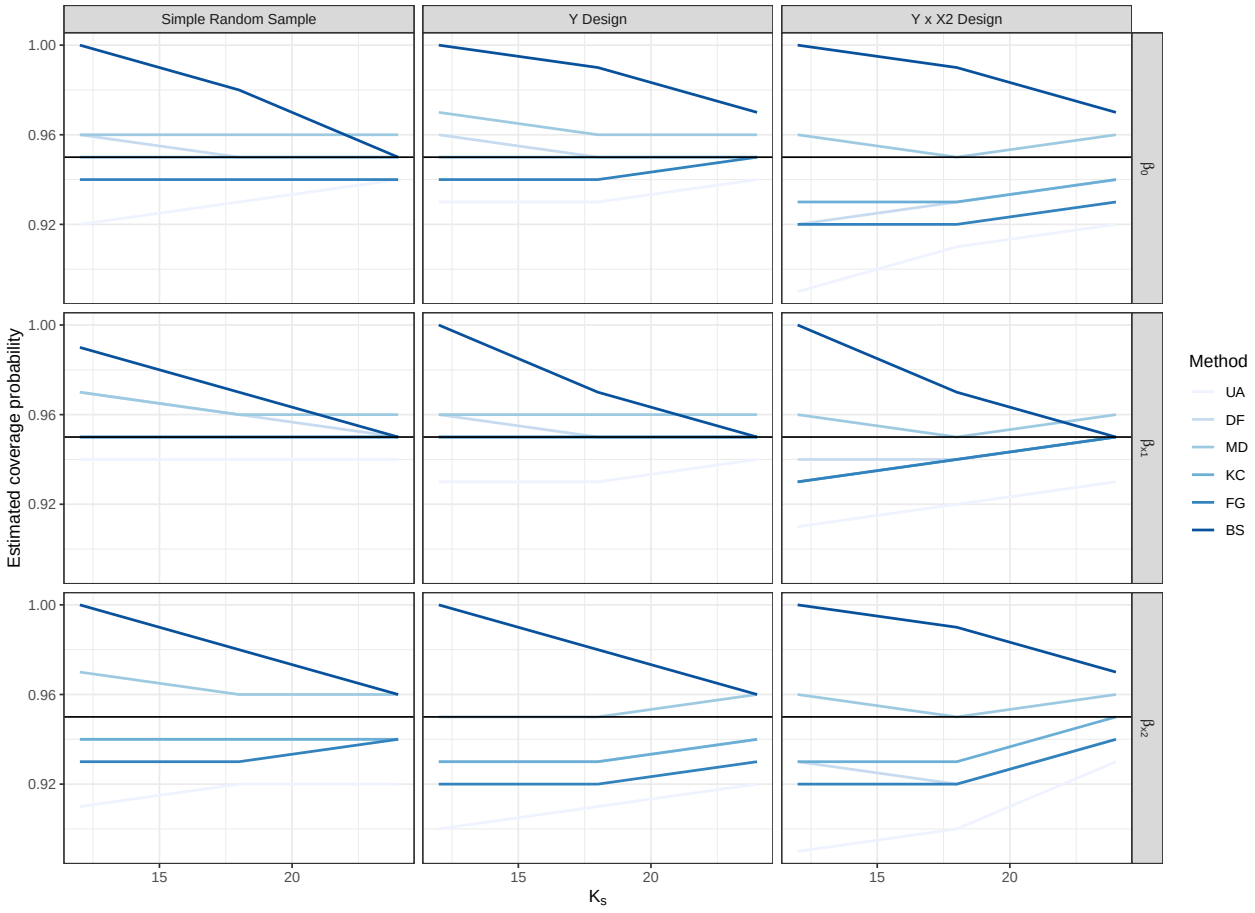

Figure A.6: Estimated coverage probabilities with $\widehat{\beta}_w$, using t distribution for confidence interval construction, σ_V=0.5

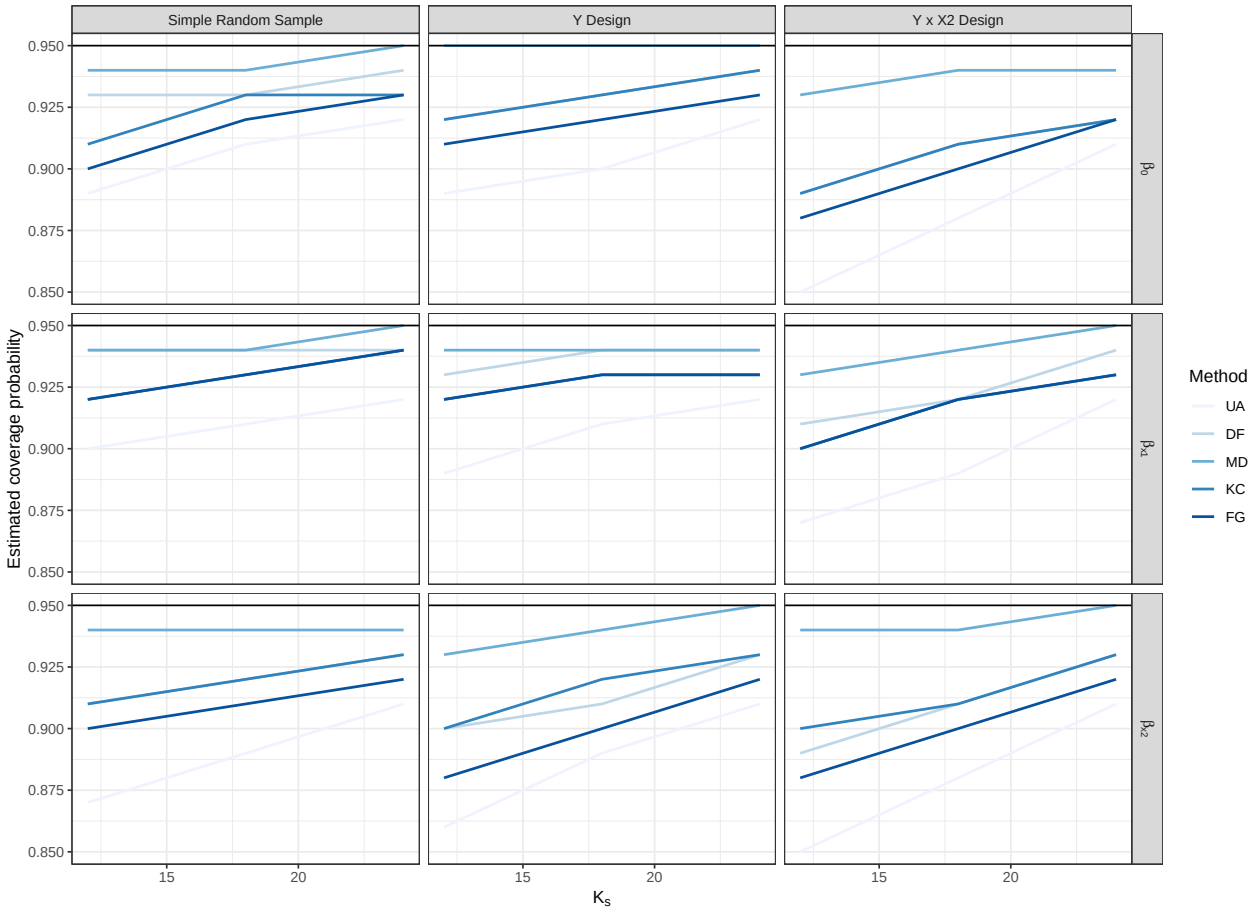

Figure A.7: Estimated coverage probabilities with $\widehat{\beta}_w^{\zeta}$, using normal distribution for confidence interval construction, σ_Y=0.5

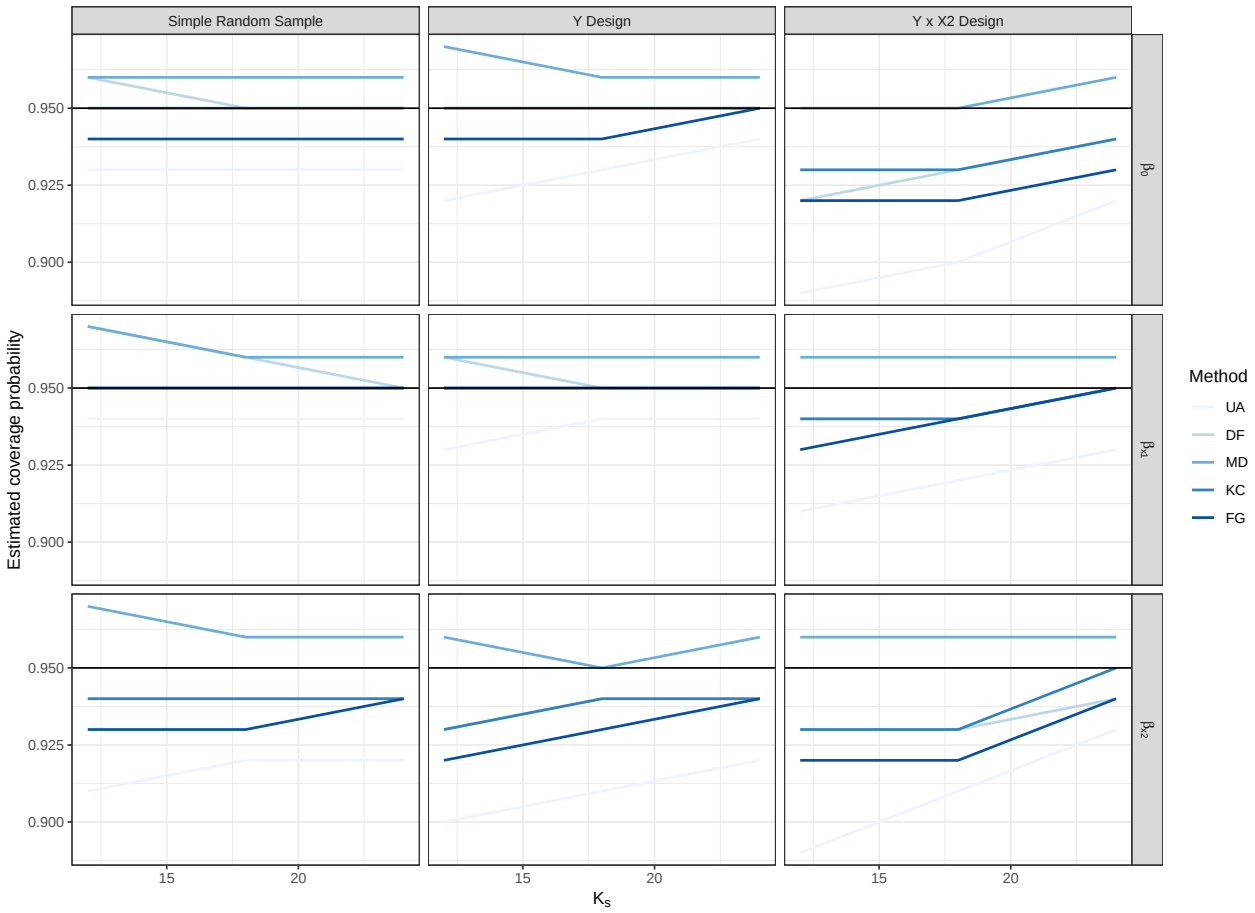

Figure A.8: Estimated coverage probabilities with $\widehat{\beta}_w^c$, using t distribution for confidence interval construction, σ_V=0.5

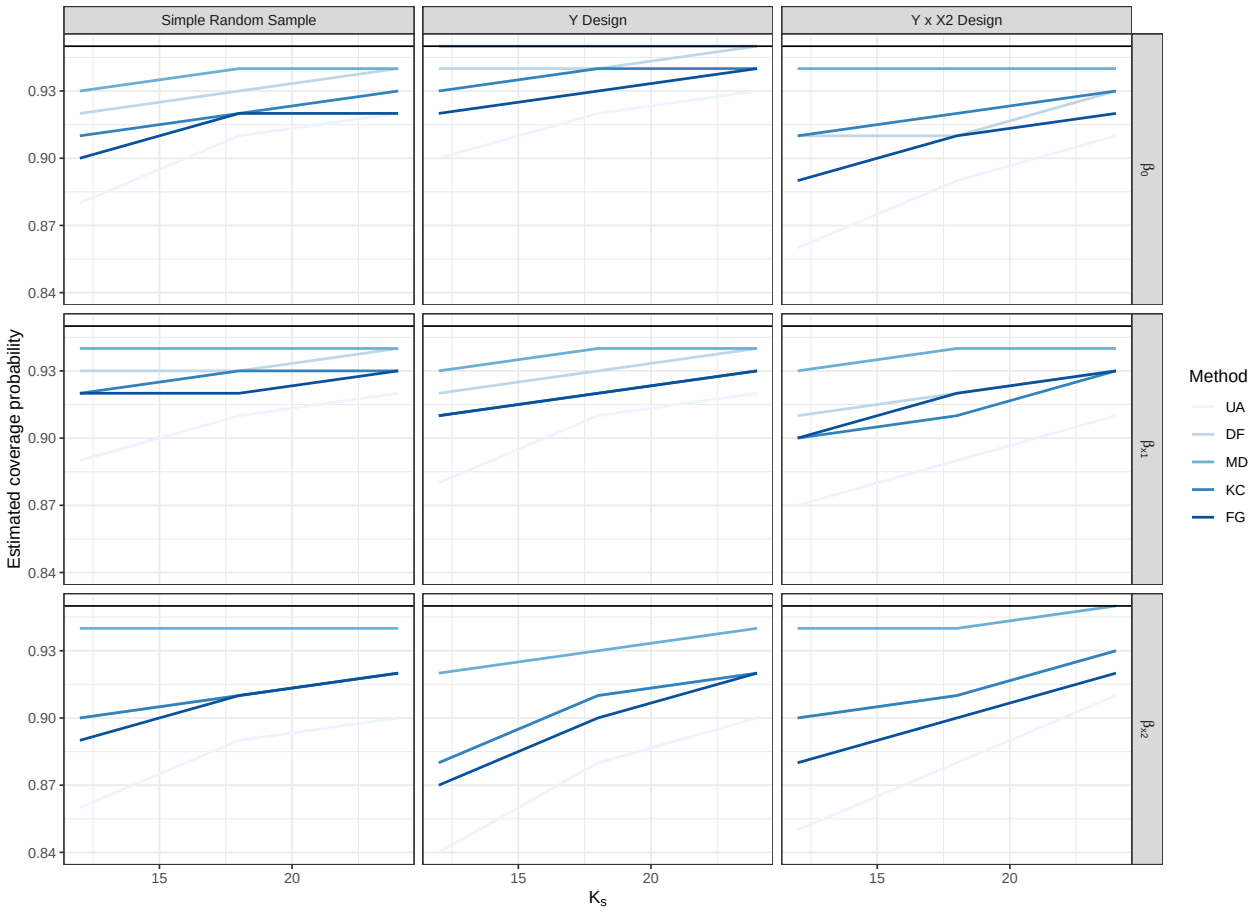

Figure A.9: Estimated coverage probabilities with $\widehat{\beta}_w$, using normal distribution for confidence interval construction, ignoring negative correlation, σ_V=0.5

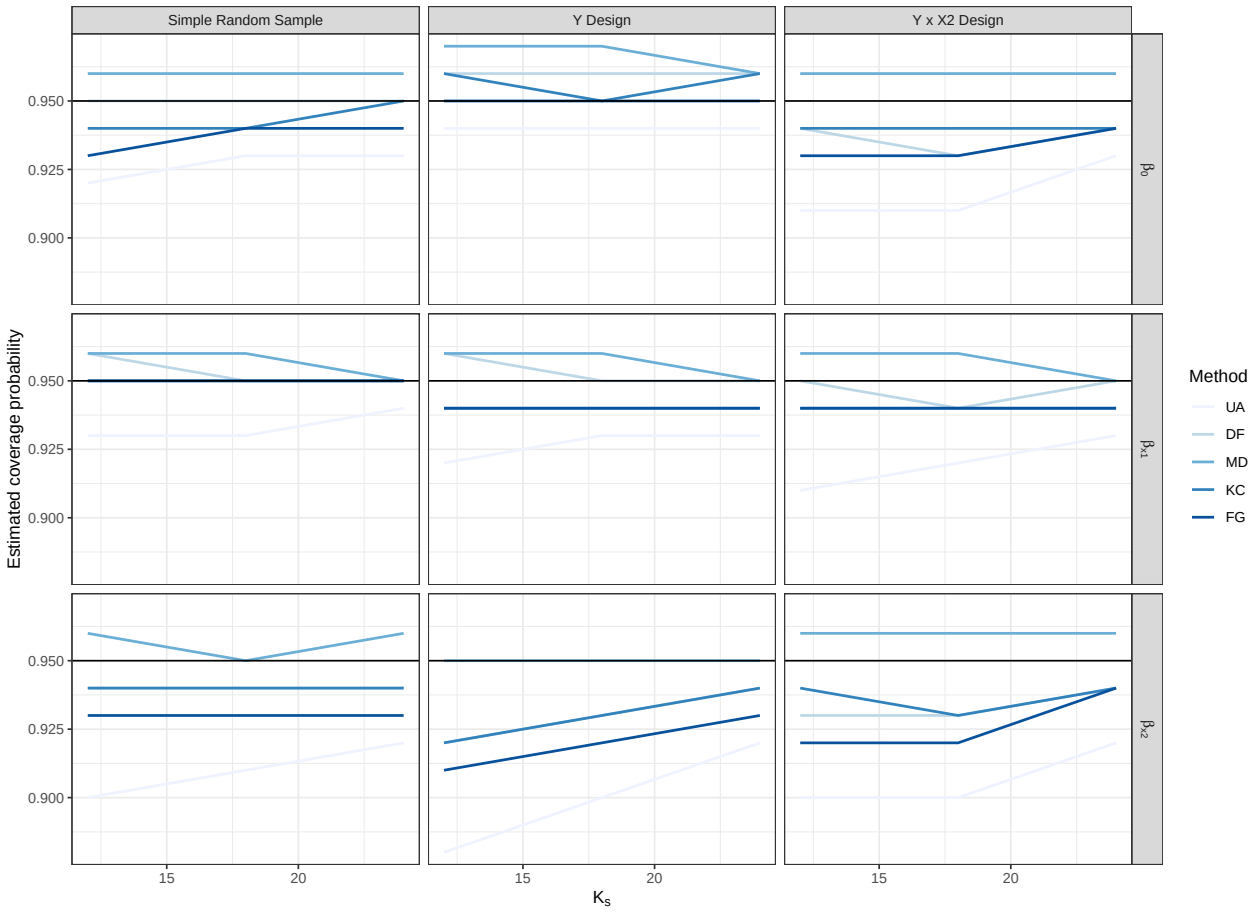

Figure A.10: Estimated coverage probabilities with $\widehat{\beta}_w$, using t distribution for confidence interval construction, ignoring negative correlation, σ_V=0.5

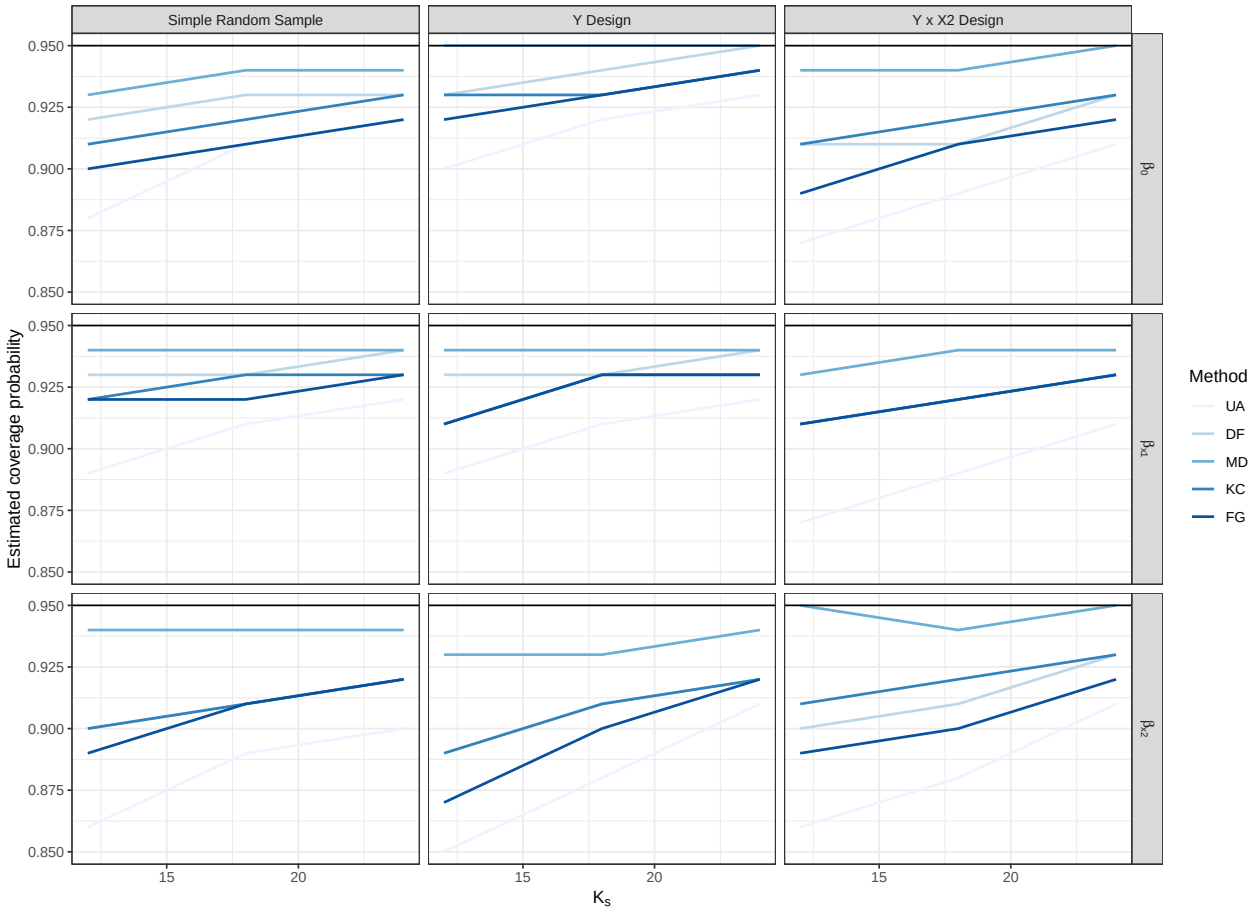

Figure A.11: Estimated coverage probabilities with $\widehat{\beta}_w^c$, using normal distribution for confidence interval construction, ignoring negative correlation, σ_V=0.5

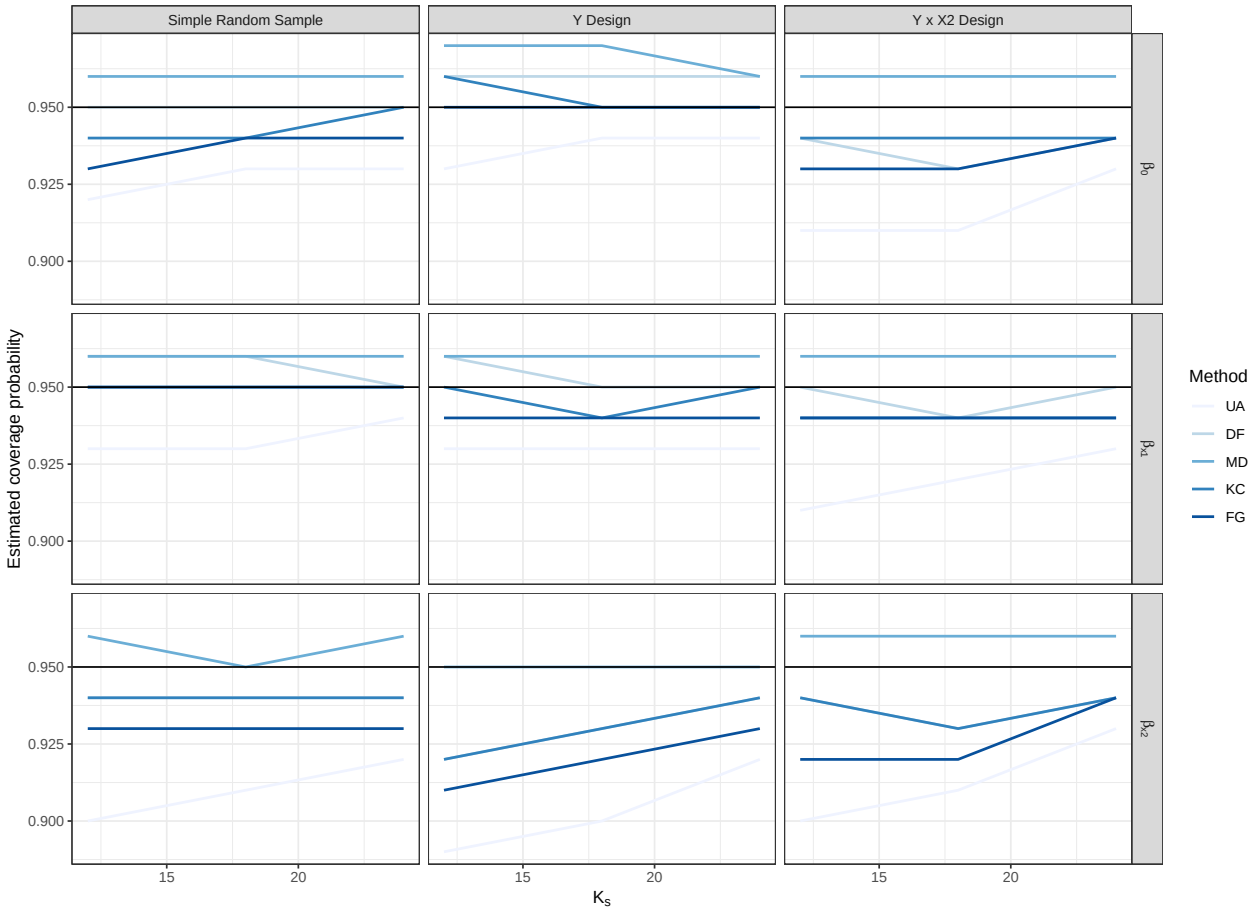

Figure A.12: Estimated coverage probabilities with $\widehat{\beta}_w^c$, using t distribution for confidence interval construction, ignoring negative correlation, σ_V=0.5

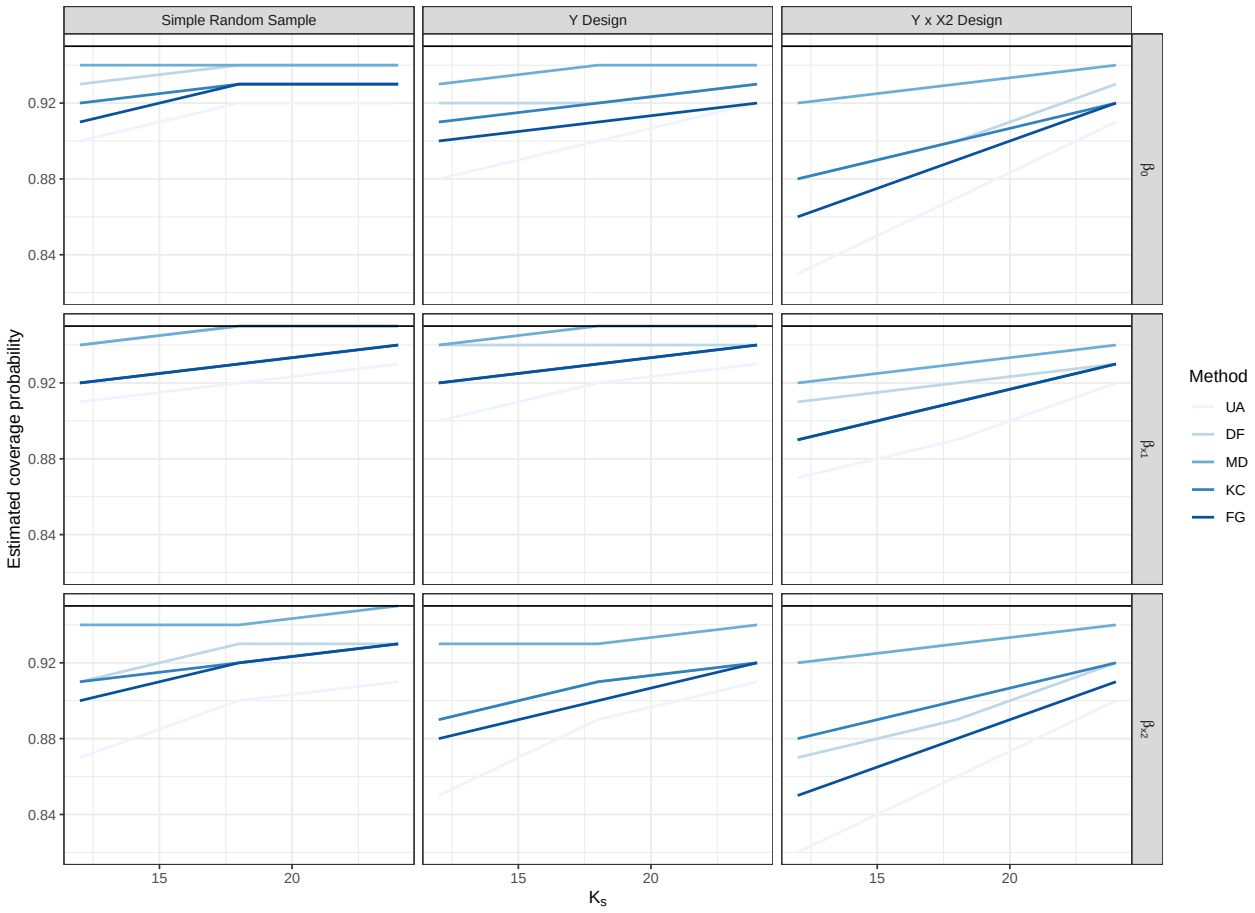

Figure A.13: Estimated coverage probabilities with $\widehat{\beta}_w$, using normal distribution for confidence interval construction, σ_V=0.25

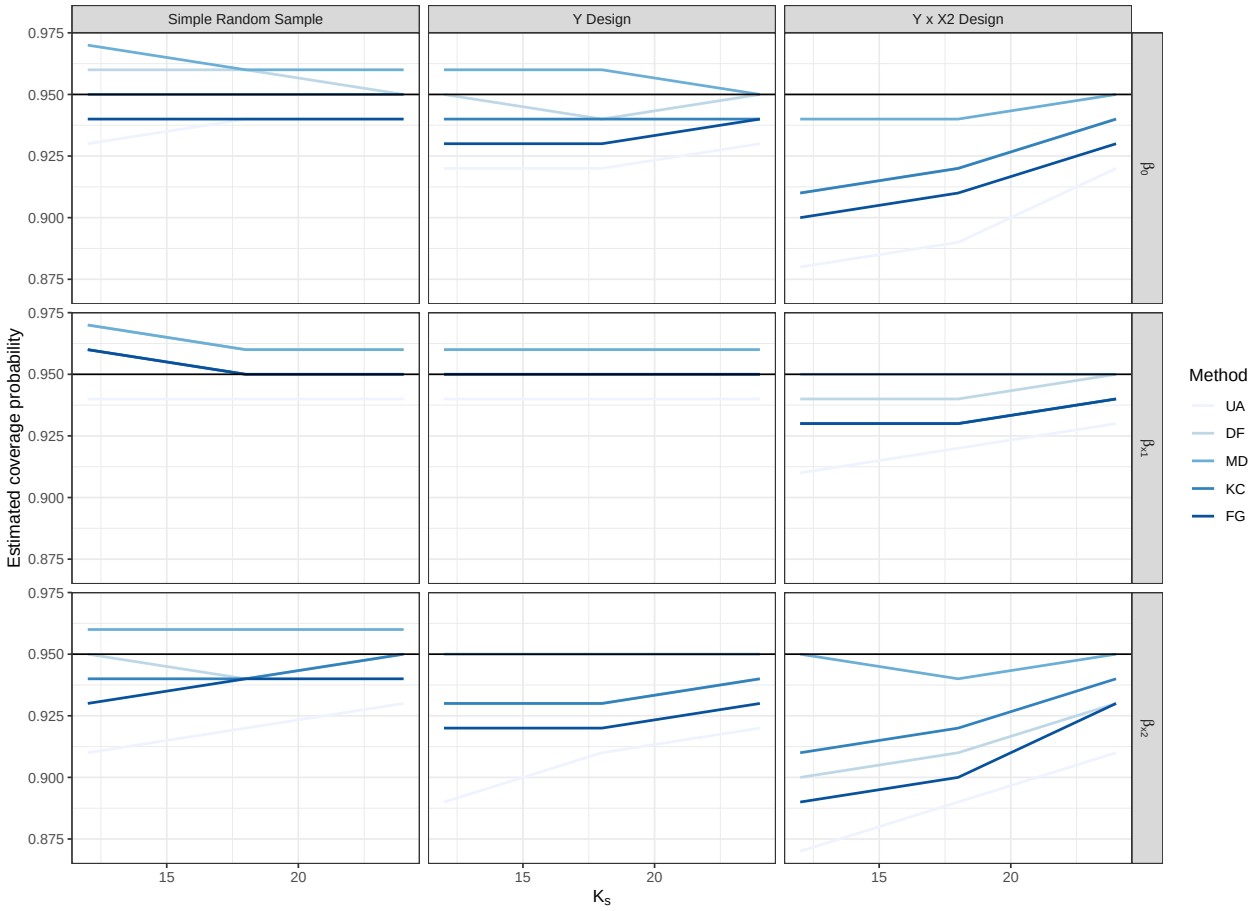

Figure A.14: Estimated coverage probabilities with $\widehat{\beta}_w$, using t distribution for confidence interval construction, σ_V=0.25

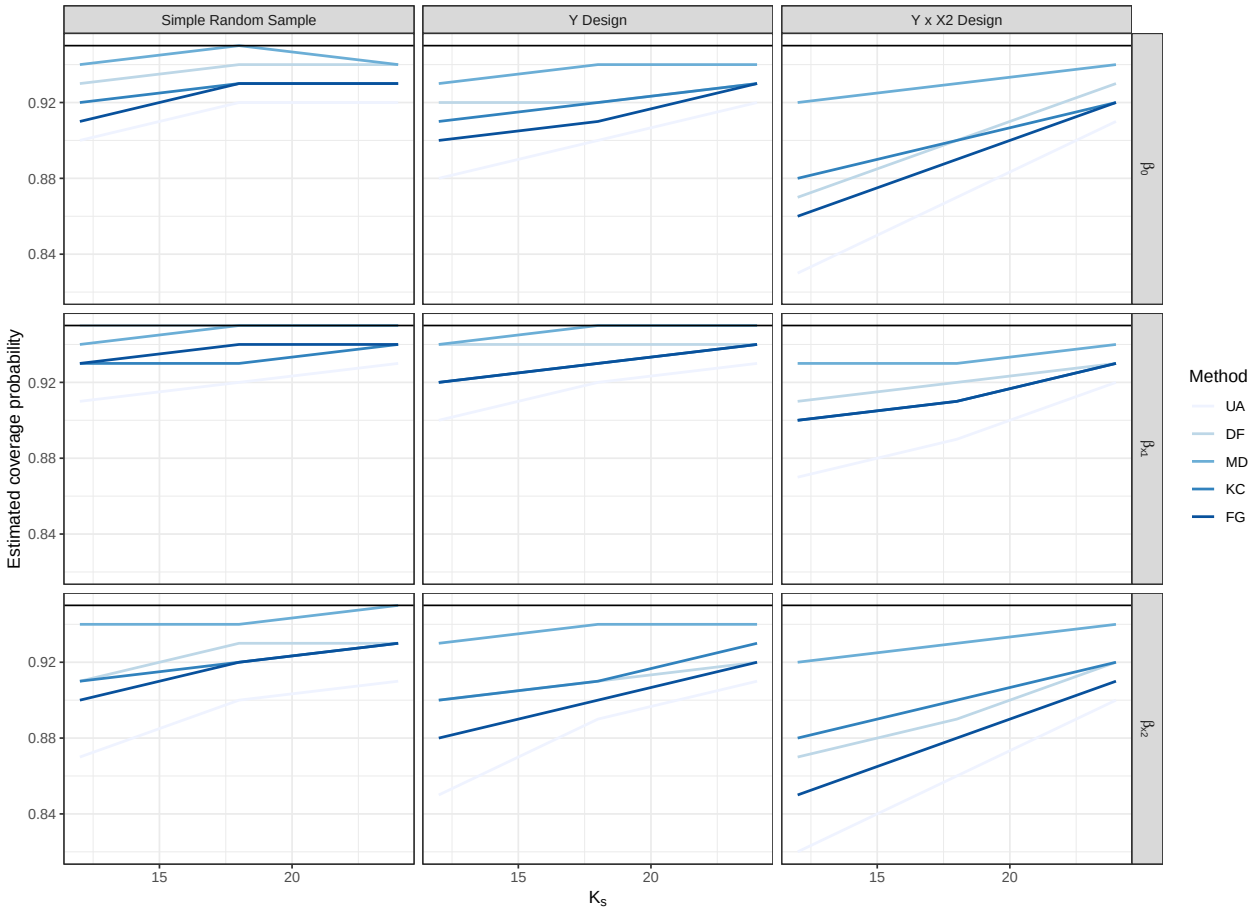

Figure A.15: Estimated coverage probabilities with $\widehat{\beta}_w^c$, using normal distribution for confidence interval construction, σ_V=0.25

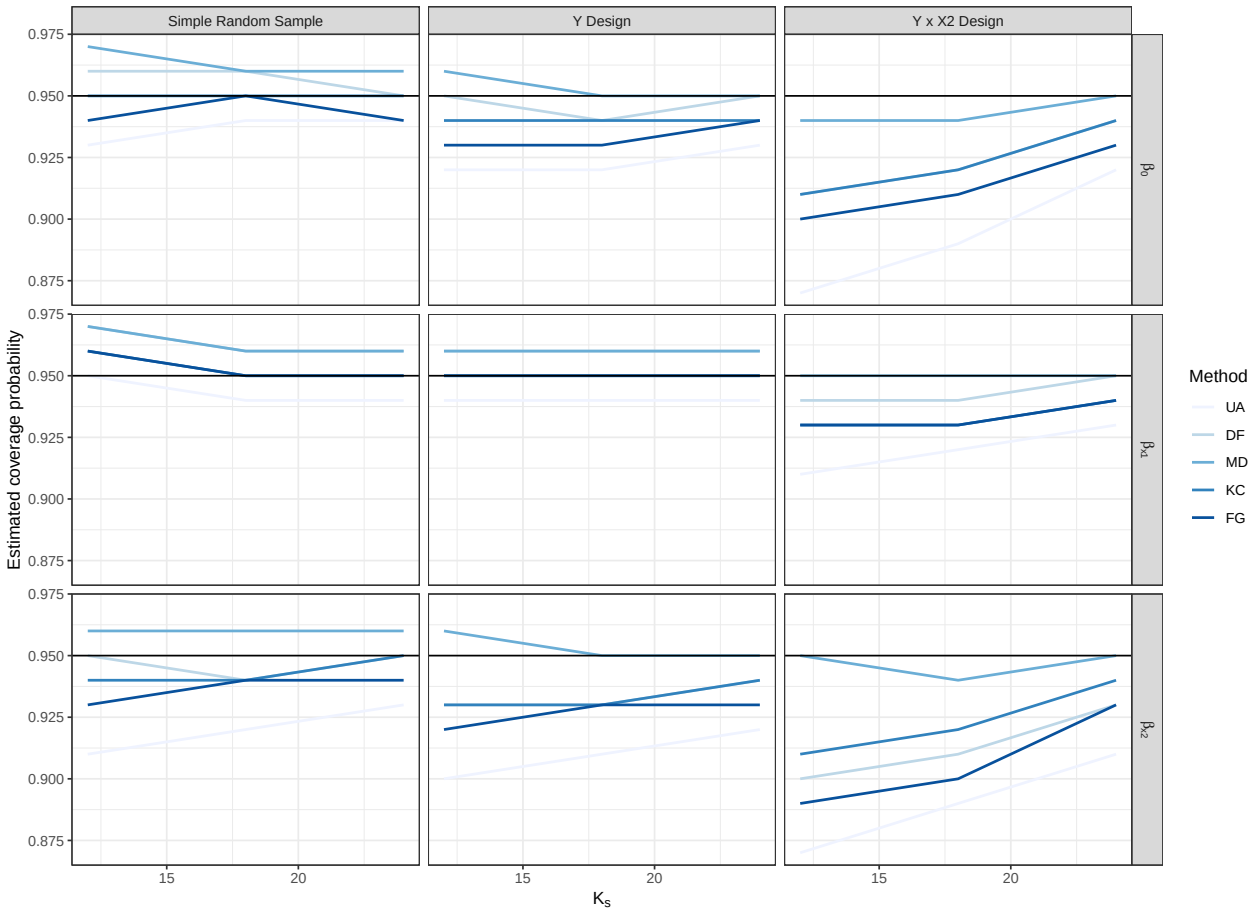

Figure A.16: Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using t distribution for confidence interval construction, σ_V=0.25

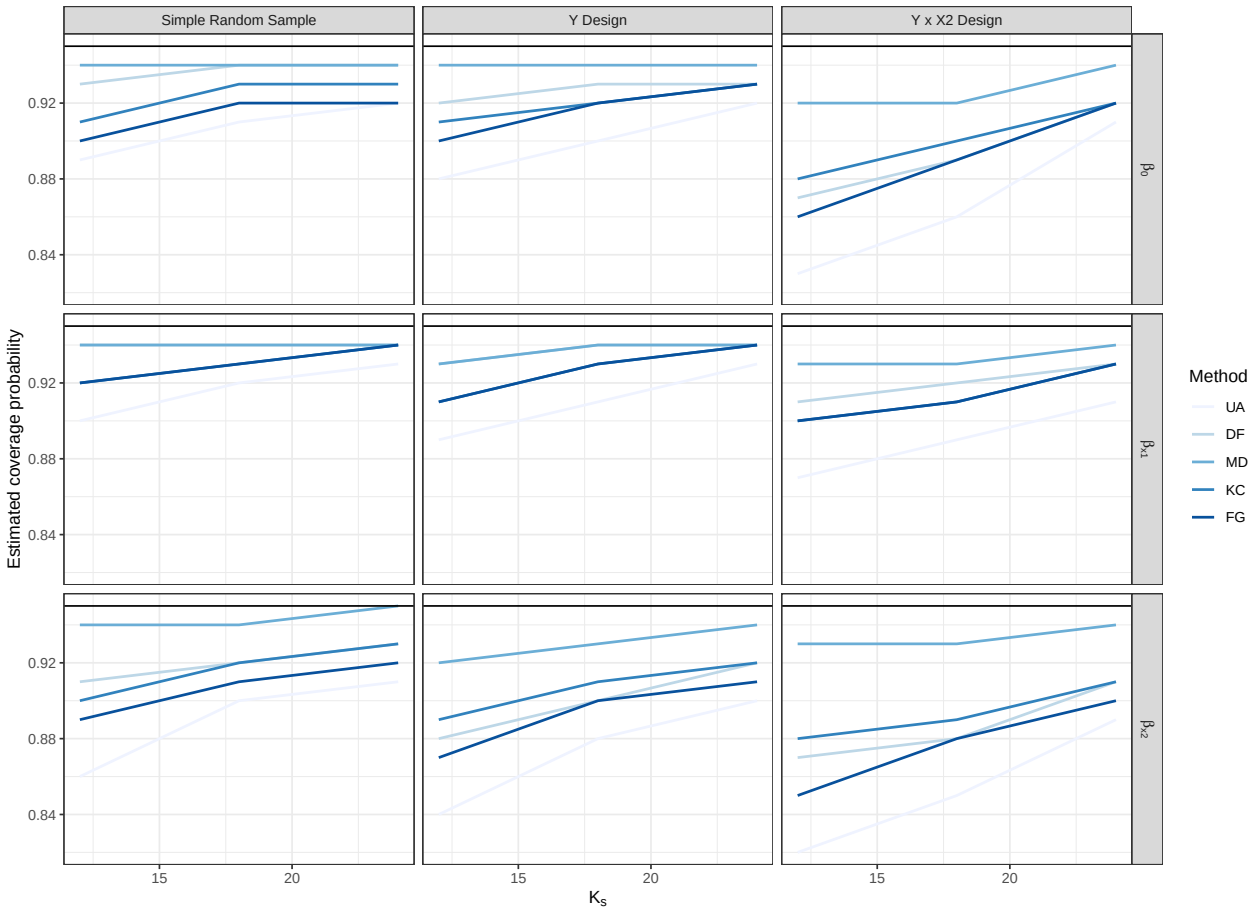

Figure A.17: Estimated coverage probabilities with $\widehat{\beta}_w$, using normal distribution for confidence interval construction, σ_V=0.25, ignoring negative correlation

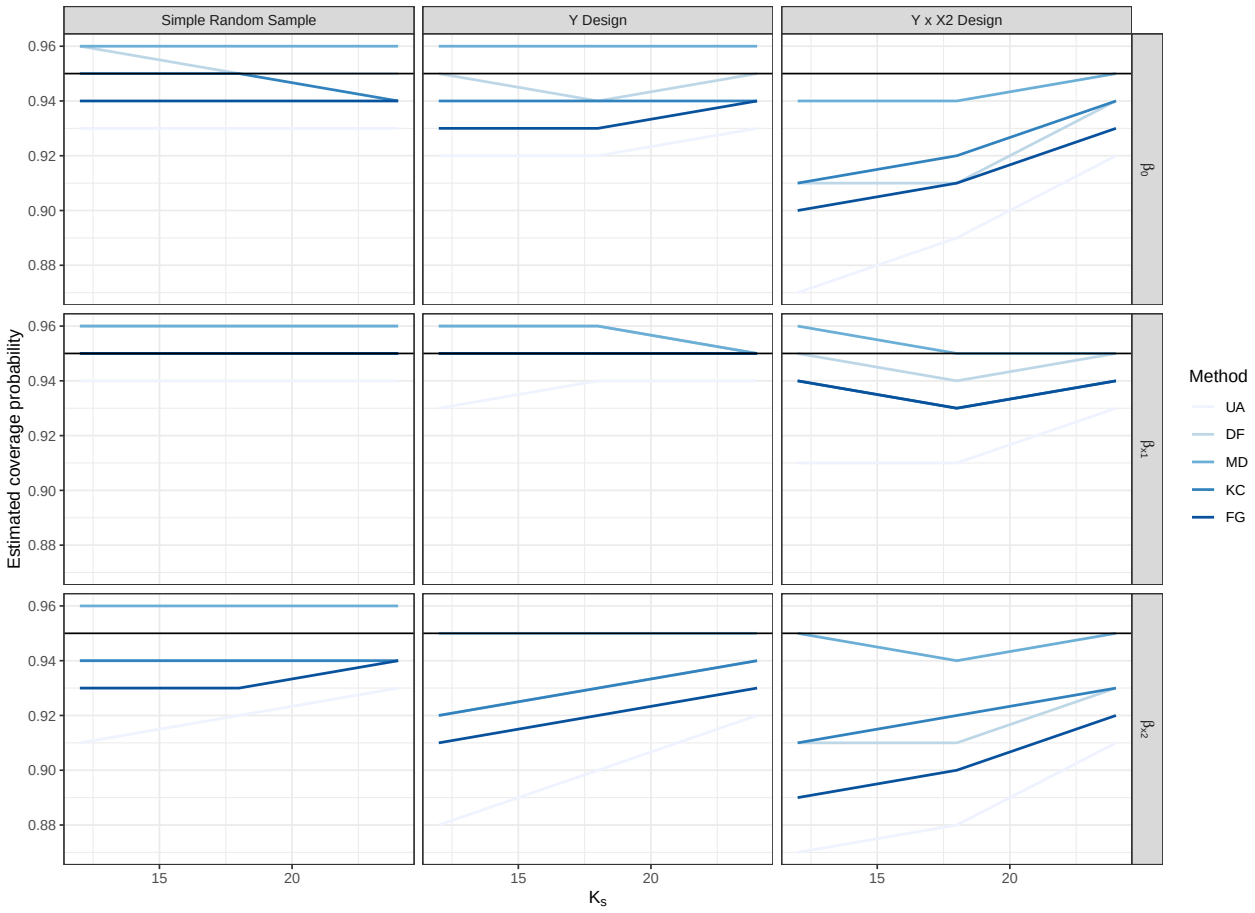

Figure A.18: Estimated coverage probabilities with $\widehat{\beta}_w$, using t distribution for confidence interval construction, σ_V=0.25, ignoring negative correlation

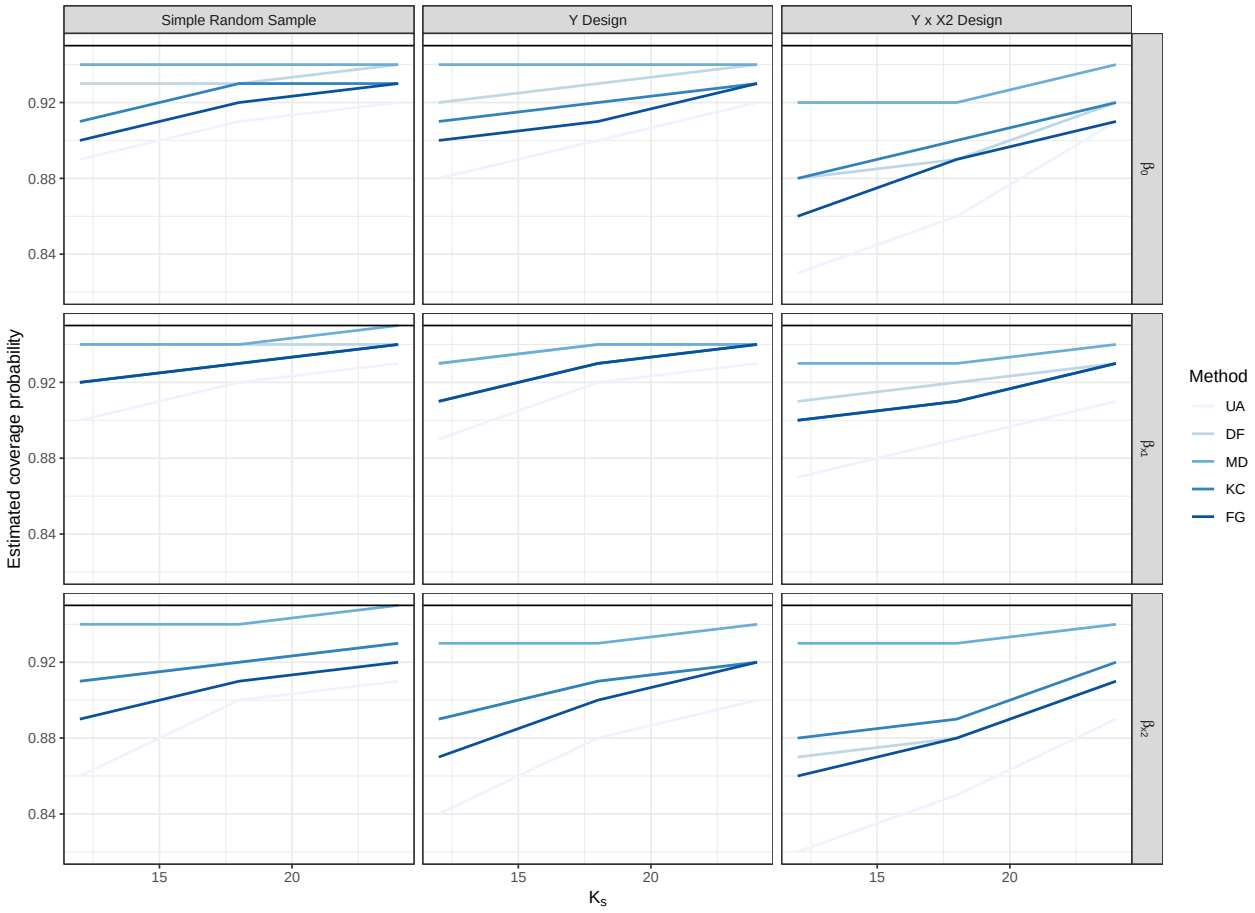

Figure A.19: Estimated coverage probabilities with $\widehat{\beta}_w^c$, using normal distribution for confidence interval construction, σ_V=0.25, ignoring negative correlation

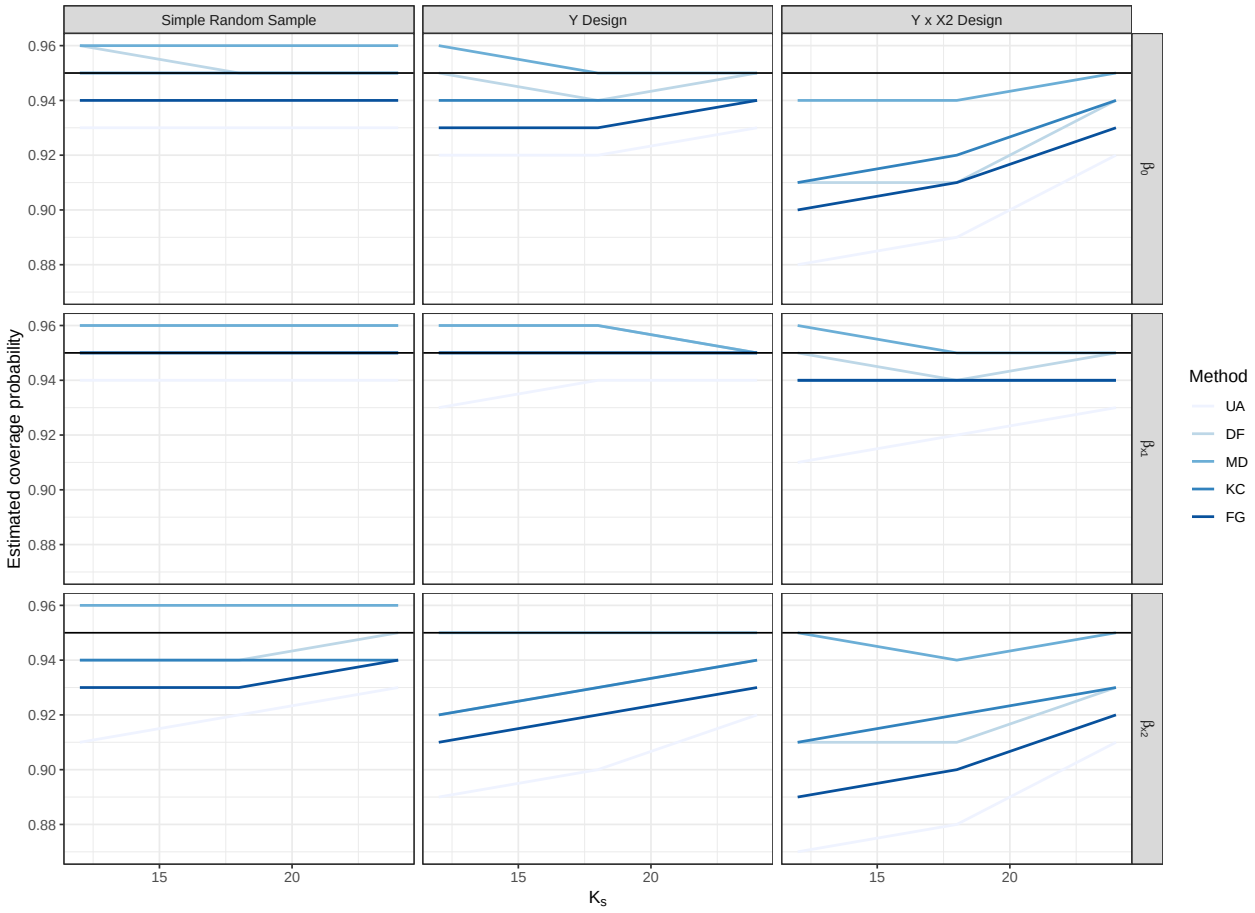

Figure A.20: Estimated coverage probabilities with $\widehat{\beta}_w^c$, using t distribution for confidence interval construction, σ_V=0.25, ignoring negative correlation

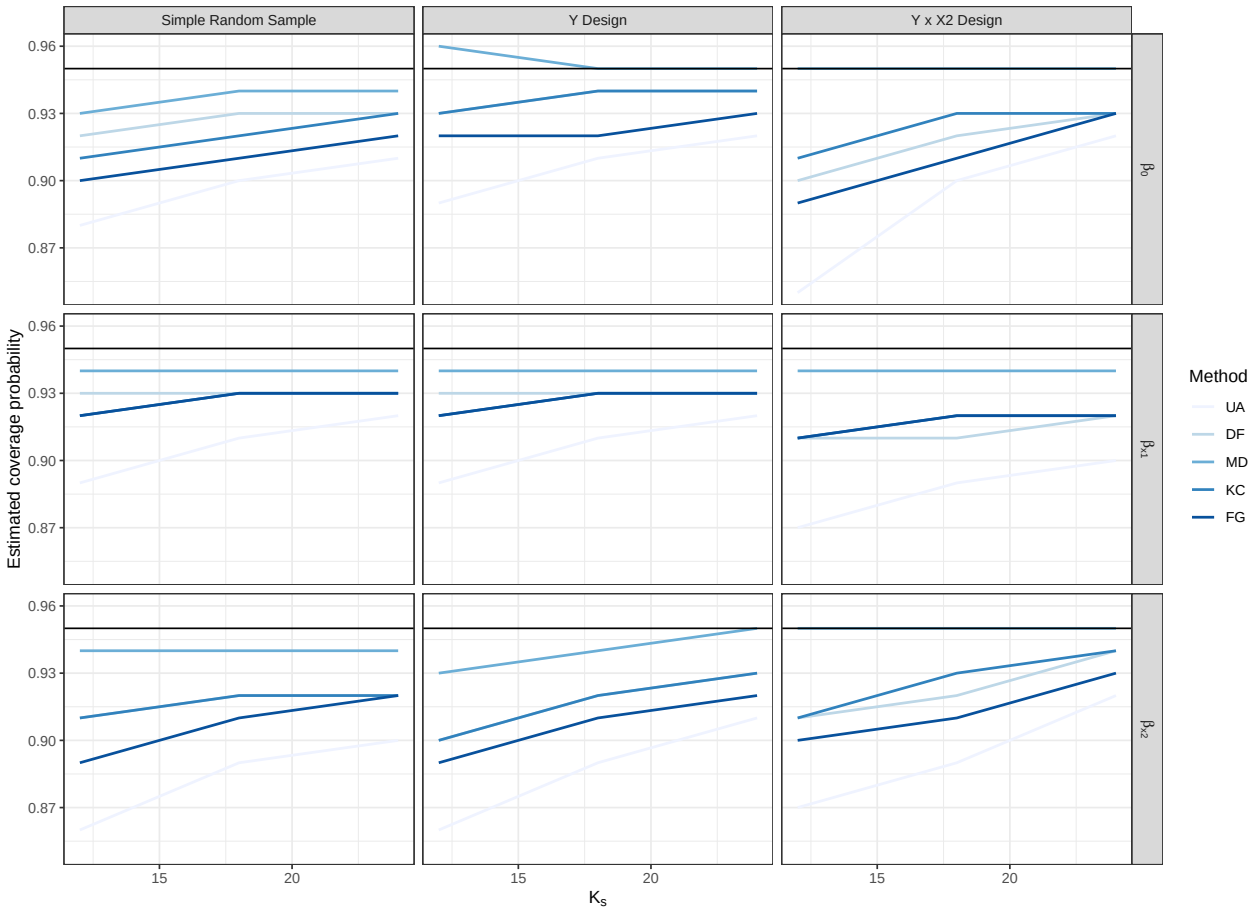

Figure A.21: Estimated coverage probabilities with $\widehat{\beta}_w$, using normal distribution for confidence interval construction, σ_V=0.75

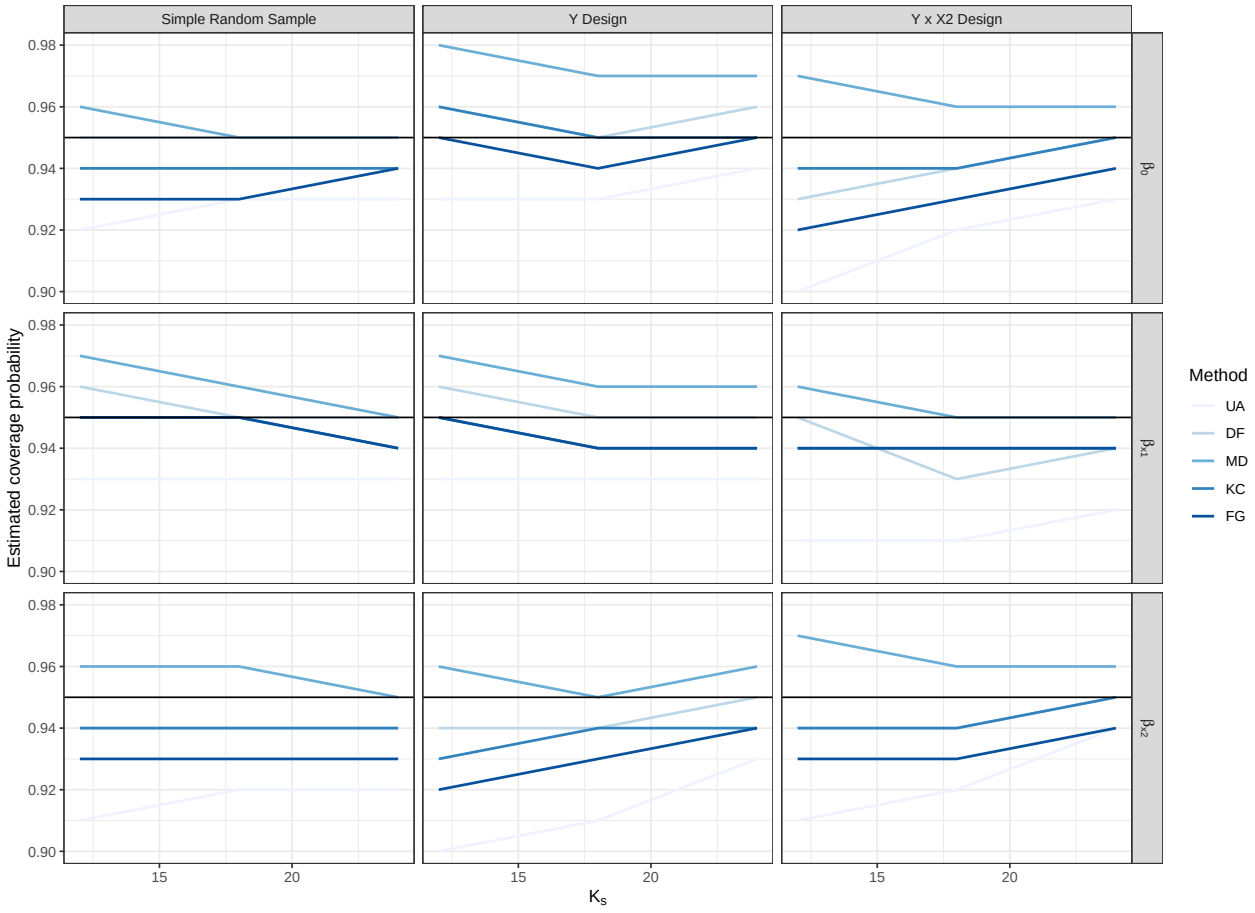

Figure A.22: Estimated coverage probabilities with $\widehat{\beta}_w$, using t distribution for confidence interval construction, σ_V=0.75

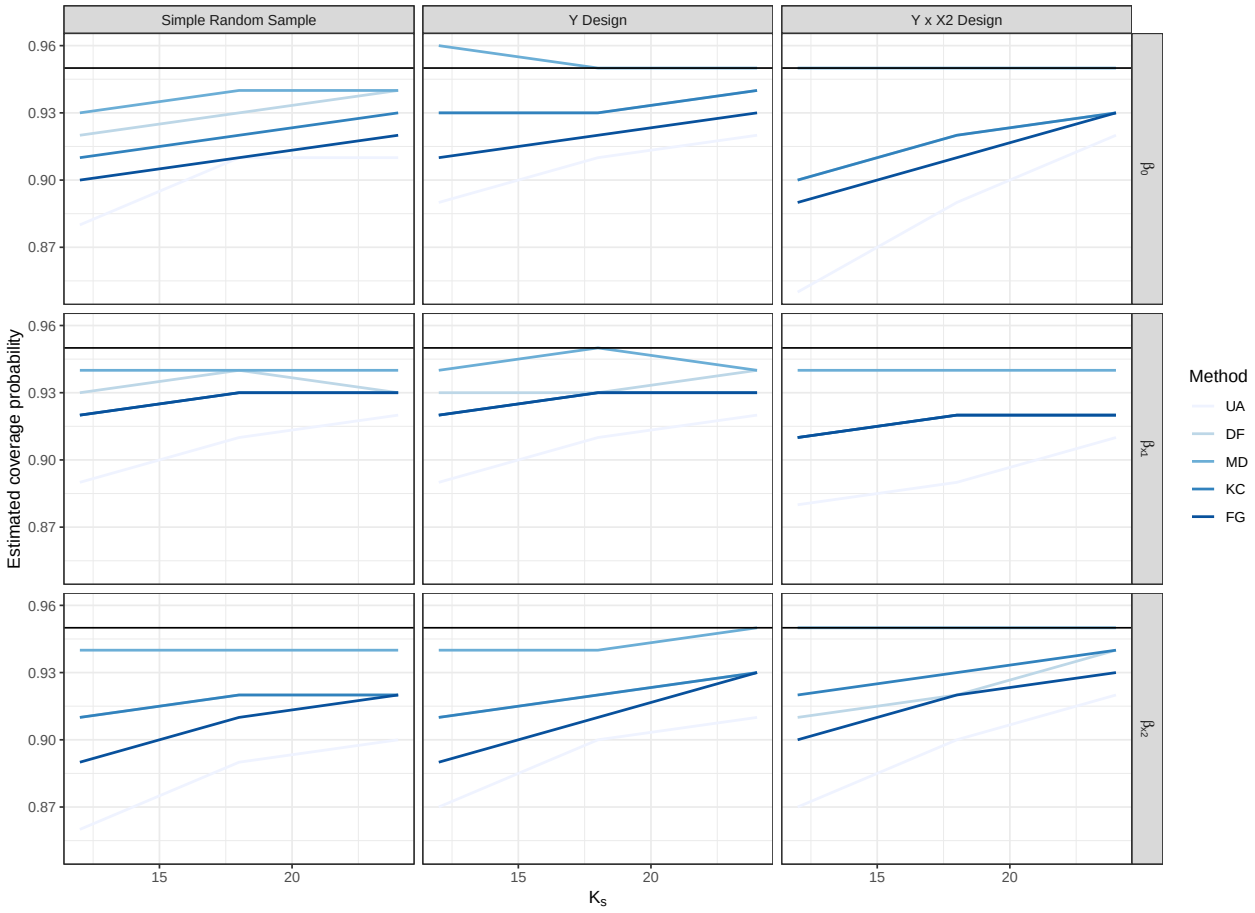

Figure A.23: Estimated coverage probabilities with $\widehat{\beta}_w^c$, using normal distribution for confidence interval construction, σ_Y=0.75

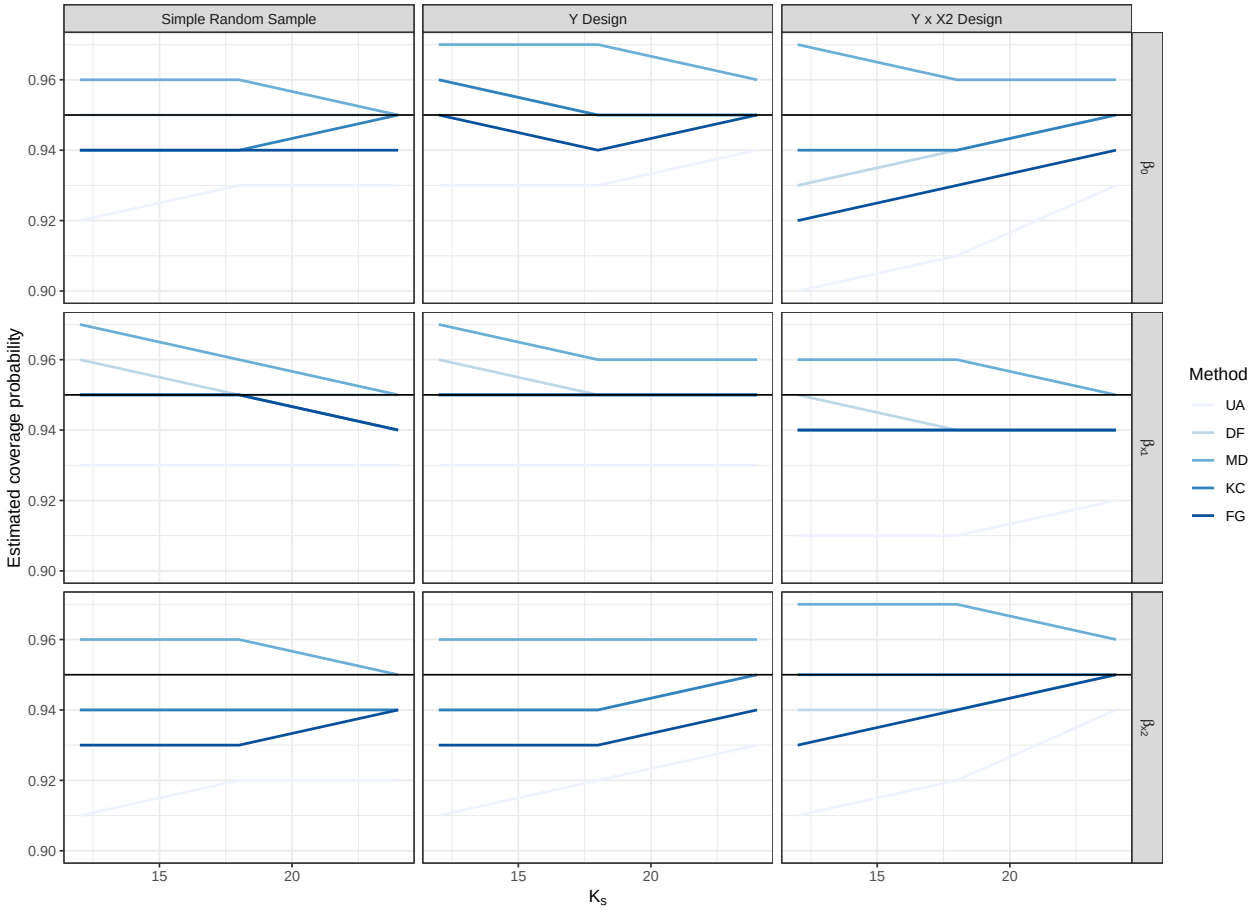

Figure A.24: Estimated coverage probabilities with $\widehat{\beta}_w^{\mathcal{L}}$, using t distribution for confidence interval construction, σ_V=0.75

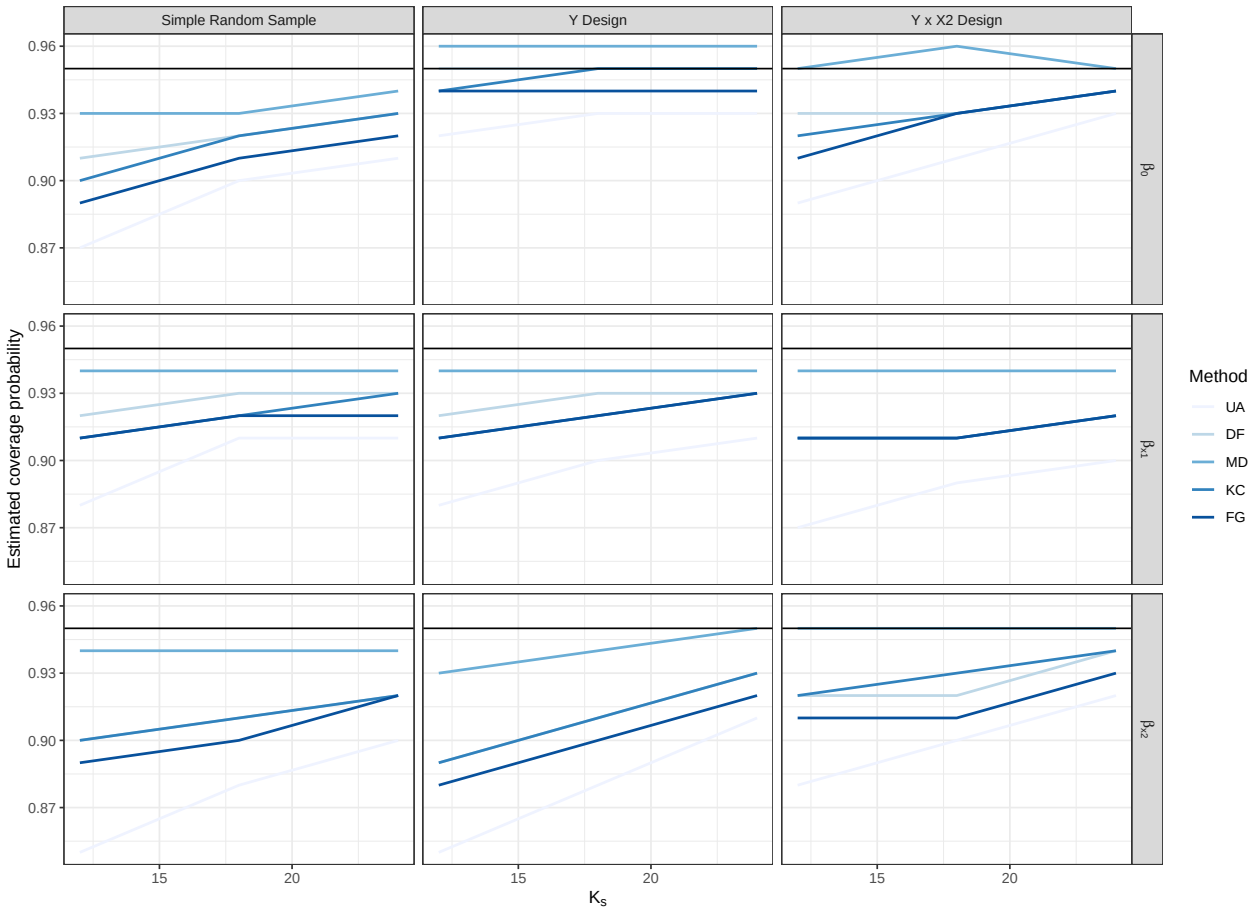

Figure A.25: Estimated coverage probabilities with $\widehat{\beta}_w$, using normal distribution for confidence interval construction, σ_V=0.75, ignoring negative correlation

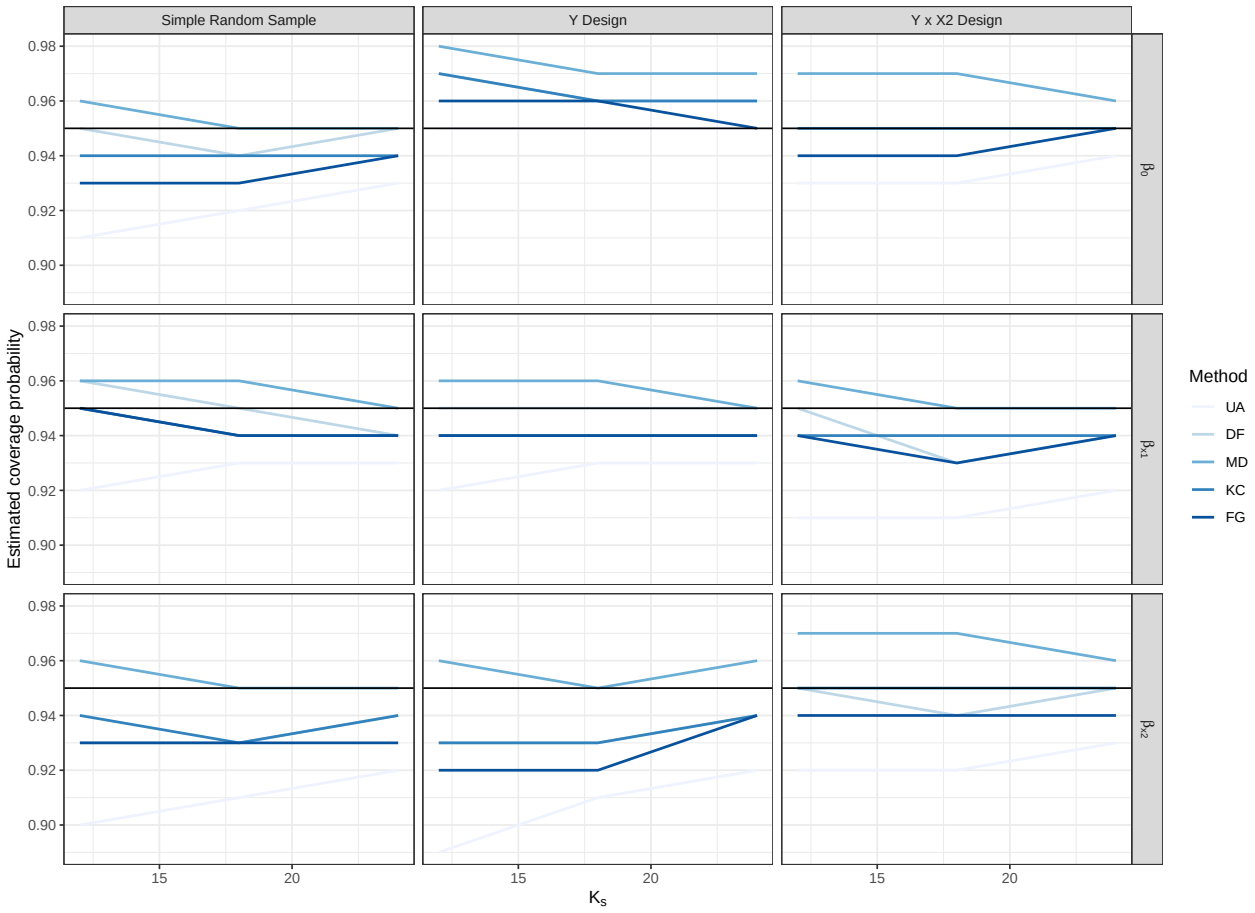

Figure A.26: Estimated coverage probabilities with $\widehat{\beta}_w$, using t distribution for confidence interval construction, σ_V=0.75, ignoring negative correlation

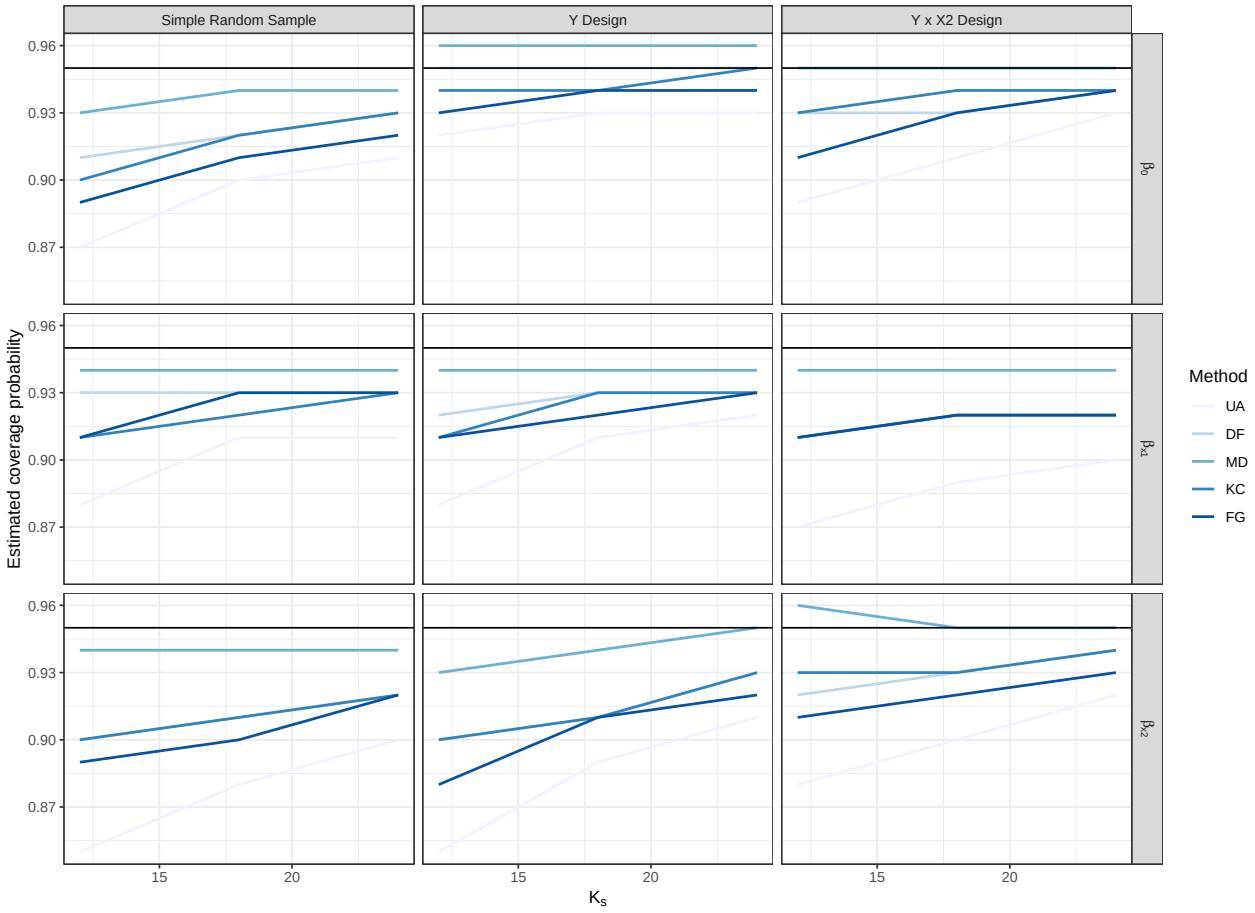

Figure A.27: Estimated coverage probabilities with $\widehat{\beta}_w^c$, using normal distribution for confidence interval construction, σ_Y=0.75, ignoring negative correlation

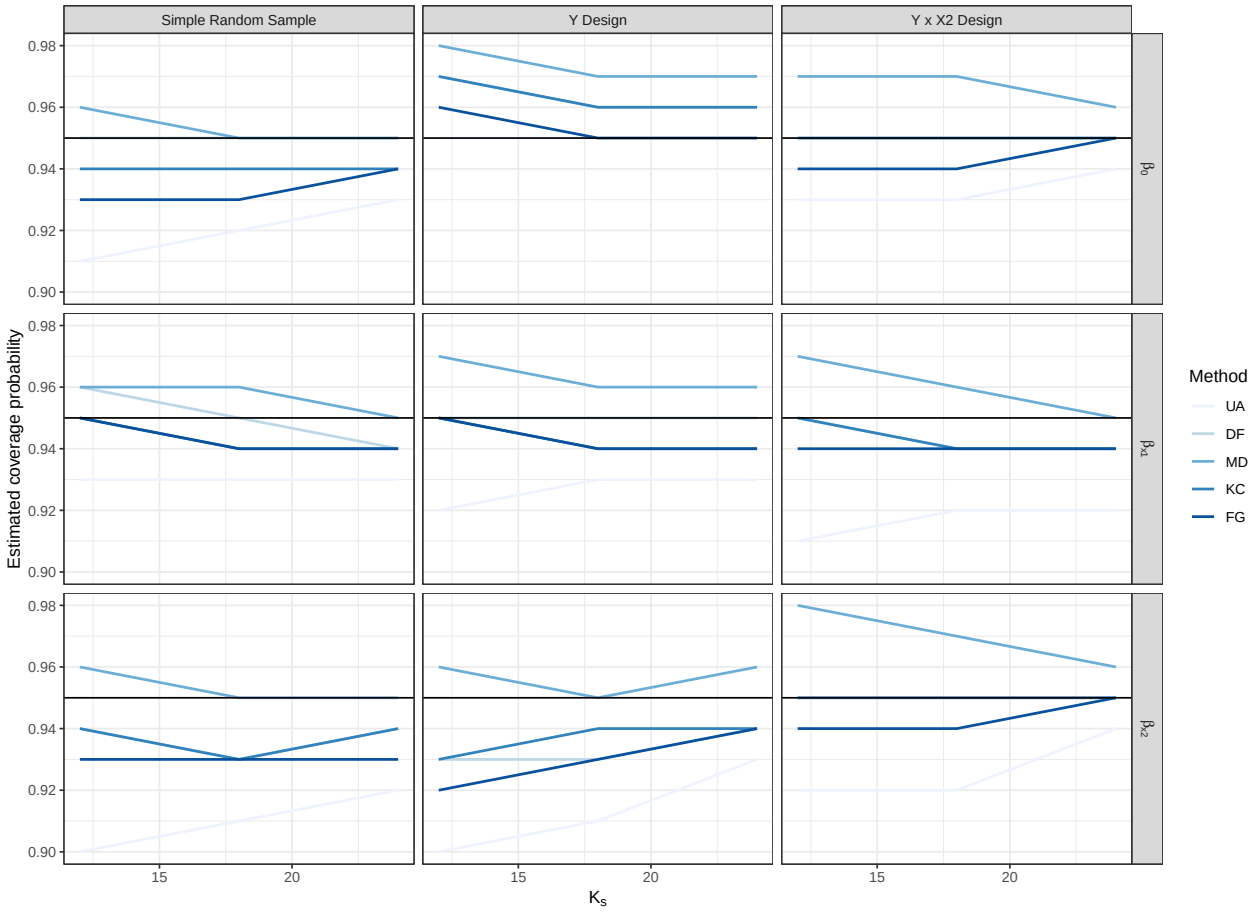

Figure A.28: Estimated coverage probabilities with $\widehat{\beta}_w^c$, using t distribution for confidence interval construction, σ_V=0.75, ignoring negative correlation

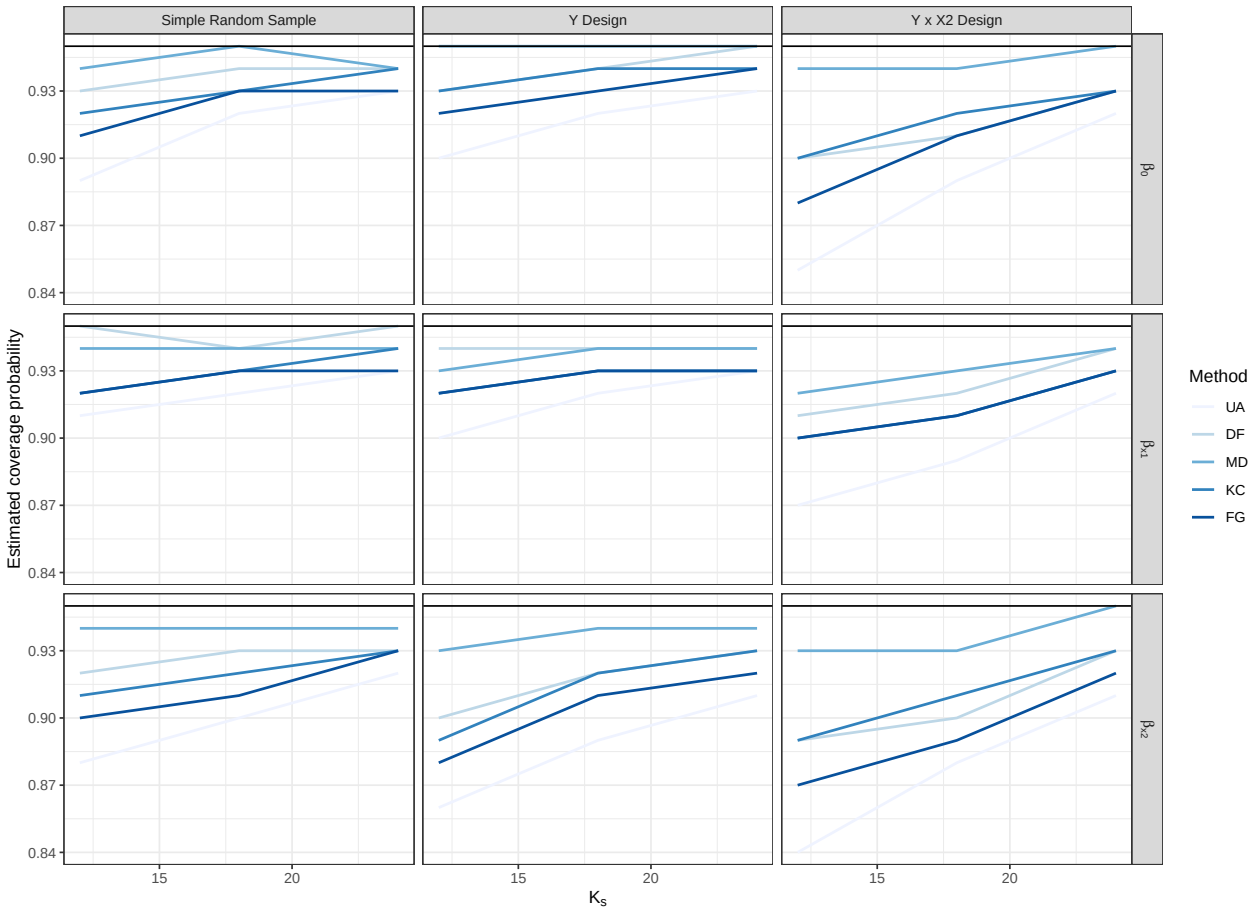

Figure A.29: Estimated coverage probabilities with $\widehat{\beta}_w$, using normal distribution for confidence interval construction, σ_V=0.5, working exchangeable

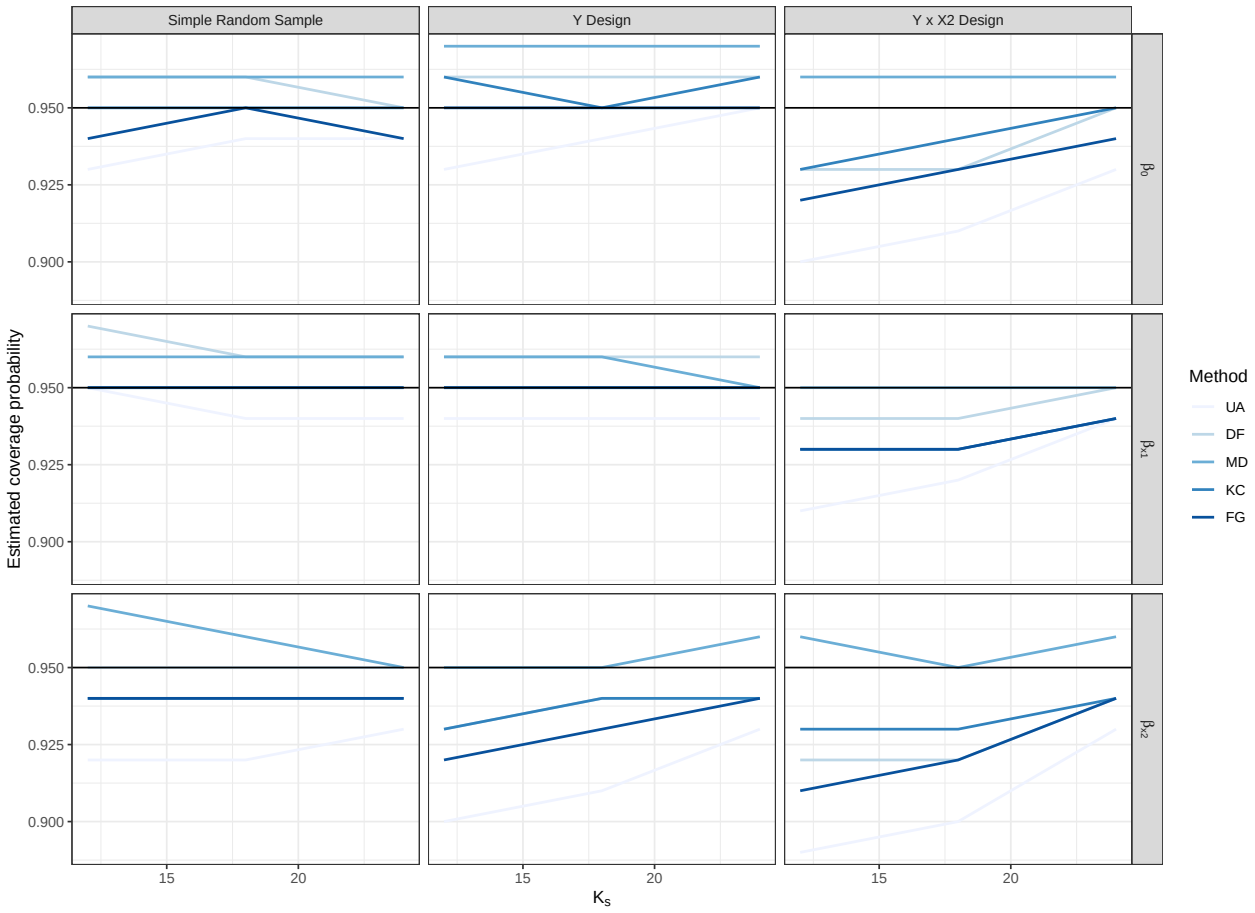

Figure A.30: Estimated coverage probabilities with $\widehat{\beta}_w$, using t distribution for confidence interval construction, σ_V=0.5, working exchangeable

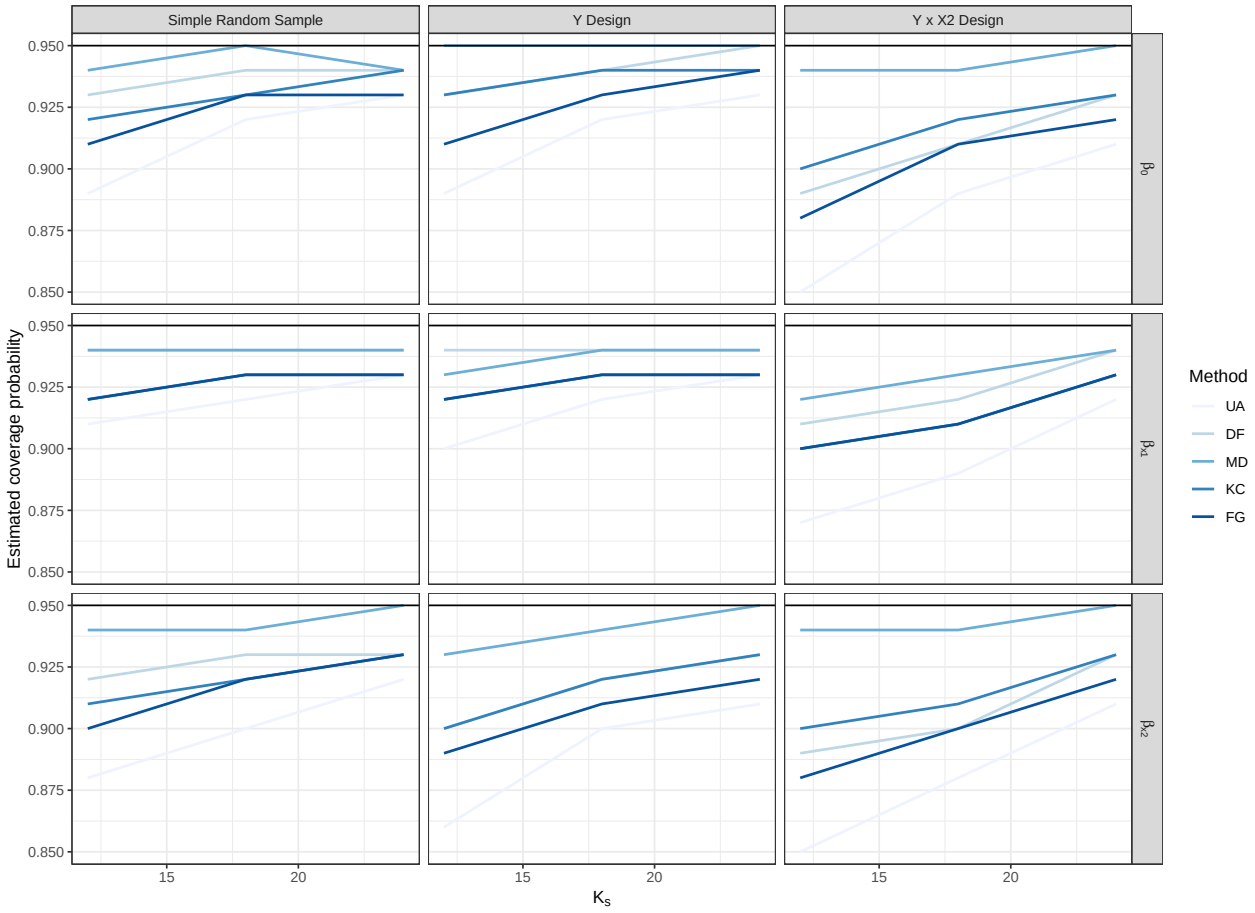

Figure A.31: Estimated coverage probabilities with $\widehat{\beta}_w^c$, using normal distribution for confidence interval construction, σ_Y=0.5, working exchangeable

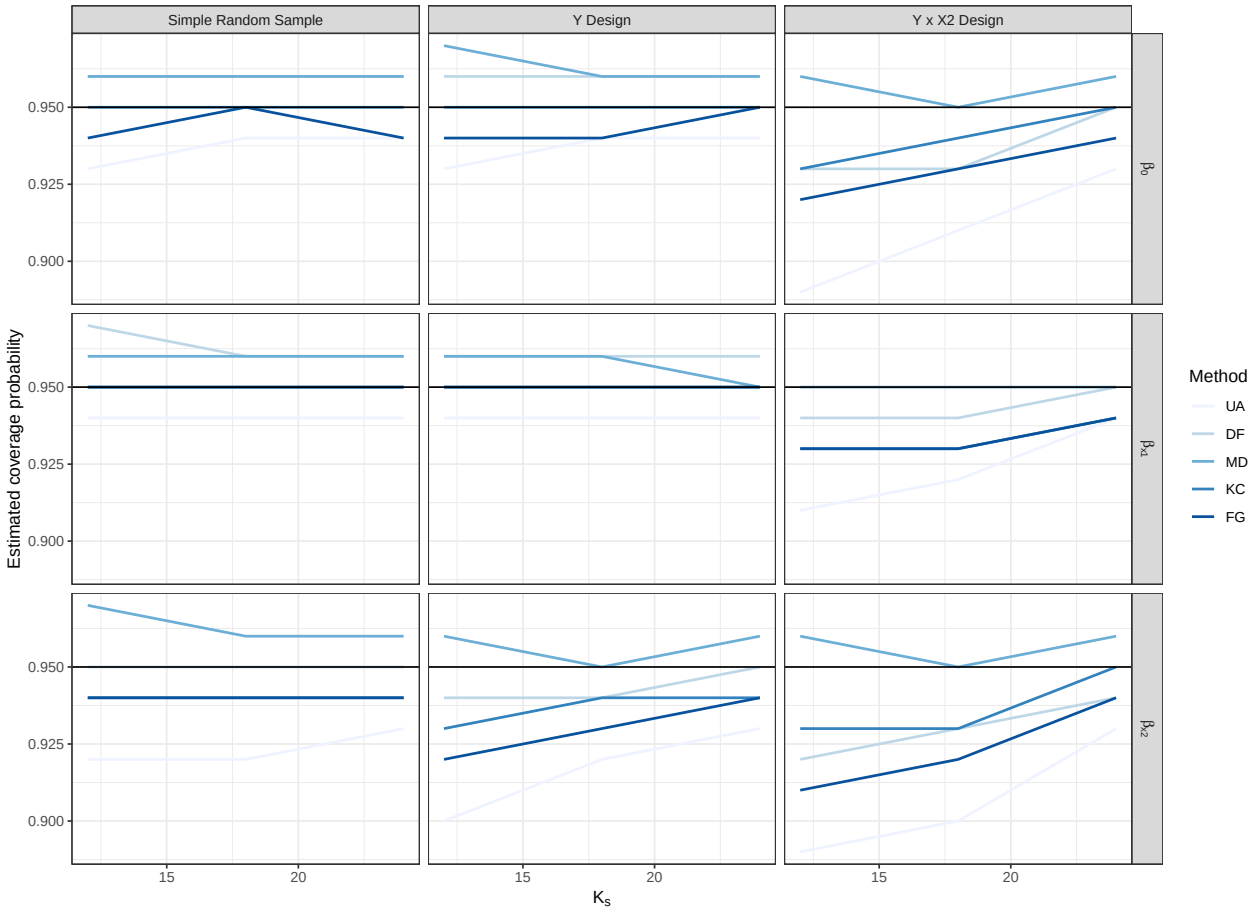

Figure A.32: Estimated coverage probabilities with $\widehat{\beta}_w^c$, using t distribution for confidence interval construction, σ_V=0.5, working exchangeable

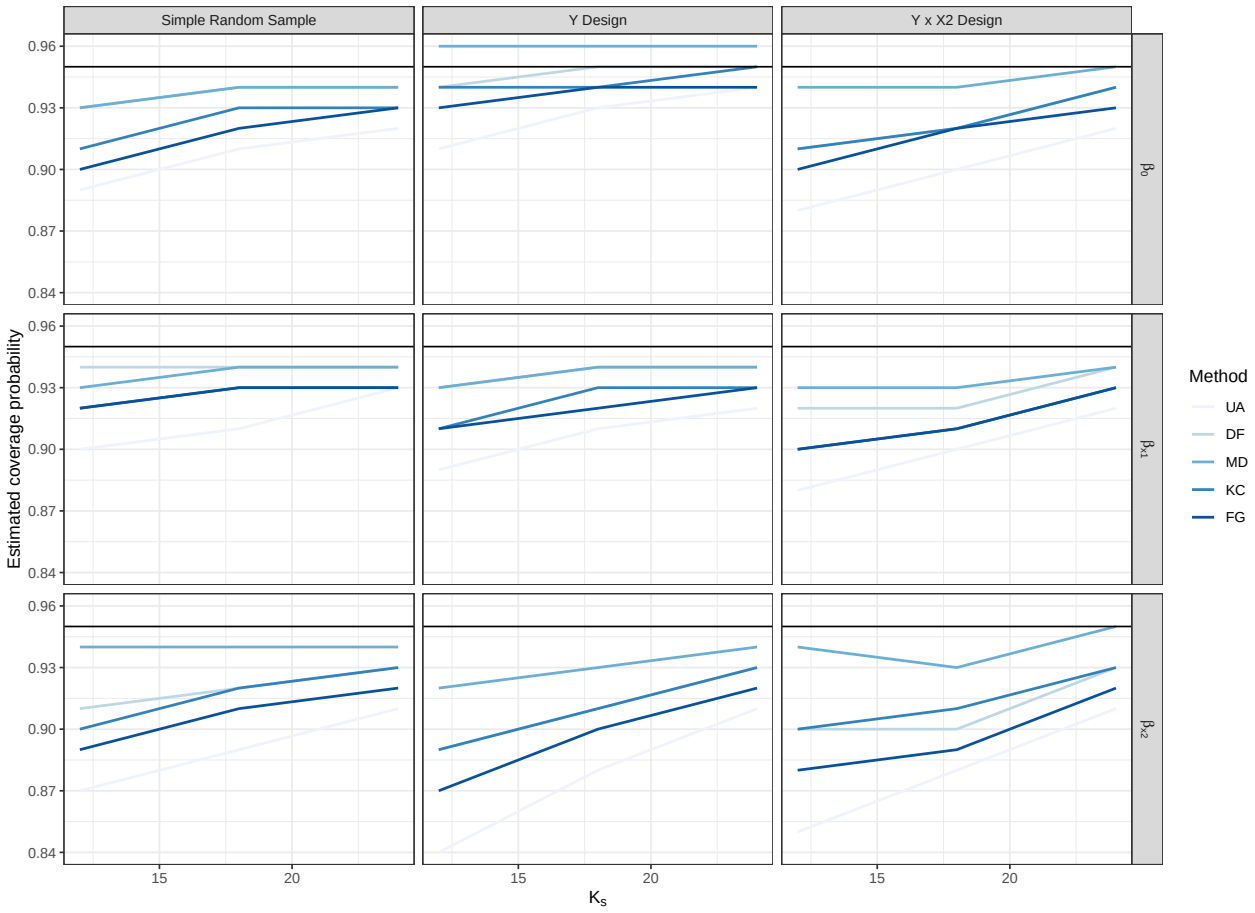

Figure A.33: Estimated coverage probabilities with $\widehat{\beta}_w$, using normal distribution for confidence interval construction, σ_V=0.5, working exchangeable, ignoring negative correlation

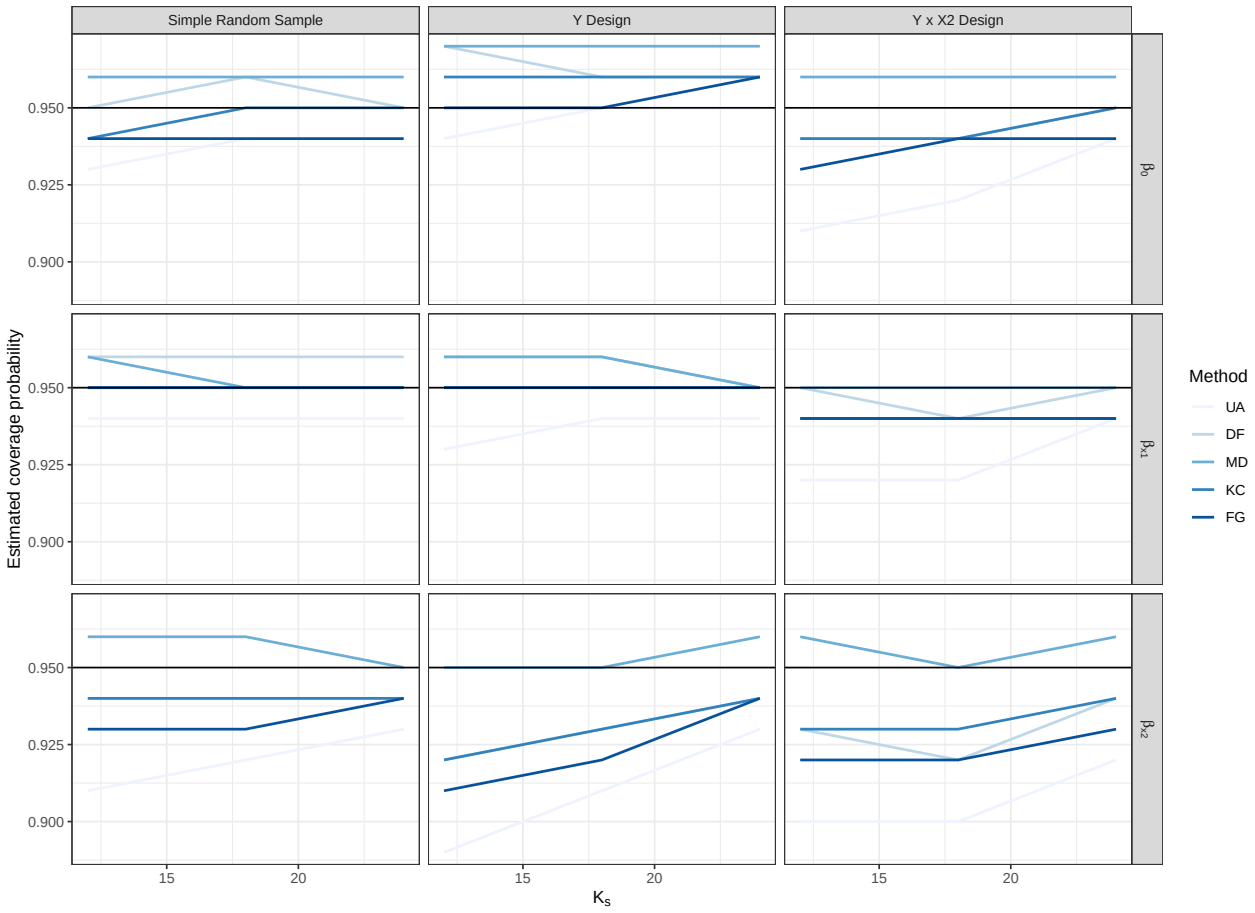

Figure A.34: Estimated coverage probabilities with $\widehat{\beta}_w$, using t distribution for confidence interval construction, σ_V=0.5, working exchangeable, ignoring negative correlation

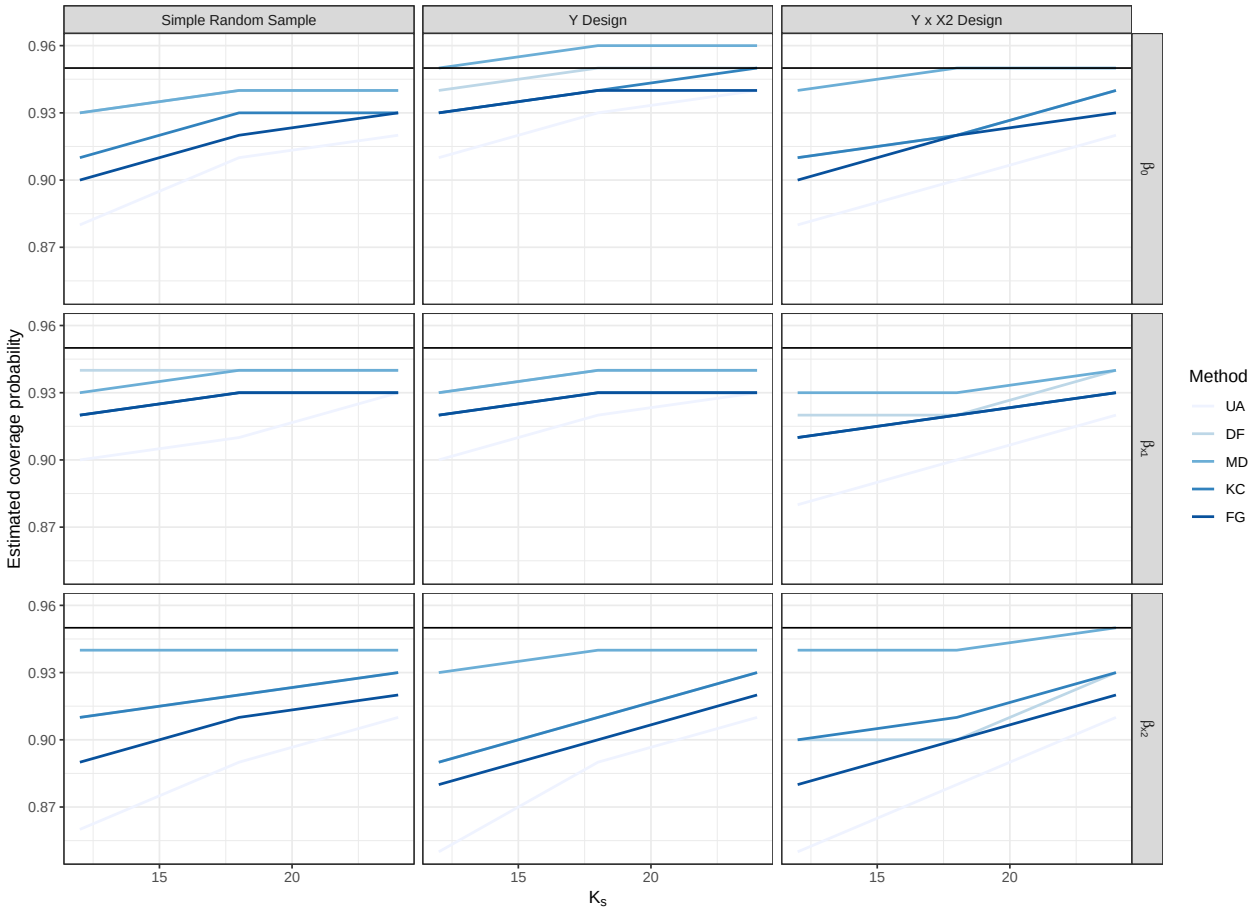

Figure A.35: Estimated coverage probabilities with $\widehat{\beta}_w^c$, using normal distribution for confidence interval construction, σ_V=0.5, working exchangeable, ignoring negative correlation

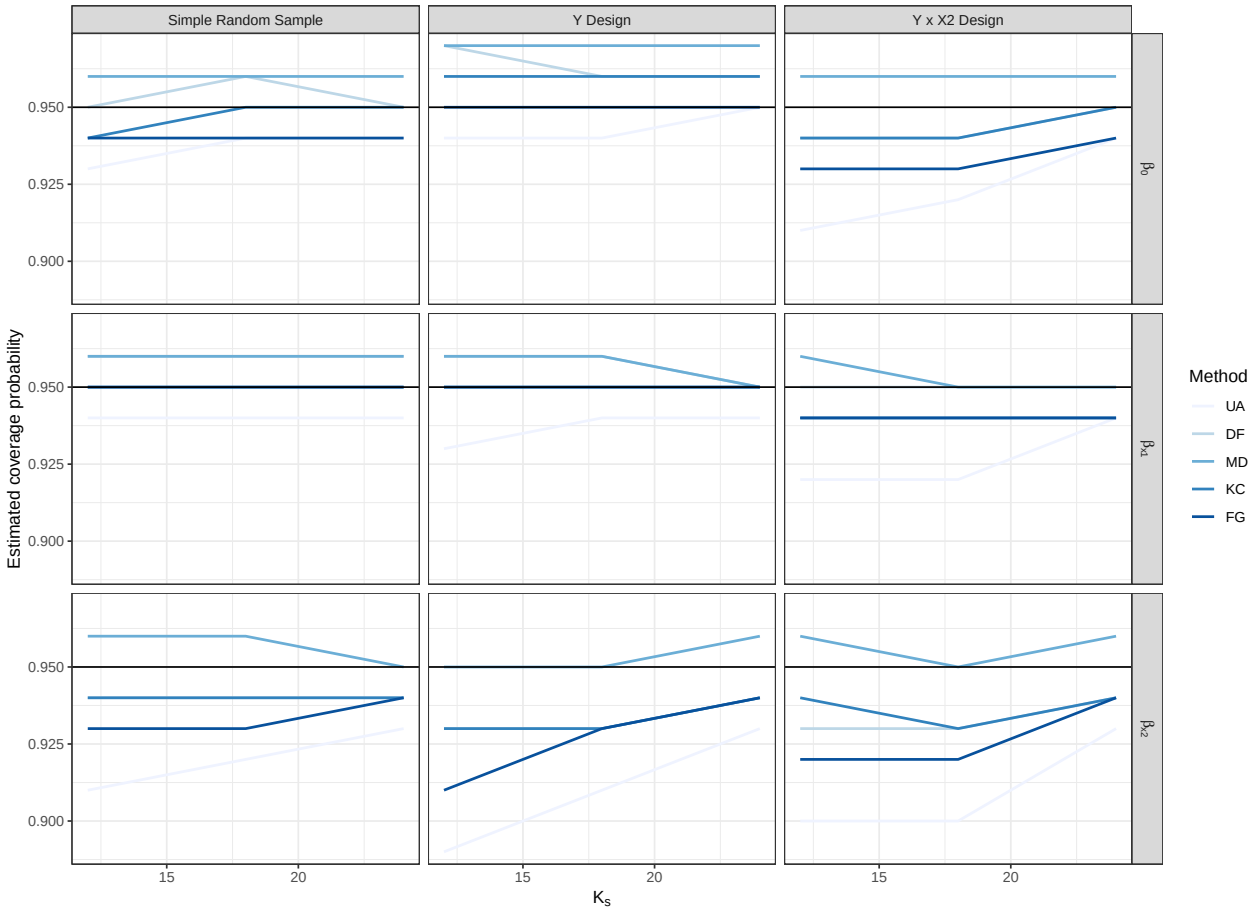

Figure A.36: Estimated coverage probabilities with $\widehat{\beta}^{c}_{w}$, using t distribution for confidence interval construction, σ_V=0.5, working exchangeable, ignoring negative correlation

B

Supplementary material to accompany Chapter 2

B.1 Initial Simulation to Investigate Potential for Efficiency Gains

As described in Sections 2.3 and 2.4 of Chapter 2, we conducted an initial simulation study using a data set of women enrolled in the Safer Deliveries Program in Zanzibar, Tanzania. For the women

in the data set, demographic and health information, obstetric history, the number of antenatal visits, and the location of delivery was collected by program-supported community health workers (CHWs). We consider a hypothetical study in which interest lies in determining what factors are associated with a woman delivering outside of a health facility. In particular, we assume that we are interested in the following marginal model for the i^{th} woman in the k^{th} shehia:

$$logit(P(Y_{ki=1})) = \beta_0 + \beta_1 X_{loc,k} + \beta_{ANC} X_{ANC,ki} + \boldsymbol{\beta}_A^T X_{A,ki} \tag{B.1}$$

Where Y_{ki} is the binary outcome denoting whether the woman delivered outside of a health facility (1/0=Yes/No), $X_{loc,k}$ is a binary cluster-level variable indicating which island the shehia of residence is located in (1/0 = Pemba/Unguja), and $X_{A,ki} = (X_{A1ki}, X_{A2ki}, X_{A3ki}, X_{A4ki}, X_{A5ki})^T$, where X_{A1} is an individual-level categorical variable giving the woman's previous location of delivery (0=NA, 1=At home/in the community, 2=On the way to a health facility, 3=Health facility), X_{A2} is an individual-level binary variable denoting whether a woman had 4 standard ANC visits (1/0=Yes/No), X_{A3} is an individual-level binary variable indicating whether a woman has cardiac disease (1/0=Yes/No), X_{A4} is an individual-level categorical variable denoting the facility type that was recommended to the woman for her delivery (1=Cottage hospital, 2=PHCU, 3=Referral hospital), and X_{A5} is an individual-level binary variable indicating whether the woman received a visit from a community health worker at 8-9 months of pregnancy (1/0=Yes/No).

In model (B.1) above, we suppose that the parameter of primary interest is β_{ANC}. While the original data set contains information on 42056 women, we restrict the data to include only women who had already given birth, with complete data on all covariates in model (B.1), which results in a data set with 28789 women. We defined 500 'designs', where each design is defined as a set of 40 shehia ids that was obtained via simple random sampling. We generated 1000 complete datasets, by simulating correlated outcomes based on model (B.1), using the `GenBinaryY()` function in the the `MMLB`

packag for R. For each generated complete dataset, we took 500 samples corresponding to the 500 designs, and for each of the 500 samples, we obtained the point estimates using generalized estimating equations (GEE). Across the 1000 iterations, we then computed the standard deviation of the 1000 point estimates for each for the 500 designs. Figure B.1 shows this for β_{ANC}:

Figure B.1 shows that there is efficiency to be gained by making a good choice regarding which clusters to sample, i.e. by choosing a good sampling design. To investigate whether there is a general rule of thumb that can be followed to determine such a design, we looked at the characteristics of the 'best' designs for the estimation of β_{ANC}: we restricted attention to the designs which yielded a mean point estimate of β_{ANC} within 5% of the gold standard, where the gold standard was taken to be the mean point estimate obtained from running the analysis on the 1000 generated complete data sets (N=28789). Among this set of designs, we then defined the 'best' designs to be the 25 designs with the lowest standard deviation in the point estimates of β_{ANC}, and the 'worst' designs to be the 25 designs with the highest standard deviation in the point estimates.

Table 1/Scenario 1 displays summary information on the best and worst designs for the estimation of β_{ANC}. We see that the best design has a substantially larger overall sample size ($n = 4606$ vs. $n = 3160$), a larger number of exposure cases (1284 vs. 672), and a larger mean number of outcome cases (1337 vs. 773).

The top panel in Figure B.2 shows all 500 designs plotted as a function of the number of X_{ANC} cases, and the mean number of outcome cases across the 1000 iterations. The dark blue points indicate the best designs among the set of 'unbiased' designs, whereas the red points indicate the worst designs in this set. We see that in comparison to the worst designs, the best designs have a higher number of X_{ANC} cases and, on average, a higher number of outcome cases. The bottom panel in Figure B.2 additionally shows that the best designs yield a larger overall sample size than the worst designs. These results are intuitive, and suggest that a general rule of thumb for determining an efficient design might involve selecting clusters with 1) a higher number of outcome cases, 2) a higher

number of X_{ANC} cases, and 3) higher overall individual sample size n.

However, it may not necessarily be the case that a cluster simultaneously satisfies the three criteria listed above; in fact, it may be the case, for example, that the clusters with a higher number of outcome cases have a lower number of exposure cases. In the original data set-up we described, the number of outcome cases and the number of X_{ANC} cases is proportional to the shehia size. We therefore also consider a scenario in which the number of outcome cases in a cluster is *inversely* proportional to cluster size. This was done by introducing into the data generation model an indicator for cluster size, where the indicator is equal to 1 if the cluser size is greater than the median cluster size, and 0 otherwise, with a coefficient of -3.

Table 1/Scenario 2 gives the characteristics of the best and worst 'unbiased' designs for the estimation of β_{ANC}, in the setting where the number of outcome cases in a cluster is inversely proportional to the cluster size. We see that in this scenario, the best (lowest standard deviation of $\widehat{\beta}_{ANC}$) design has a smaller overall sample size than the design with the highest standard deviation of $\widehat{\beta}_{ANC}$ ($n = 3846$ vs. $n = 4621$), as well as a lower number of exposure cases, but a higher average number of outcome cases.

Figure B.3 shows the overall results for this second scenario: we see that the best designs, in comparison to the worst designs, generally (though not always) yield samples with a higher average number of outcome cases, while there is no clear difference between the number of X_{ANC} cases or the overall sample size n: i.e. some of the designs that have a higher number of X_{ANC} cases or that yield a larger overall sample size are less efficient in the estimation of β_{ANC} than designs that have a lower number of X_{ANC} cases or that yield a smaller overall sample size. This suggests that there is no general rule of thumb that can be reliably used to determine an efficient design, as an efficient design depends on the interplay of a number of factors. Therefore, while there is demonstrated potential for efficiency gain through the wise selection of clusters, these results suggest the need for the development of a framework for selecting an efficient design, which is the objective of this paper.

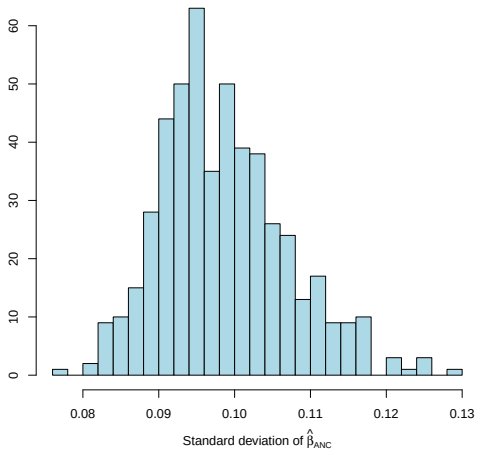

Figure B.1: Distribution of the standard errors of β_{ANC} under the 500 simple random sampling designs across 1000 iterations.

Table B.1: Characteristics of the best (lowest standard deviation of $\widehat{\beta}_{ANC}$) and worst (highest standard deviation of $\widehat{\beta}_{ANC}$) designs among the set of 'unbiased' designs for the estimation of β_{ANC}. In Scenario 1, the number of outcome cases is proportional to the cluster size, while in Scenario 2, the number of outcome cases is inversely proportional to the cluster size.

	Design	n	X_{ANC}	Y
Scenario 1				
Lowest sd($\widehat{\beta}_{ANC}$)	#431	4606	1284	1337
Highest sd($\widehat{\beta}_{ANC}$)	#284	3160	672	773
Scenario 2				
Lowest sd($\widehat{\beta}_{ANC}$)	#156	3846	1130	368
Highest sd($\widehat{\beta}_{ANC}$)	# 316	4621	1234	211

B.2 Complete Simulation Study Results

This section presents the complete simulation results for the simulation studies described in Section 5 of Chapter 2.

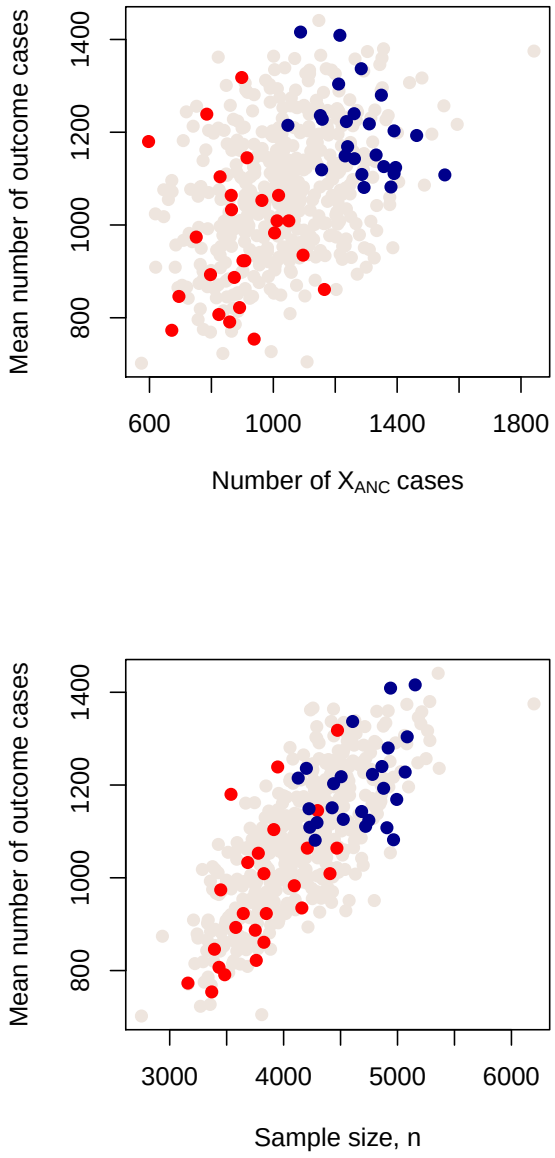

Figure B.2: The top panel shows the 500 designs plotted as a function of the number of X_{ANC} cases and the mean number of outcome cases across the samples. The dark blue points represent the best (lowest standard deviation of $\widehat{\beta}_{ANC}$) designs, and the red points represent the worst (highest standard deviation of $\widehat{\beta}_{ANC}$) designs. In this setting, the number of outcome cases is proportional to the cluster size.

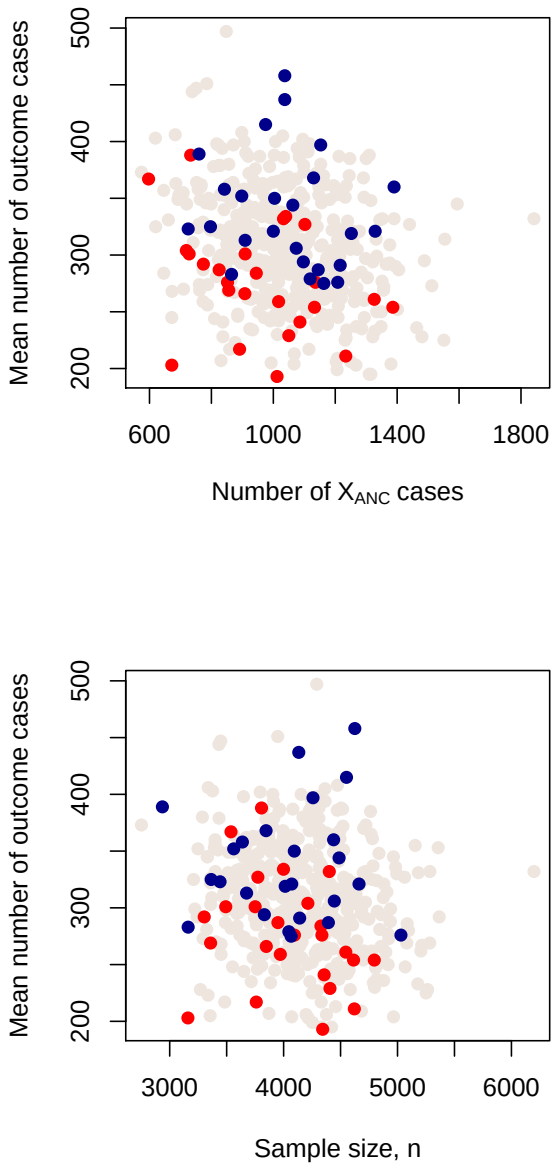

Figure B.3: The top panel shows the 500 designs plotted as a function of the number of X_{ANC} cases and the mean number of outcome cases across the samples. The dark blue points represent the best (lowest standard deviation of $\widehat{\beta}_{ANC}$) designs, and the red points represent the worst (highest standard deviation of $\widehat{\beta}_{ANC}$) designs. In this setting, the number of outcome cases is inversely proportional to the cluster size.

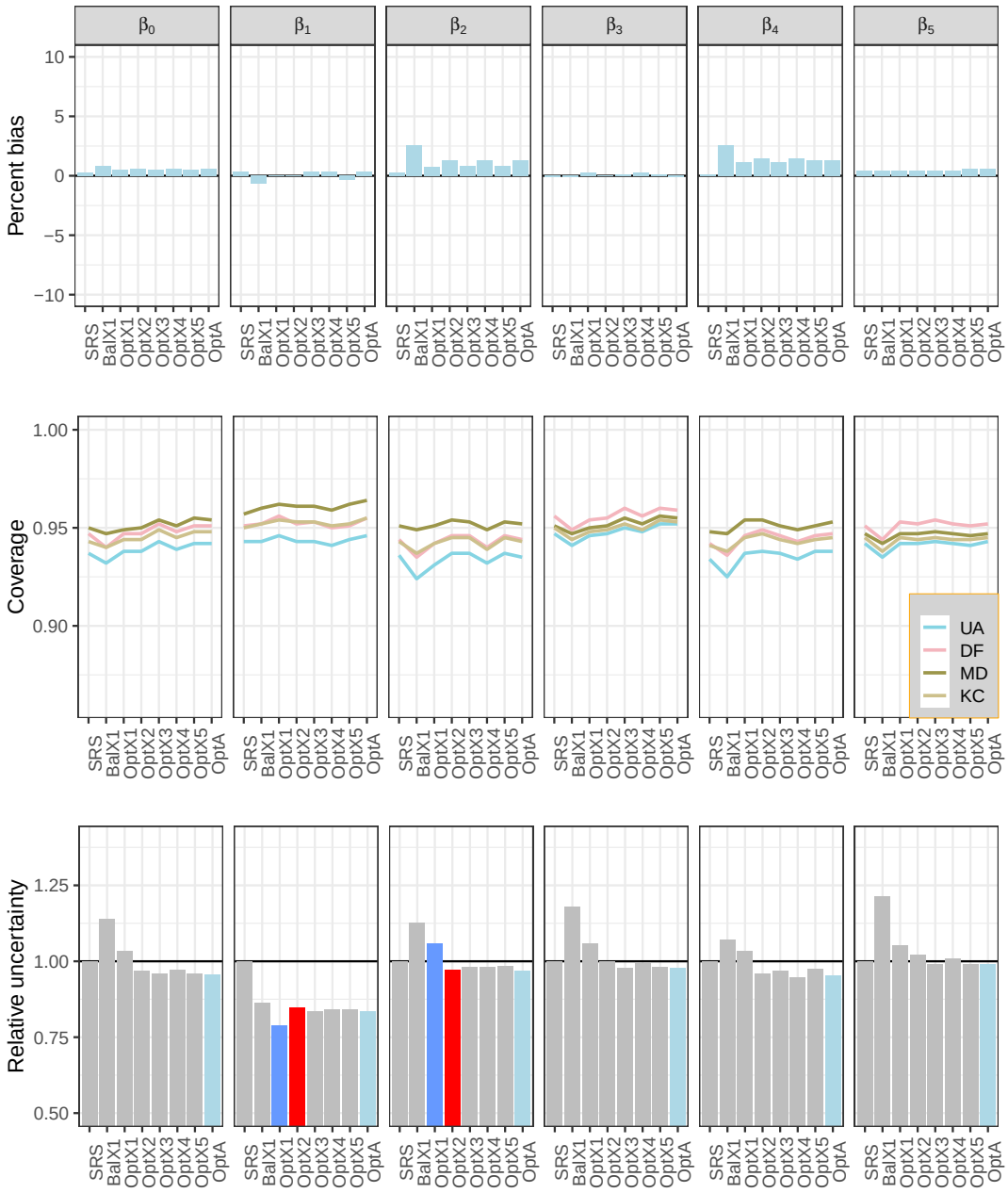

Figure B.4: Baseline scenario: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$.

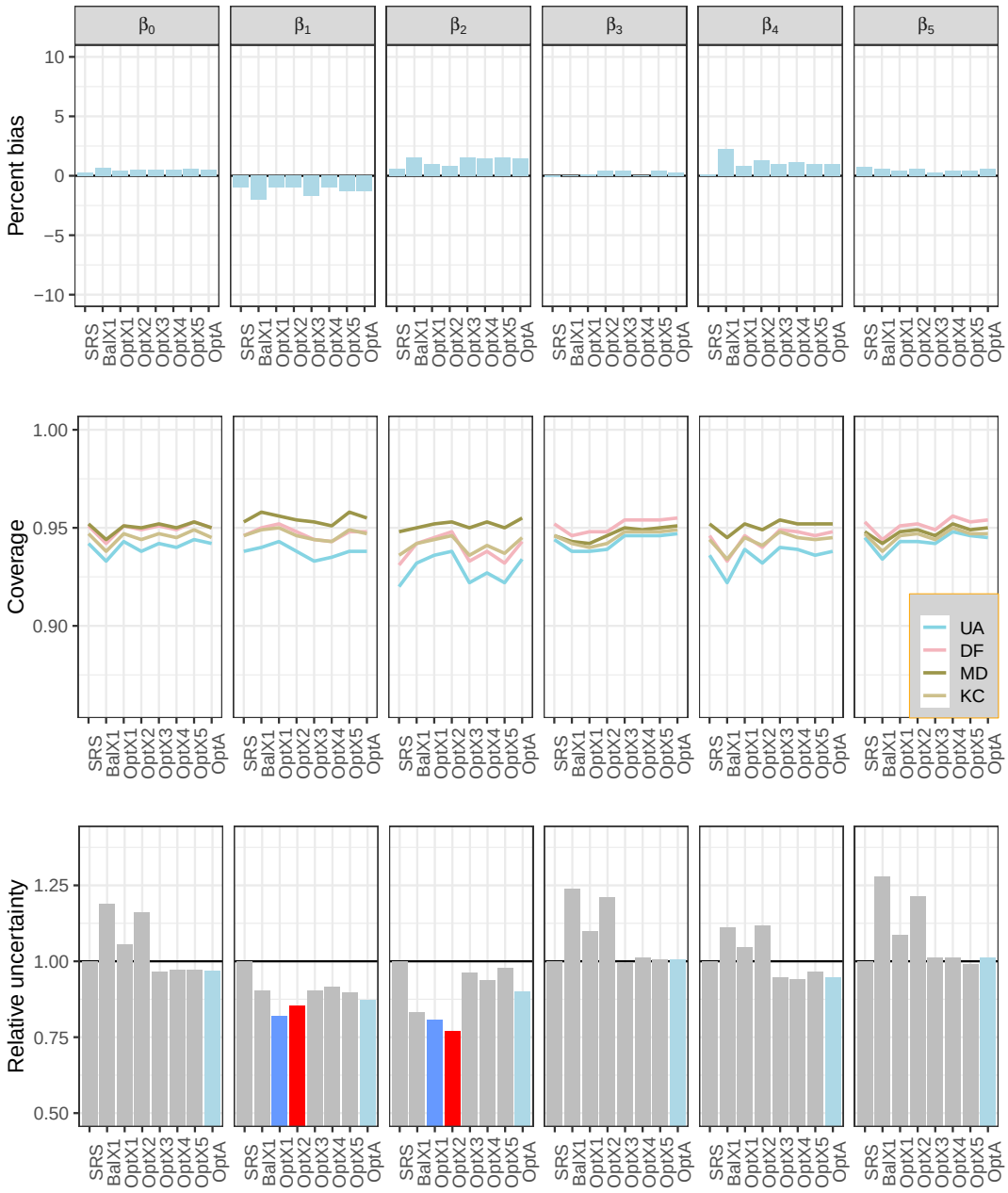

Figure B.5: Positive association X_1 and X_2: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$.

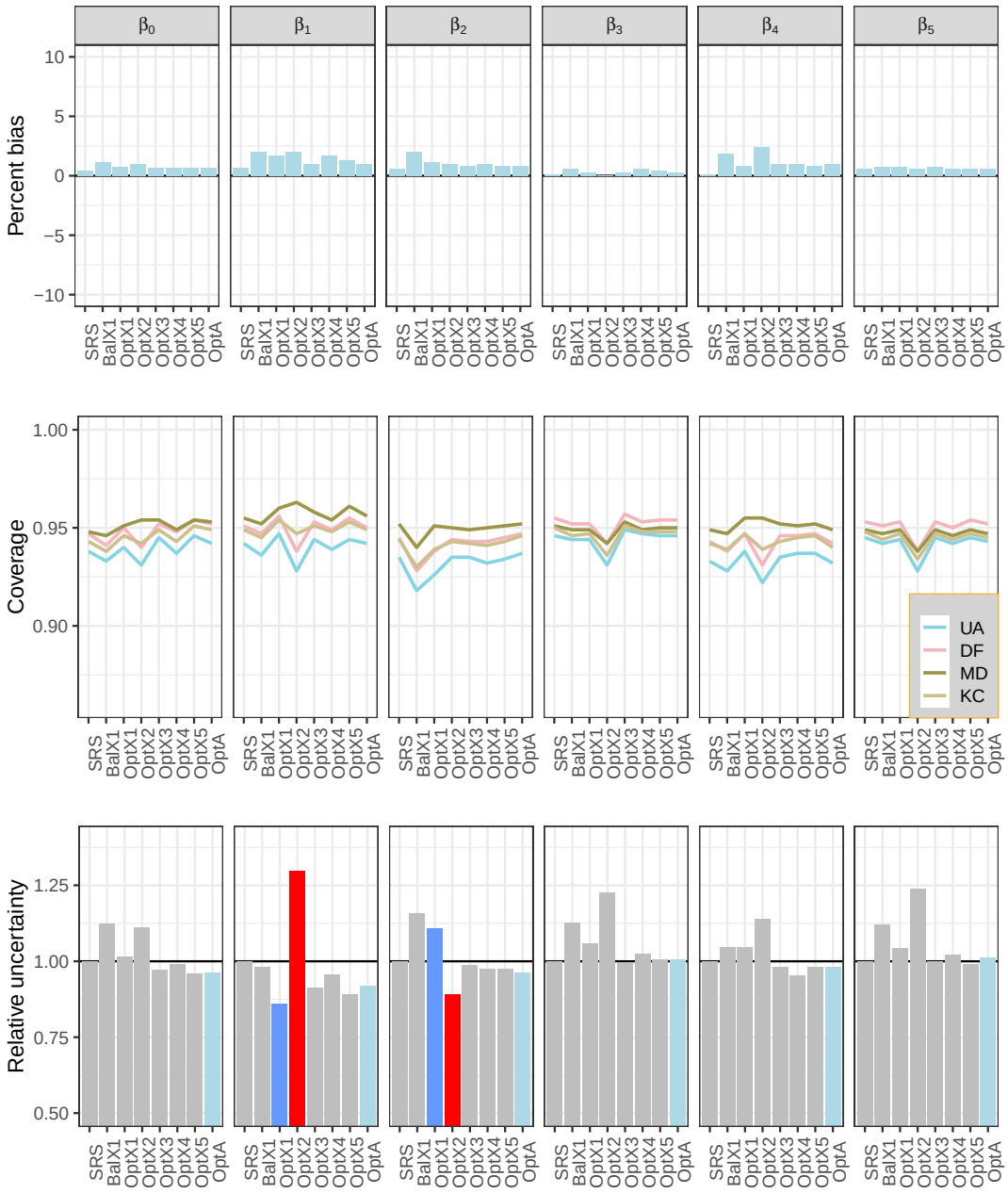

Figure B.6: Negative association X_1 and X_2: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$.

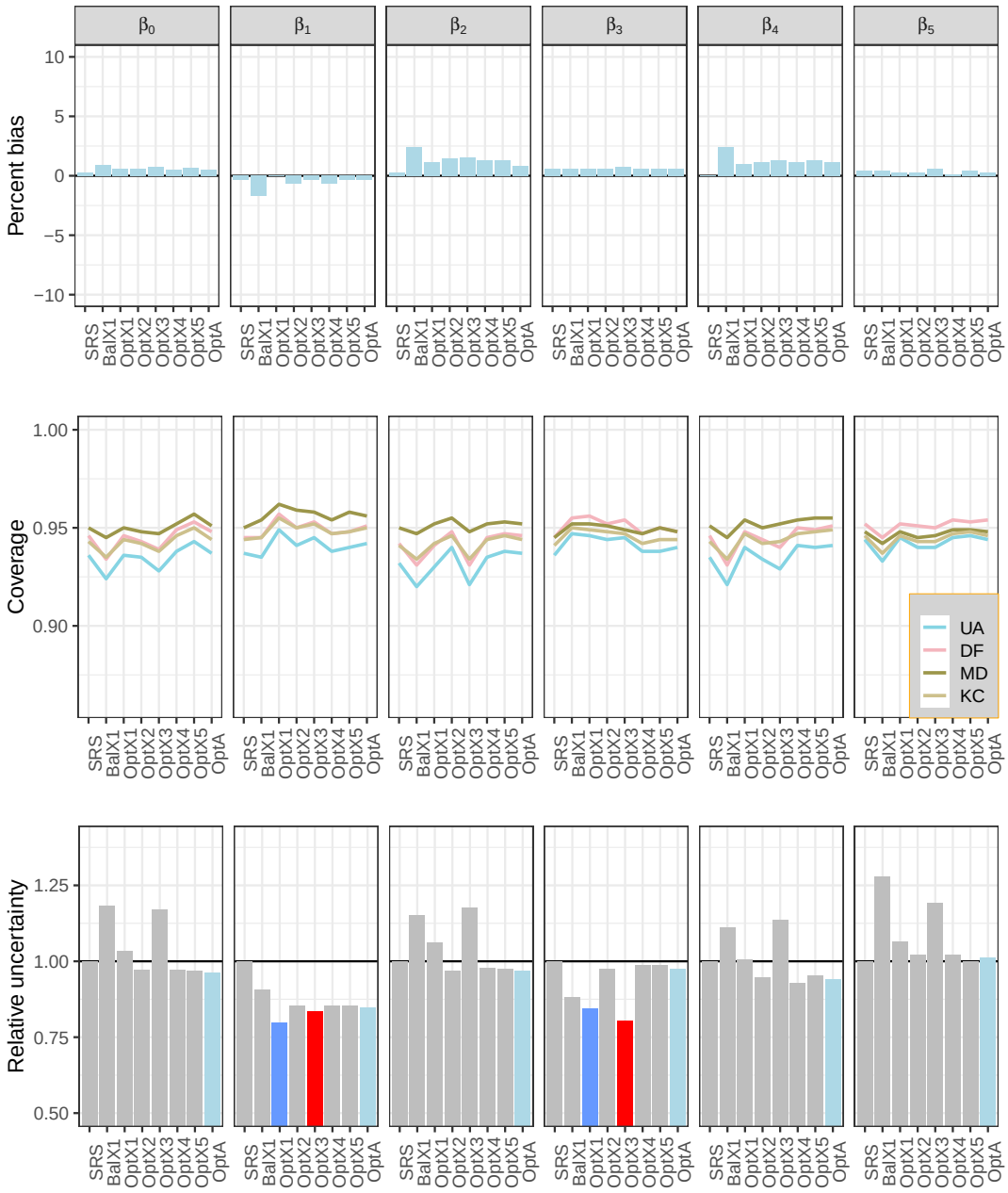

Figure B.7: Positive association X_1 and X_3: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$.

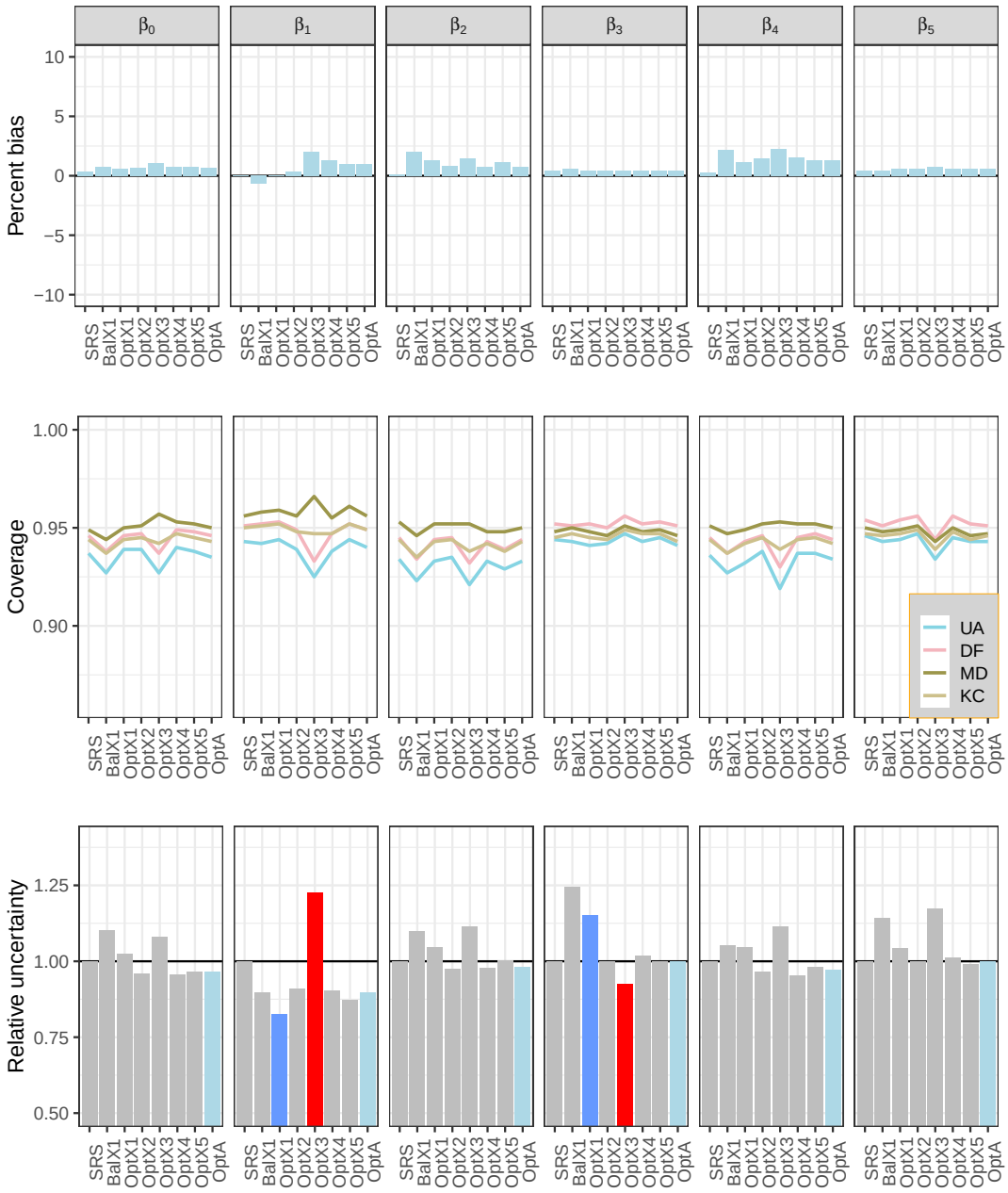

Figure B.8: Negative association X_1 and X_3: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$.

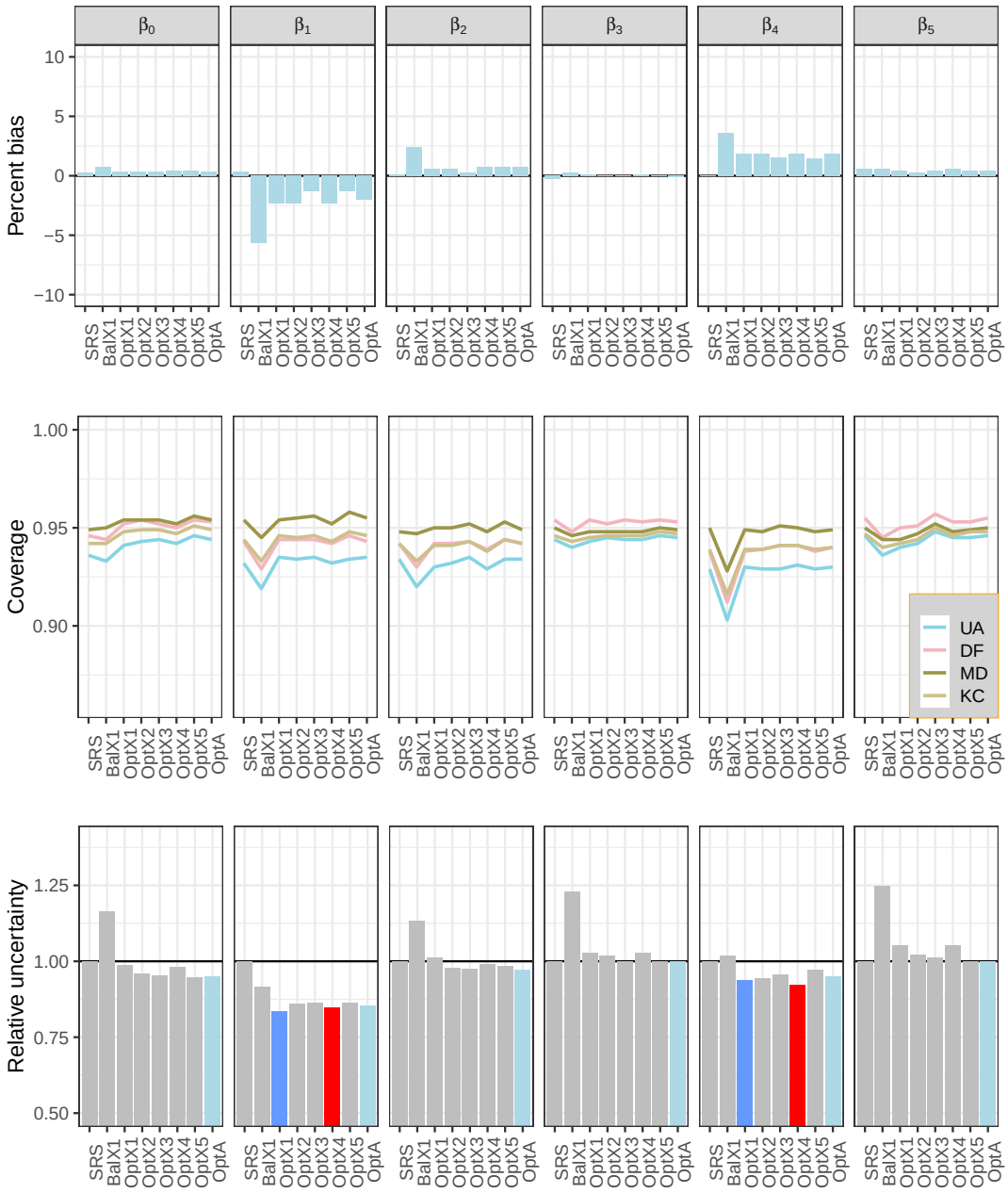

Figure B.9: Positive association X_1 and X_4: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$.

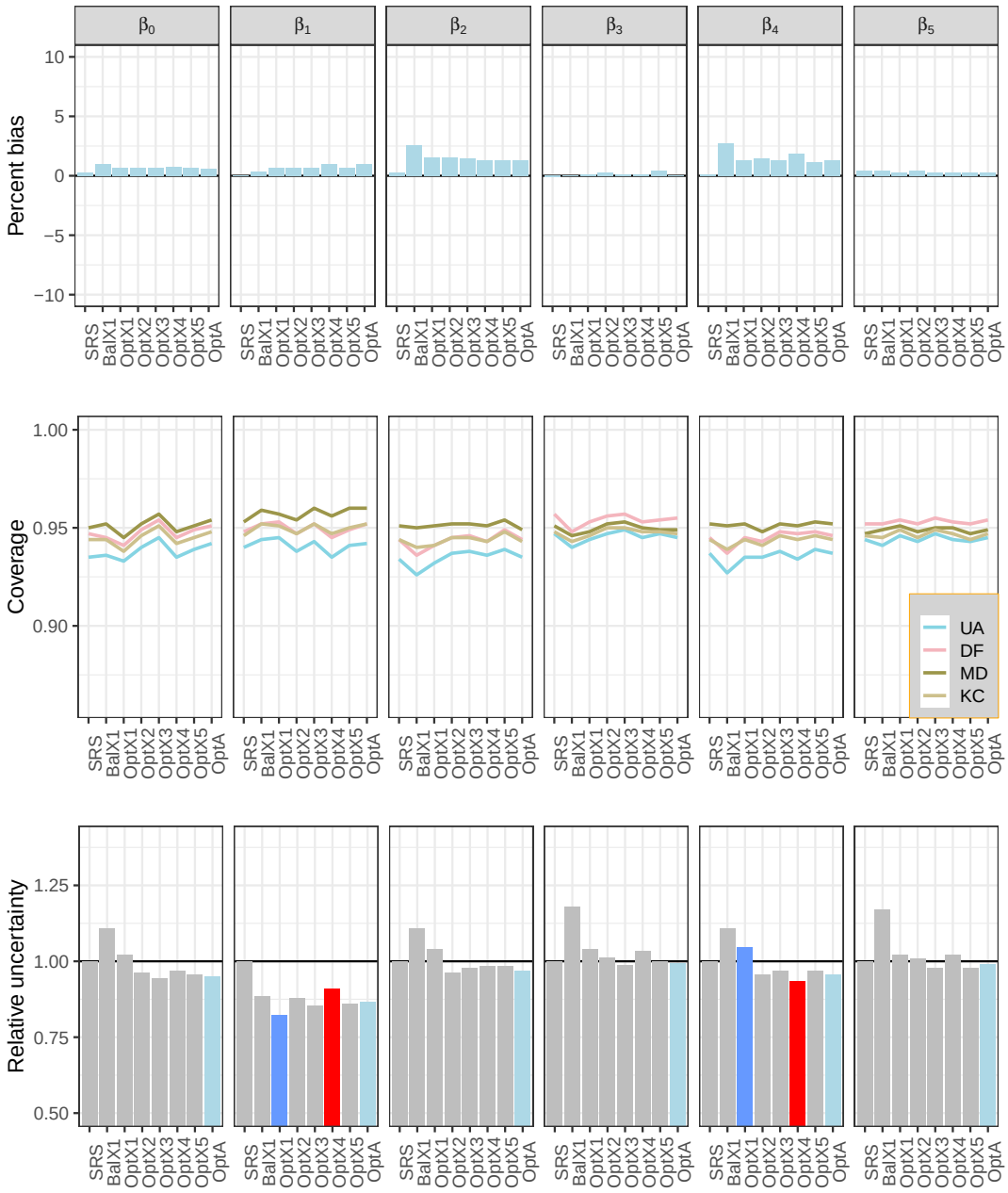

Figure B.10: Negative association X_1 and X_4: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$.

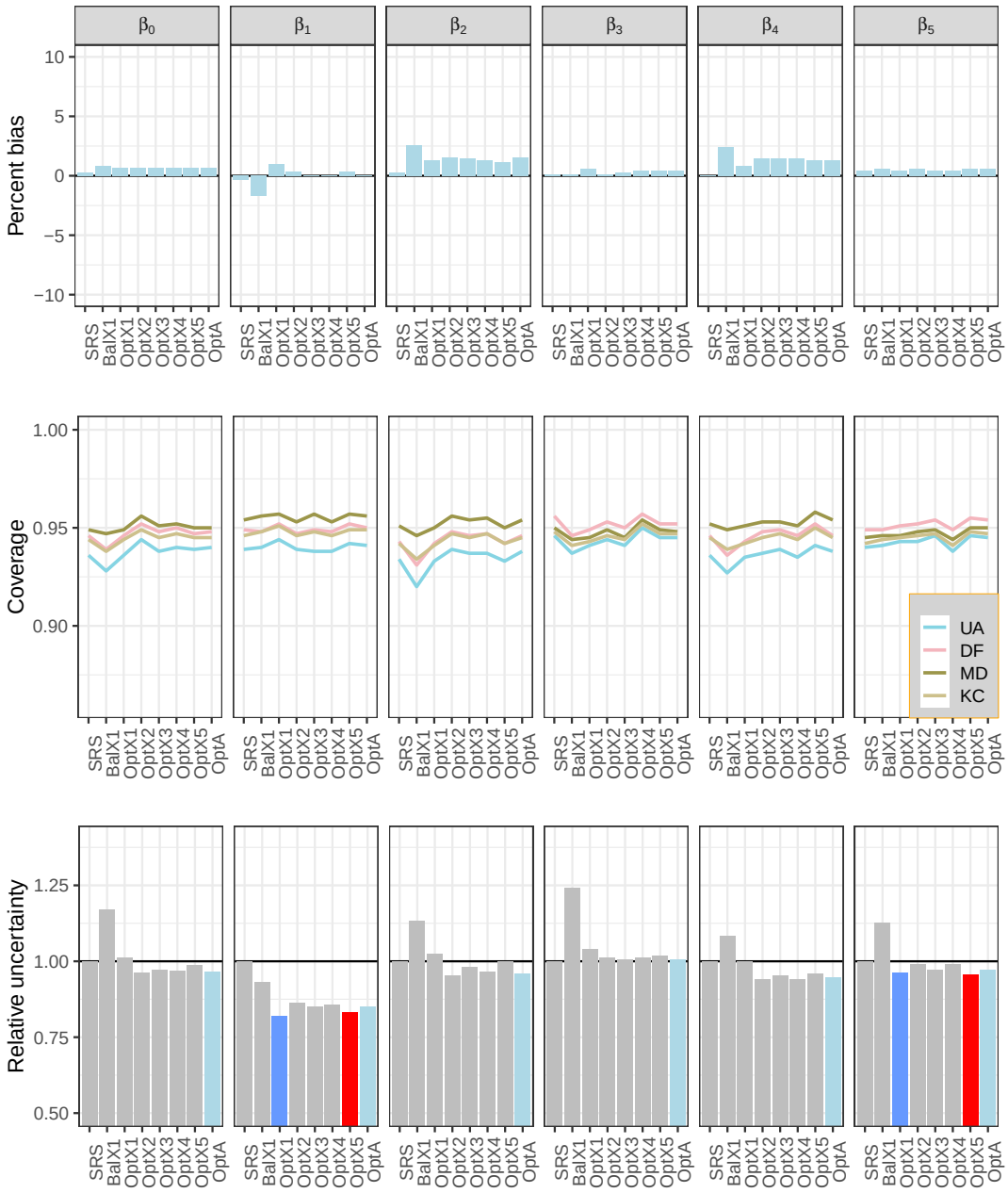

Figure B.11: Positive association X_1 and X_5: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$.

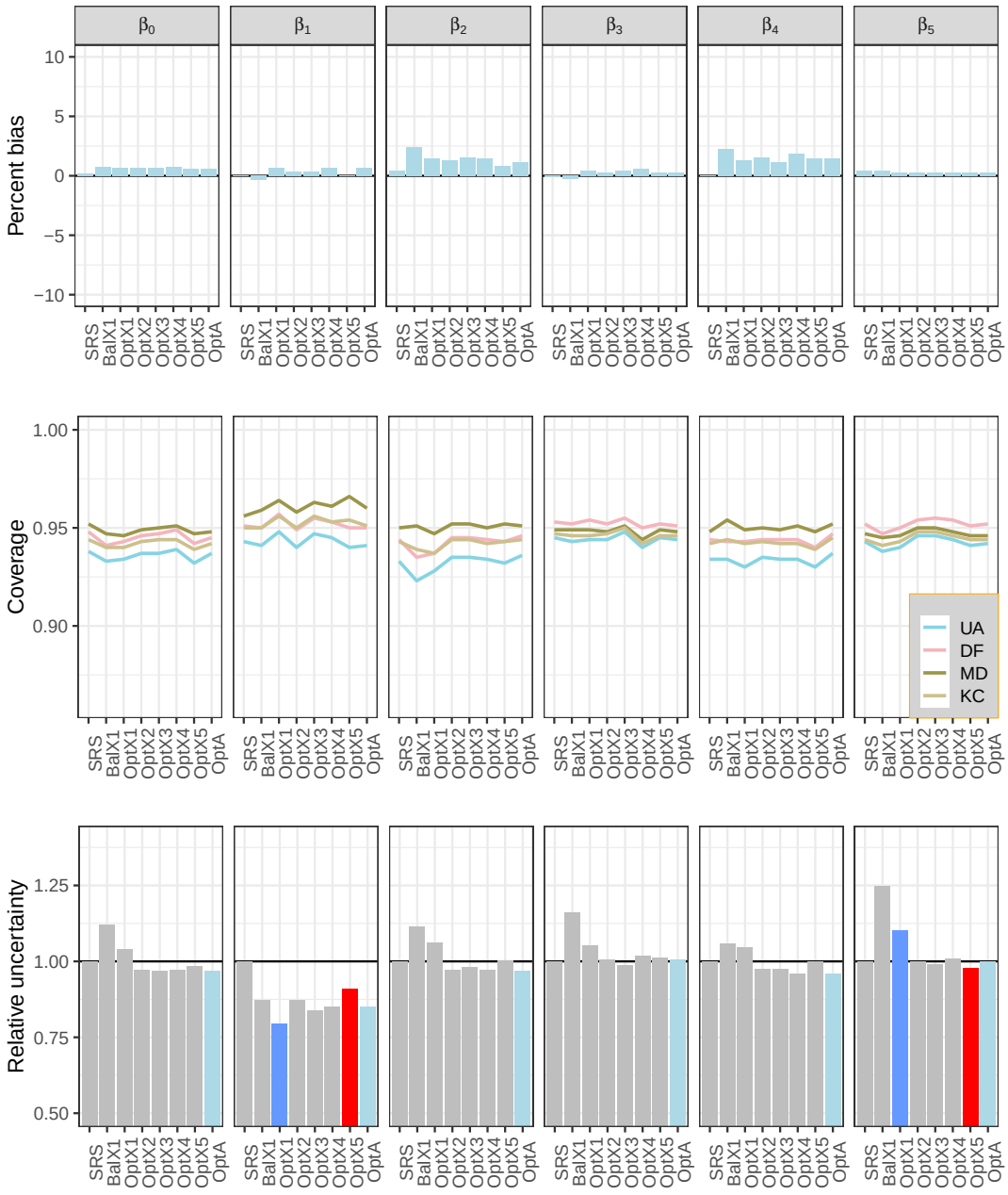

Figure B.12: Negative association X_1 and X_5: K=280, $N_k = 40\ \forall k = 1, \ldots, K, K_s = 80, \sigma_V = 0.5$.

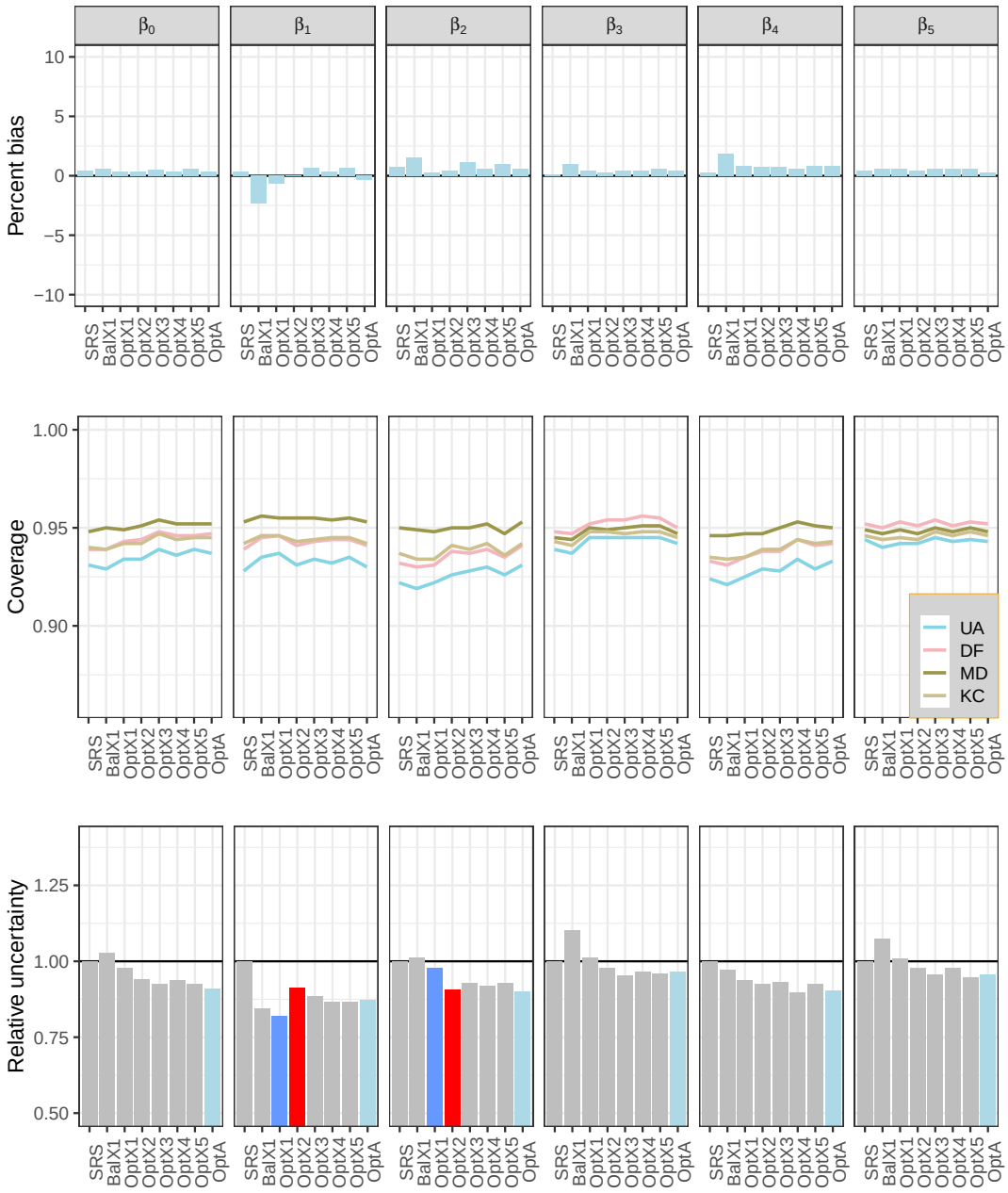

Figure B.13: Baseline scenario: K=280, varying $N_k, K_s = 80, \sigma_V = 0.5$.

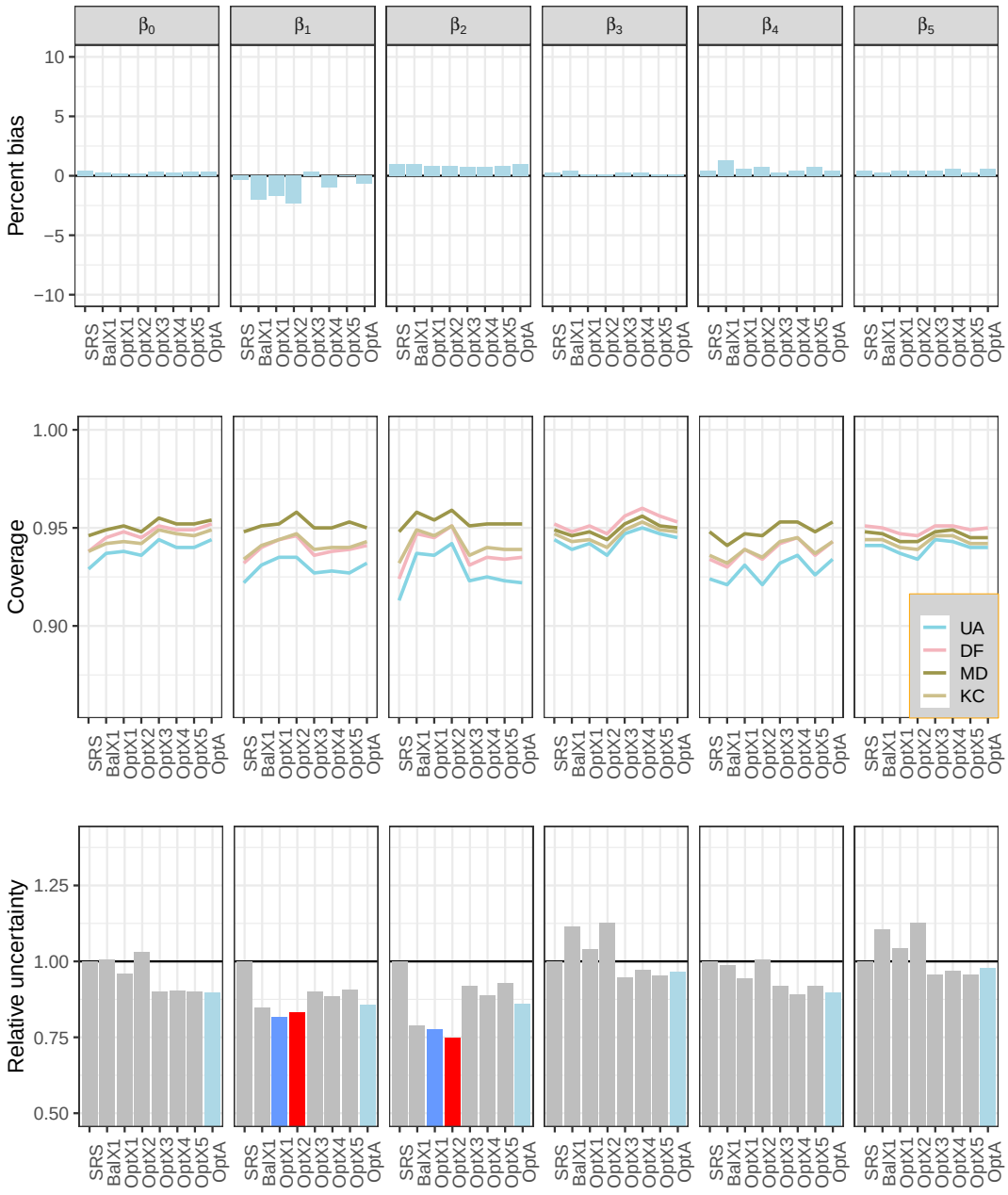

Figure B.14: Positive association X_1 and X_2: K=280, varying N_k, $K_s = 80$, $\sigma_V = 0.5$.

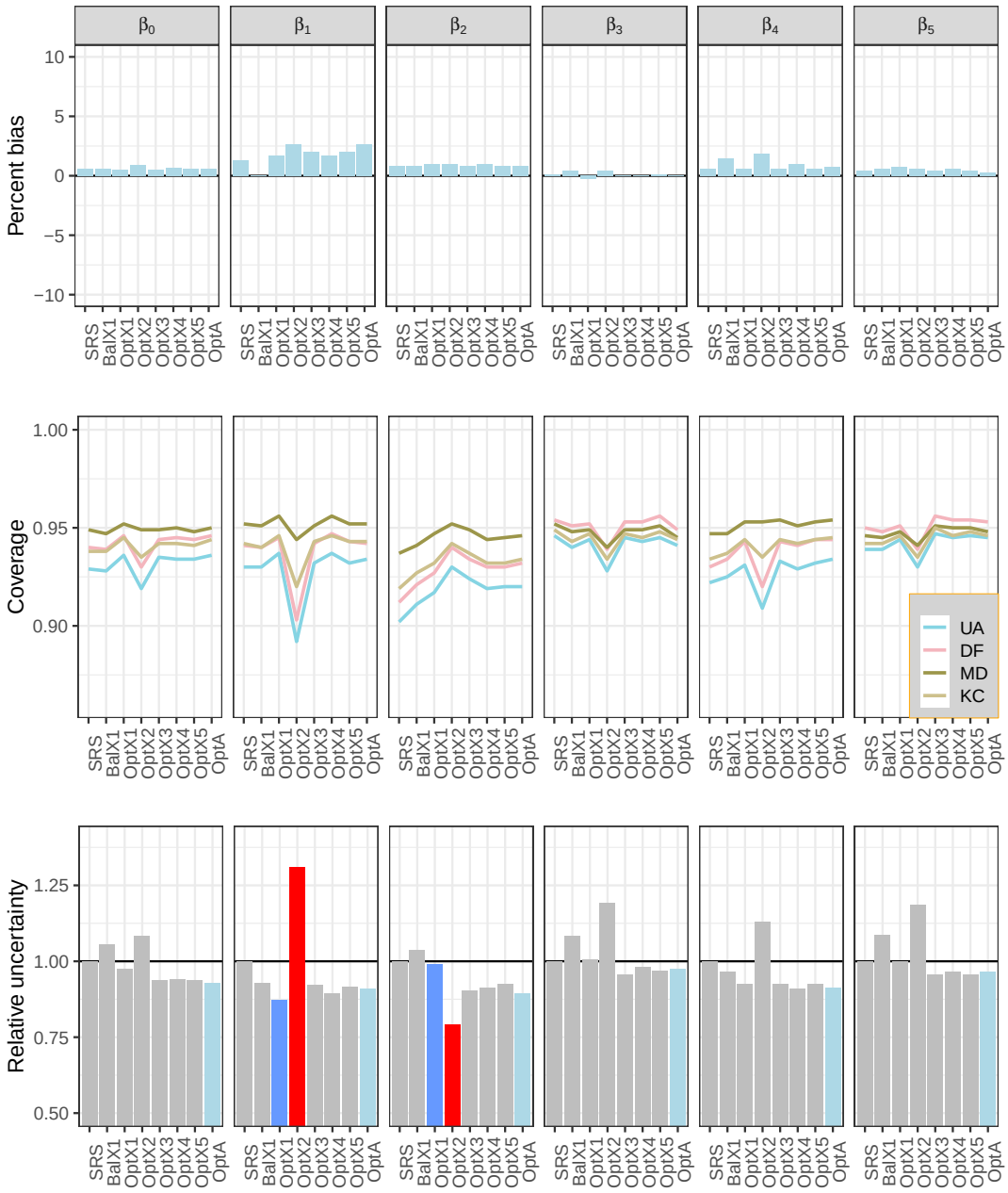

Figure B.15: Negative association X_1 and X_2: K=280, varying N_k, $K_s = 80$, $\sigma_V = 0.5$.

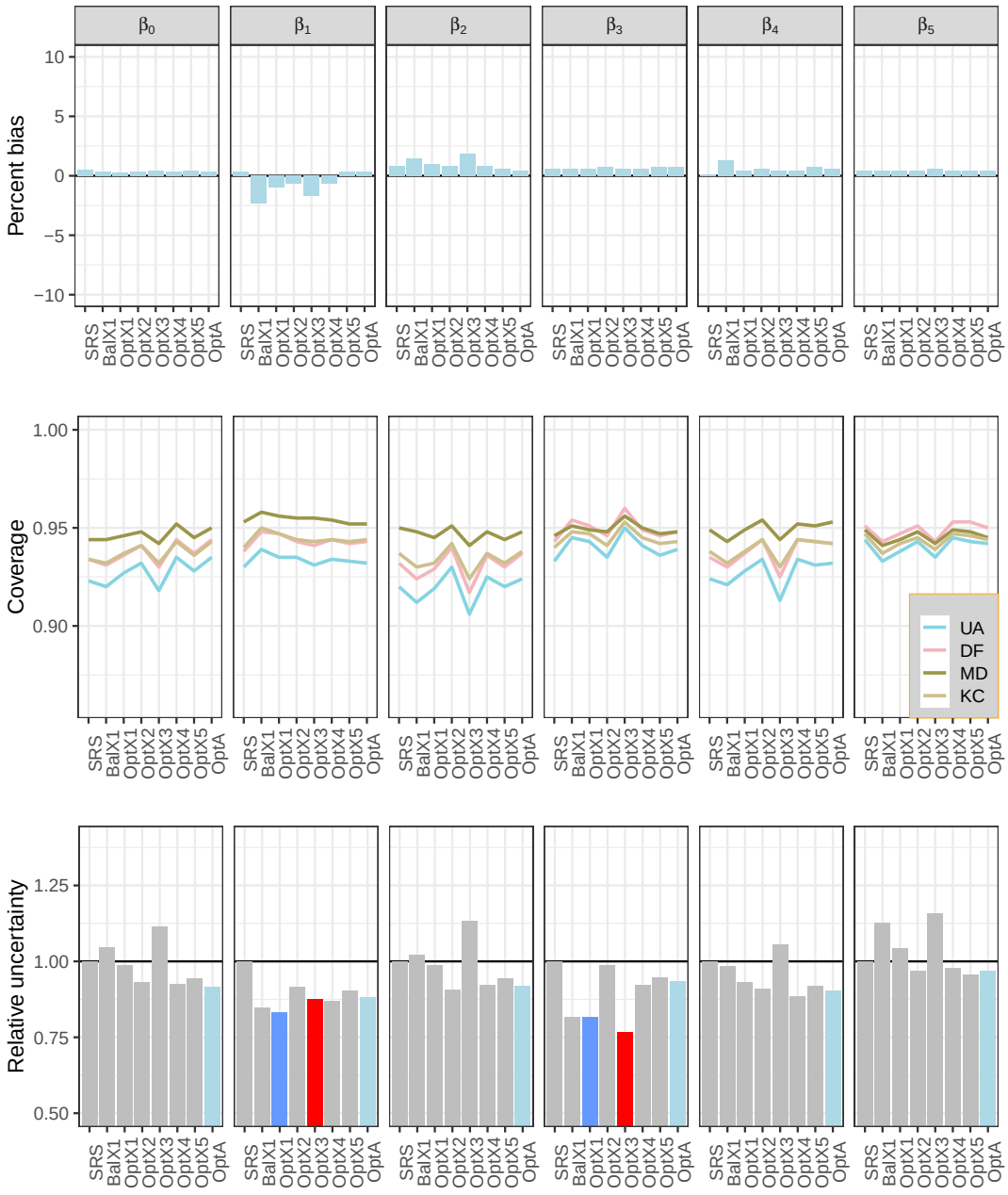

Figure B.16: Positive association X_1 and X_3: K=280, varying N_k, $K_s = 80$, $\sigma_V = 0.5$.

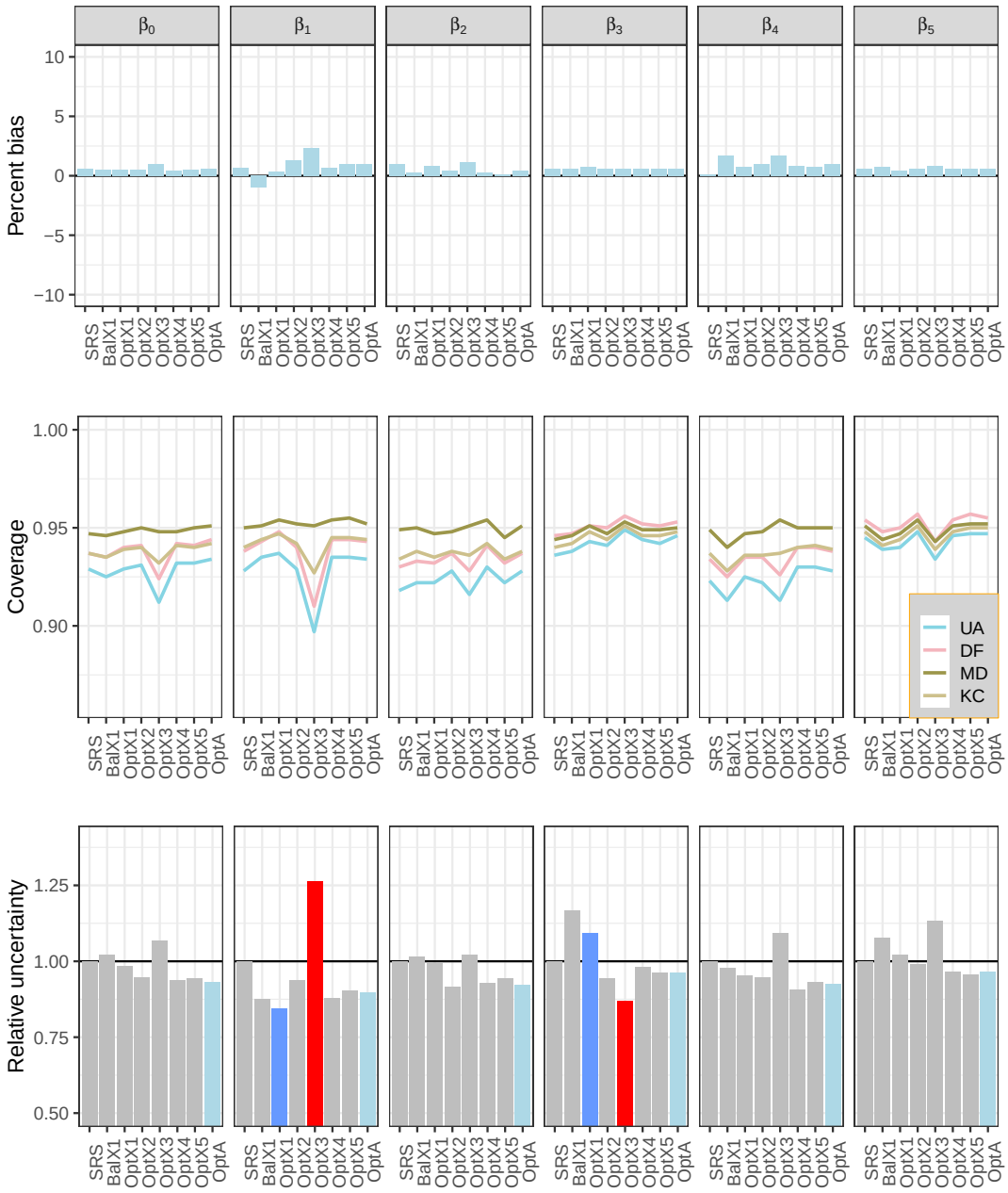

Figure B.17: Negative association X_1 and X_3: K=280, varying N_k, $K_s = 80$, $\sigma_V = 0.5$.

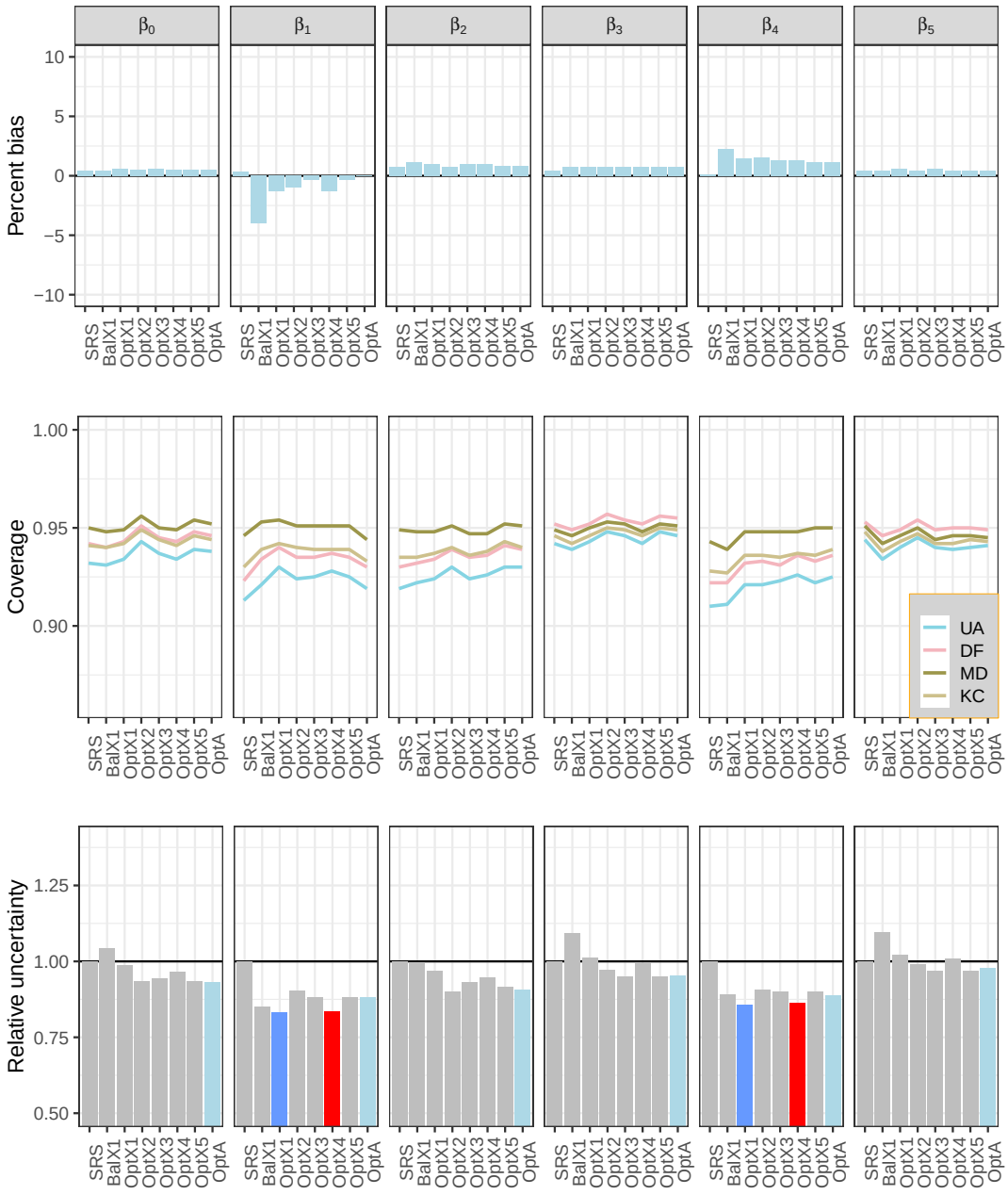

Figure B.18: Positive association X_1 and X_4: K=280, varying N_k, $K_s = 80$, $\sigma_V = 0.5$.

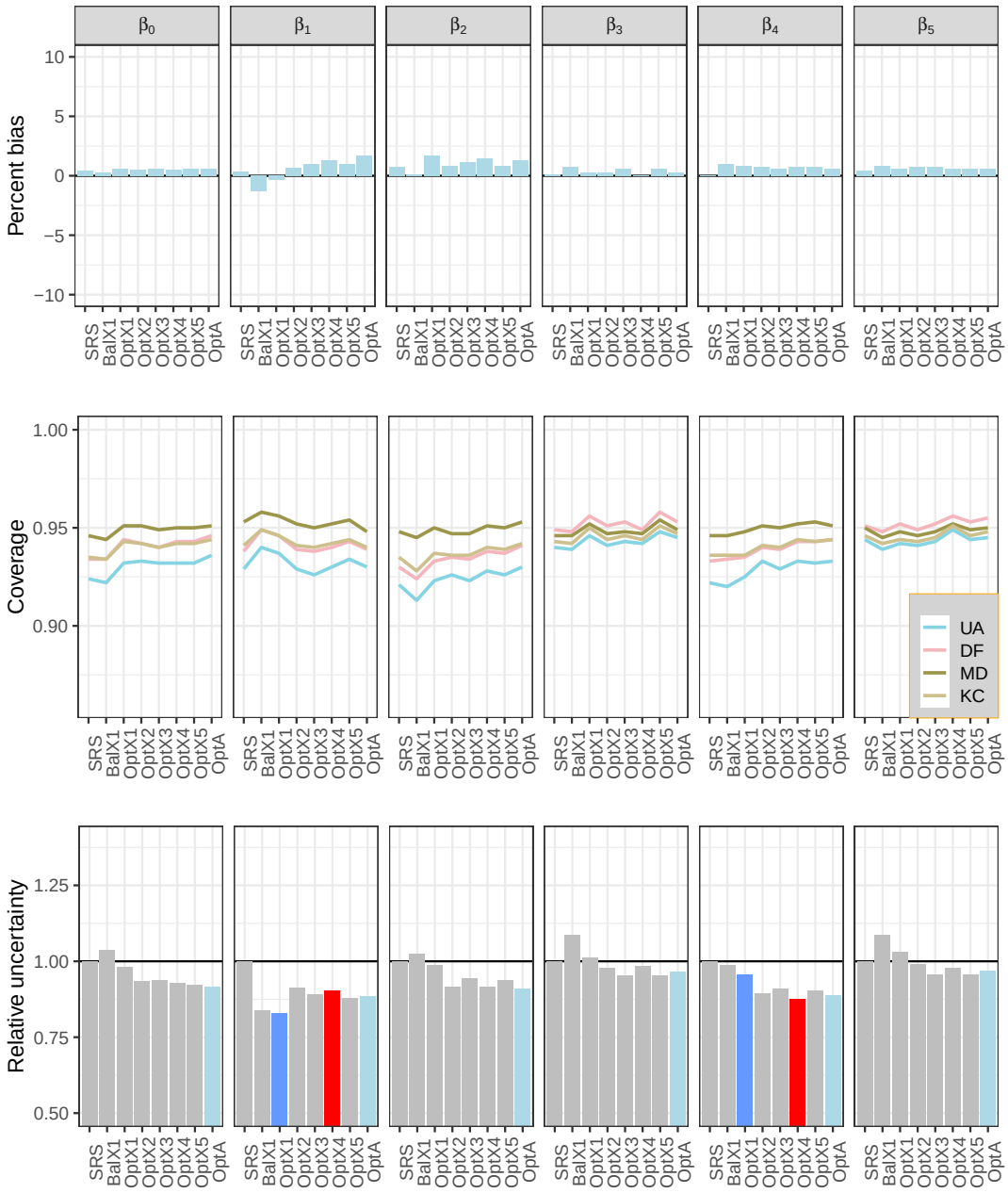

Figure B.19: Negative association X_1 and X_4: K=280, varying N_k, $K_s = 80$, $\sigma_V = 0.5$.

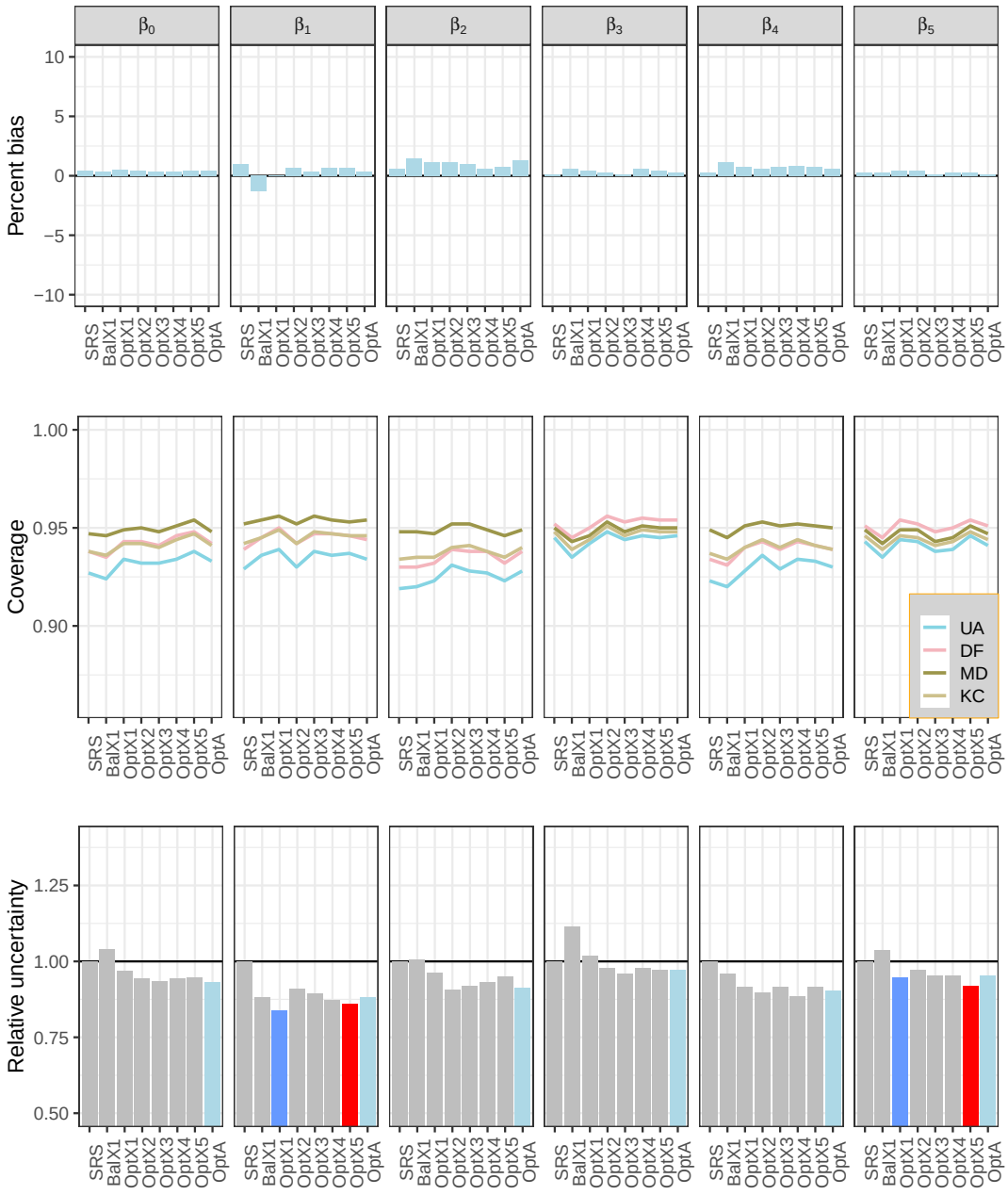

Figure B.20: Positive association X_1 and X_5: K=280, varying N_k, $K_s = 80$, $\sigma_V = 0.5$.

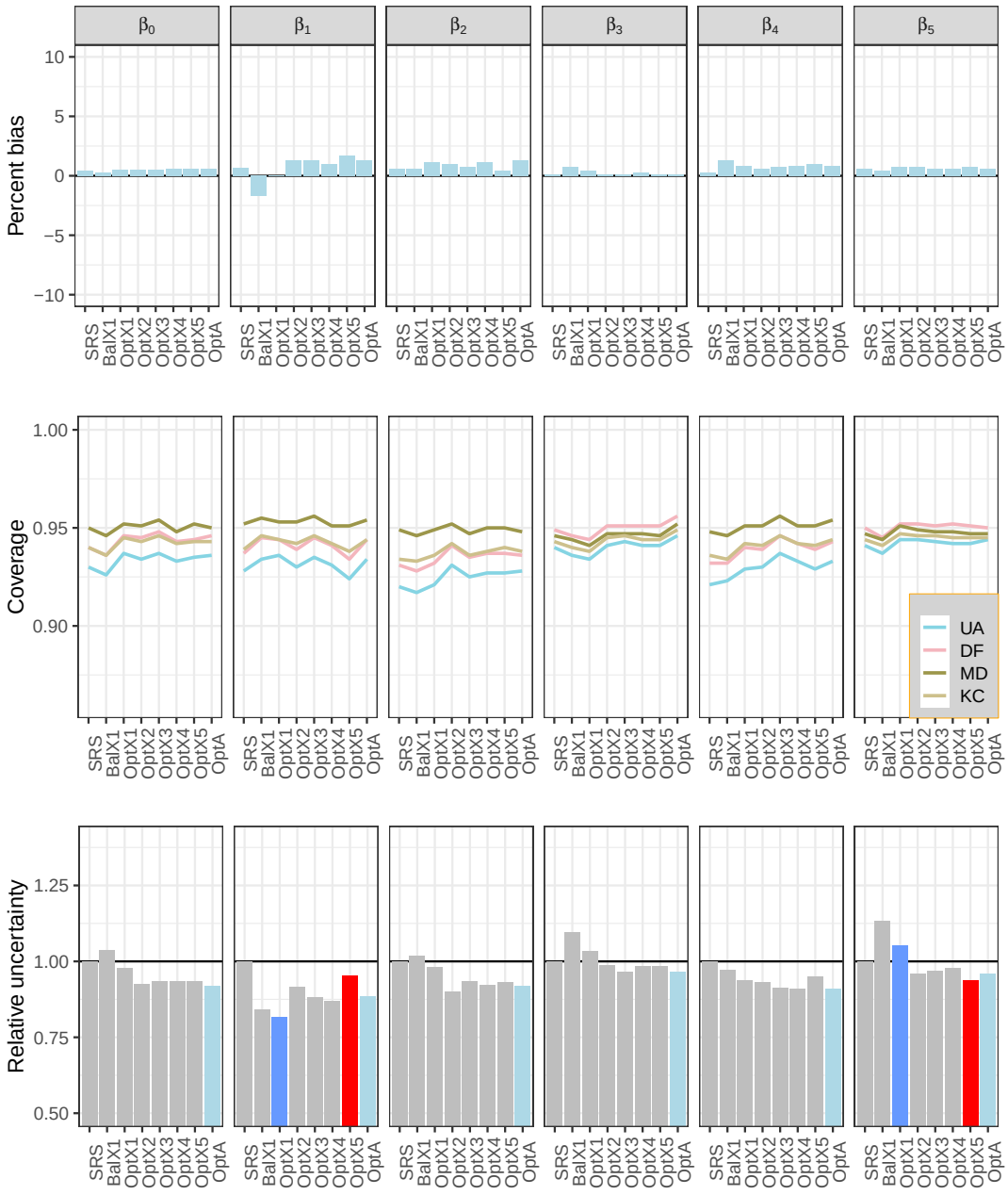

Figure B.21: Negative association X_1 and X_5: K=280, varying N_k, $K_s = 80$, $\sigma_V = 0.5$.

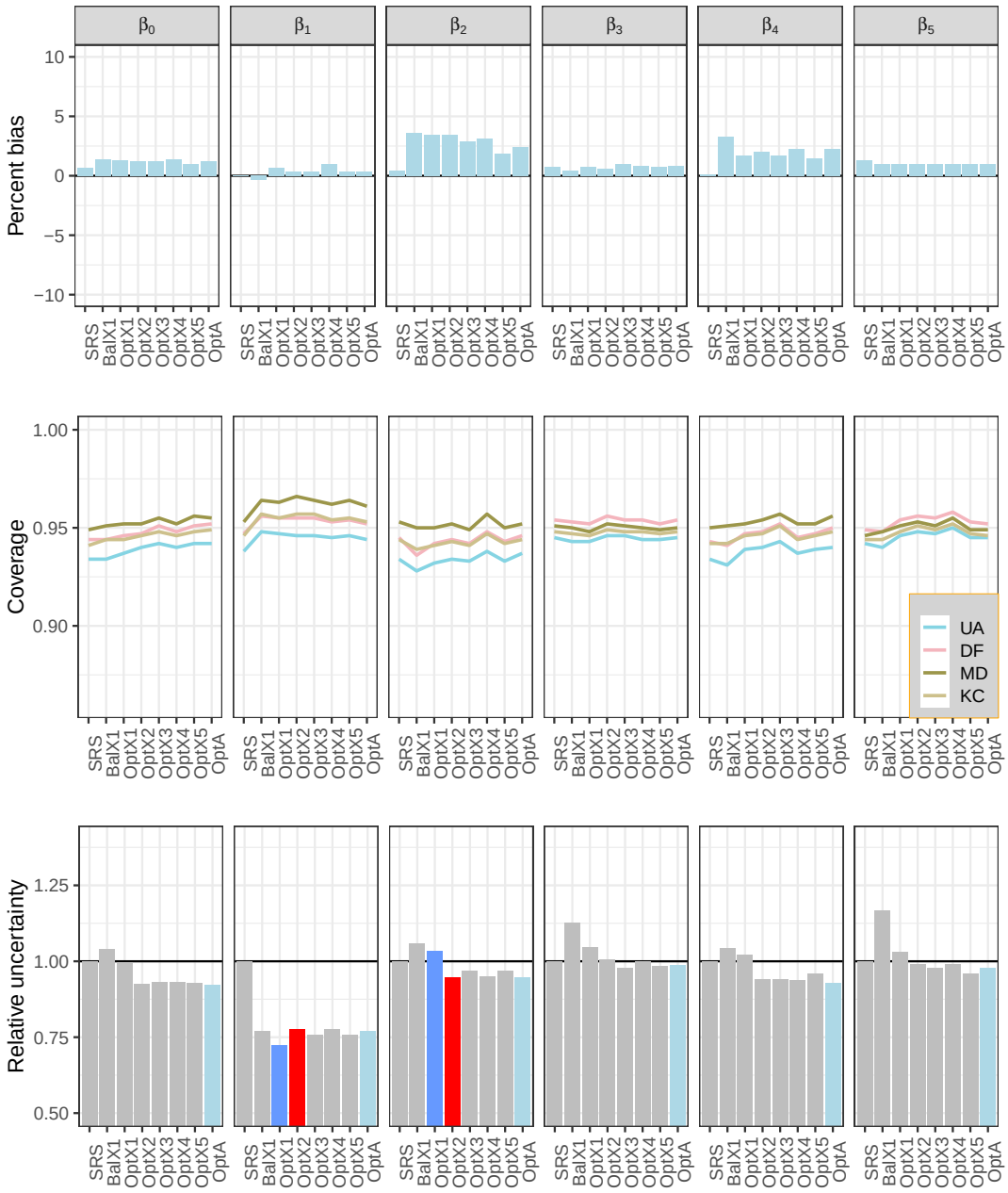

Figure B.22: Baseline scenario: K=280, equal N_k=40, $K_s = 80$, $\sigma_V = 1$.

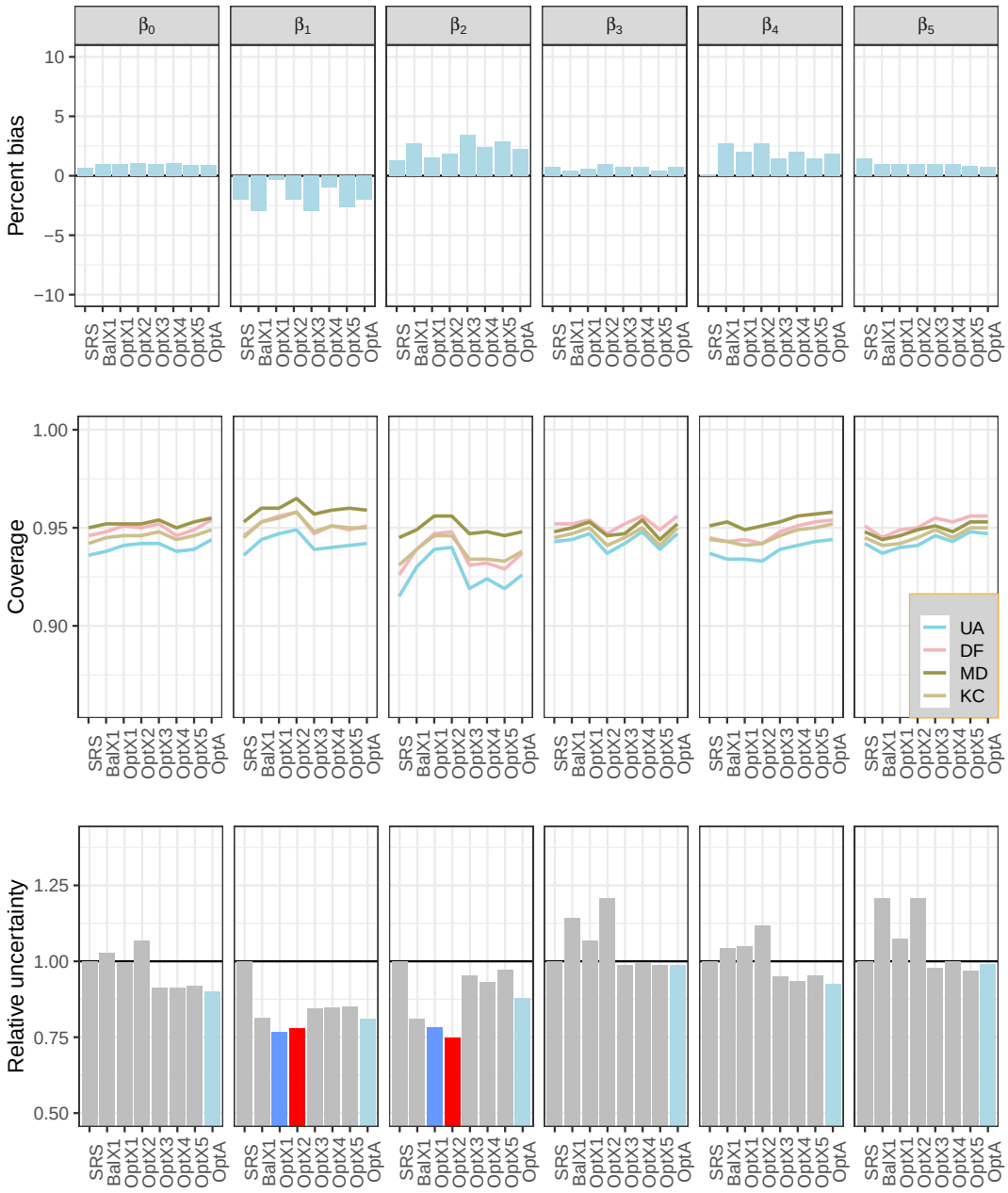

Figure B.23: Positive association X_1 and X_2: K=280, equal N_k=40, $K_s = 80$, $\sigma_V = 1$.

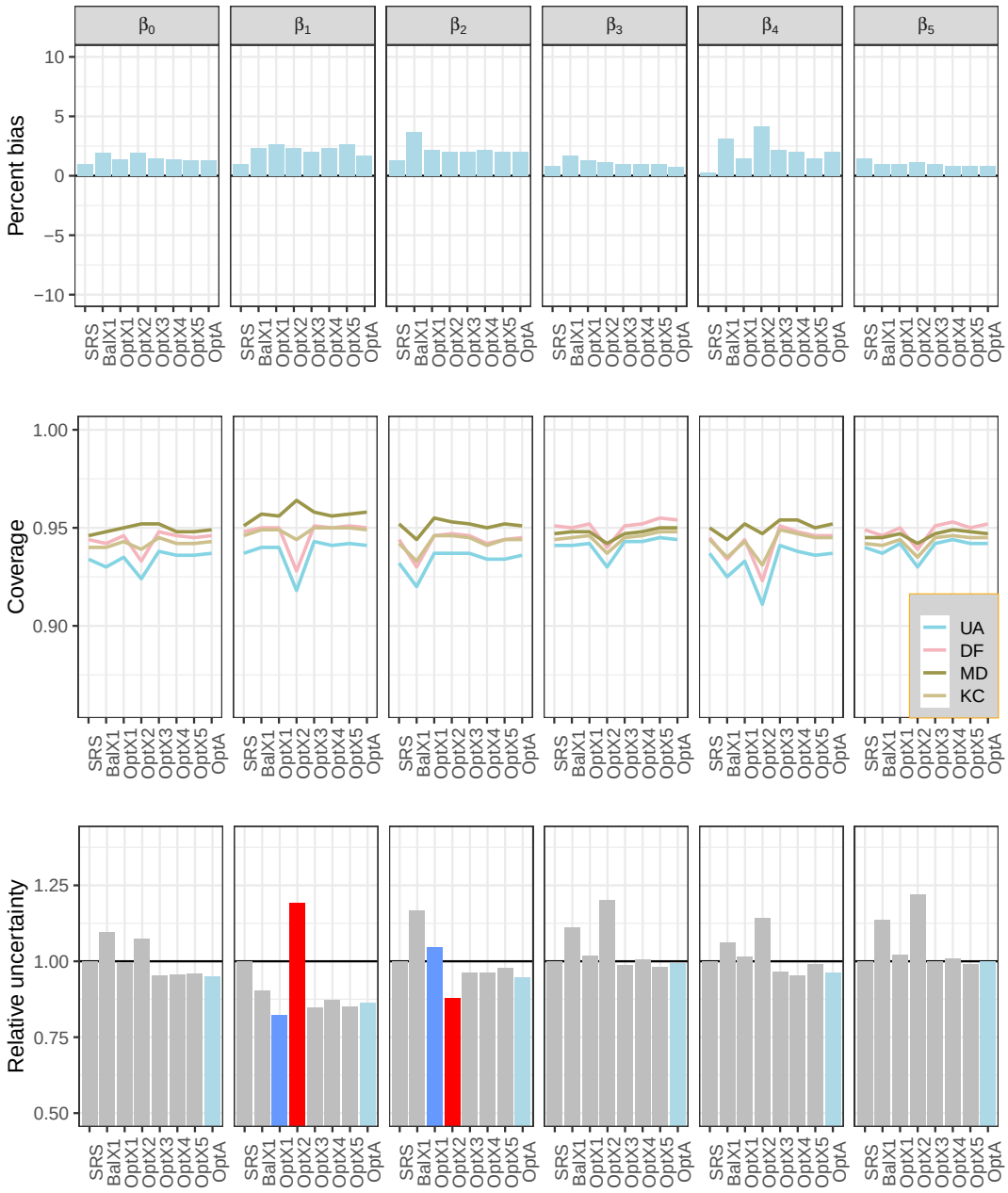

Figure B.24: Negative association X_1 and X_2: K=280, equal N_k=40, $K_s = 80$, $\sigma_V = 1$.

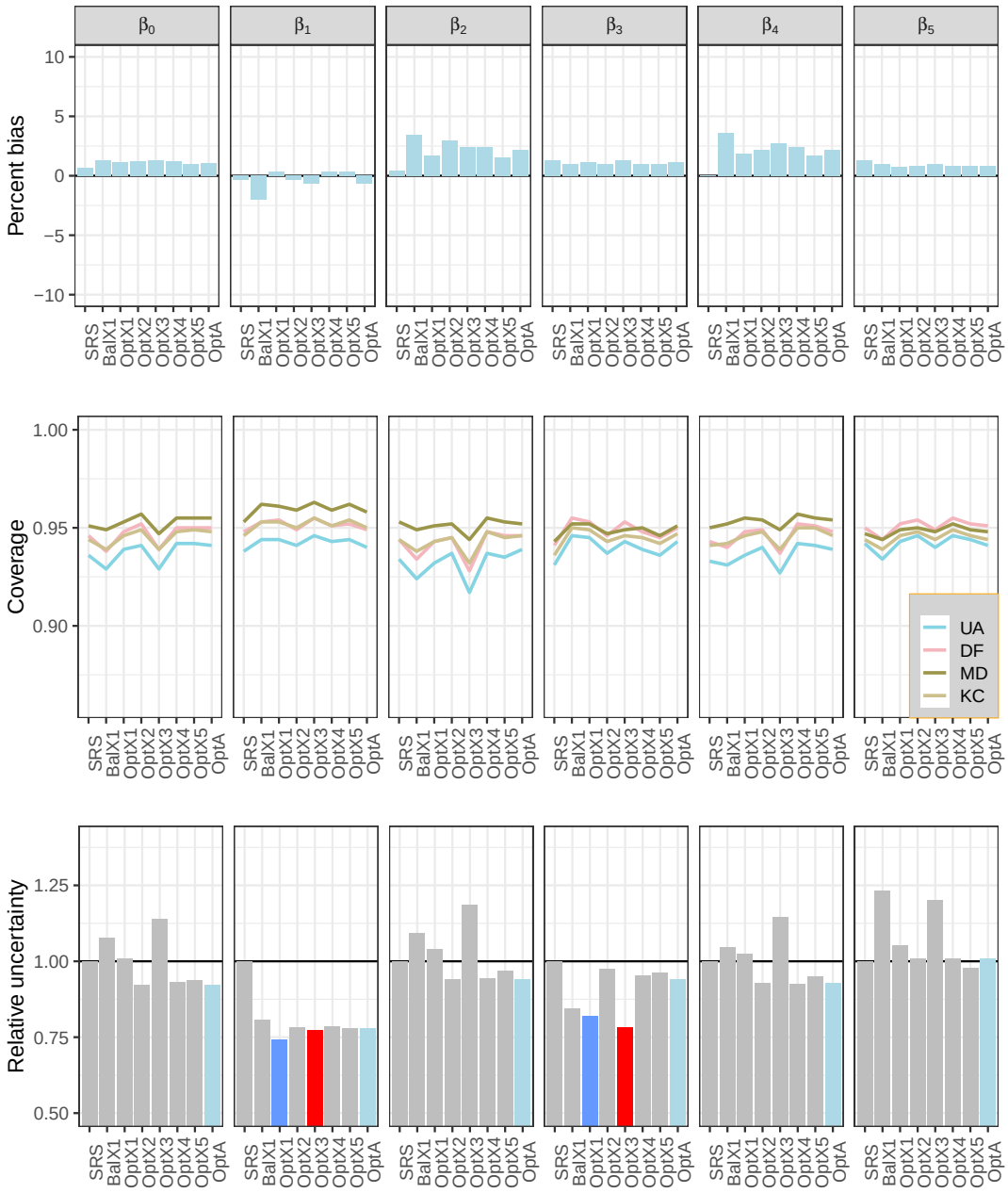

Figure B.25: Positive association X_1 and X_3: K=280, equal N_k=40, $K_s = 80$, $\sigma_V = 1$.

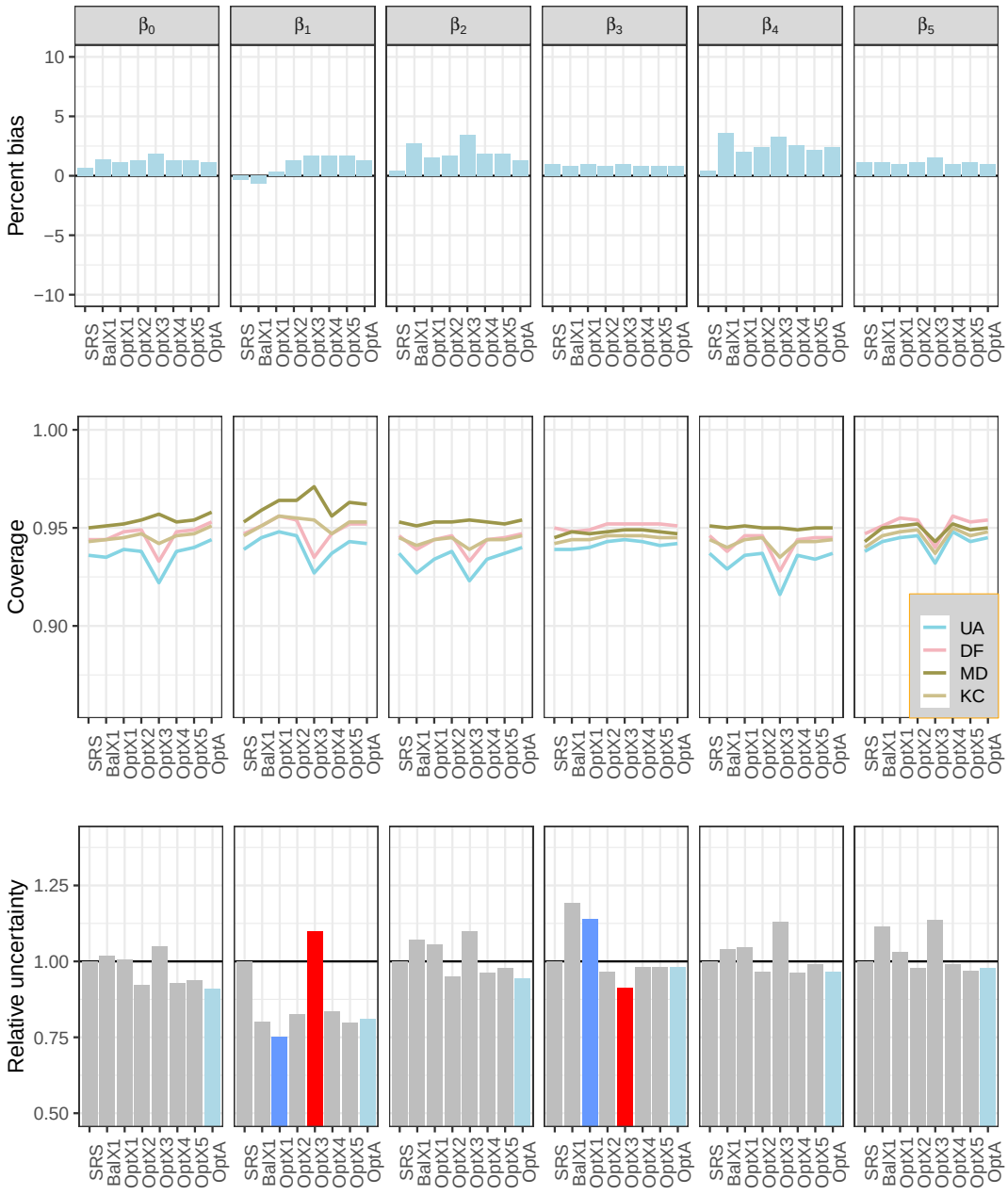

Figure B.26: Negative associaiton X_1 and X_3: K=280, equal N_k=40, $K_s = 80, \sigma_V = 1$.

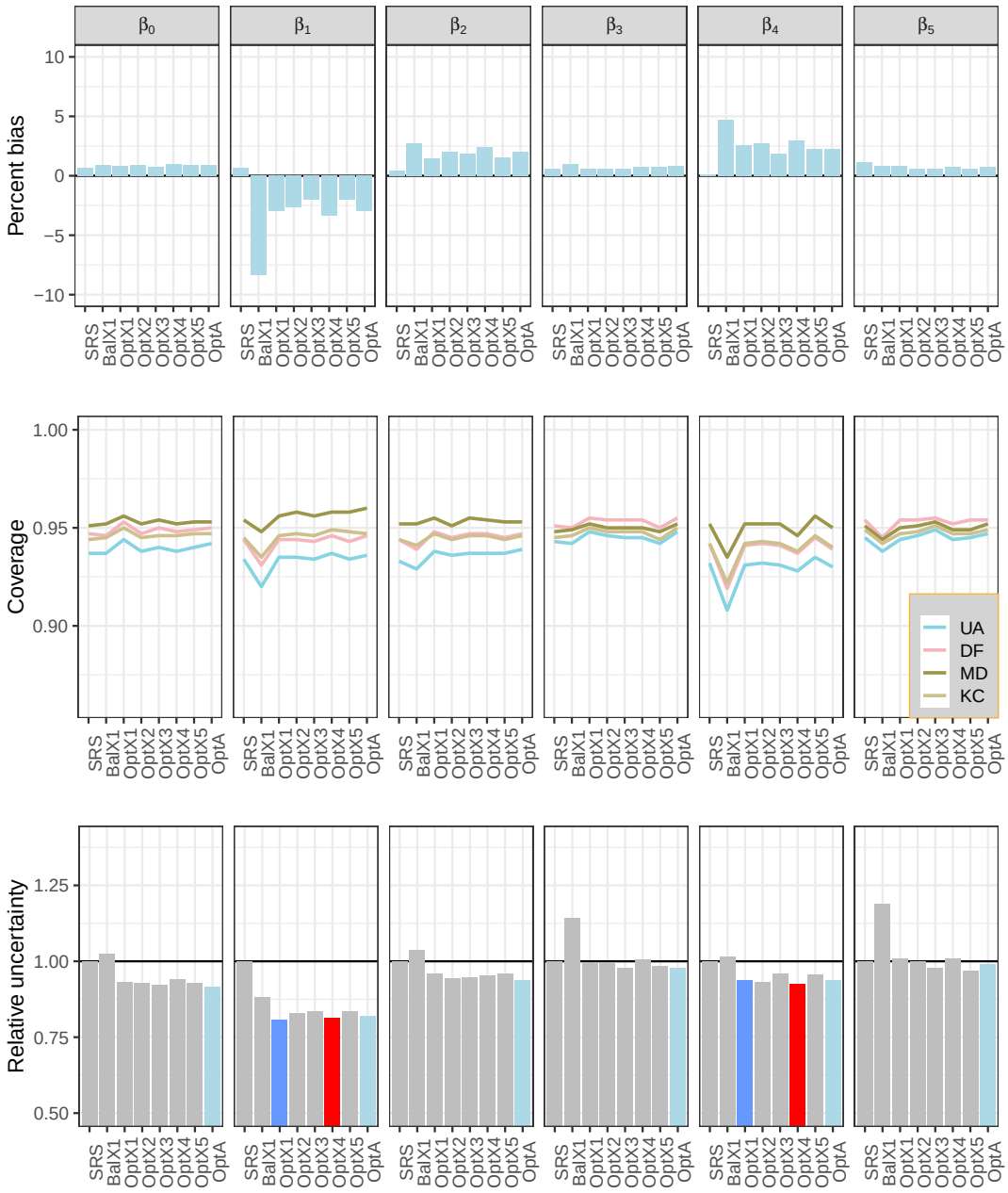

Figure B.27: Positive association X_1 and X_4: K=280, equal N_k=40, $K_s = 80$, $\sigma_V = 1$.

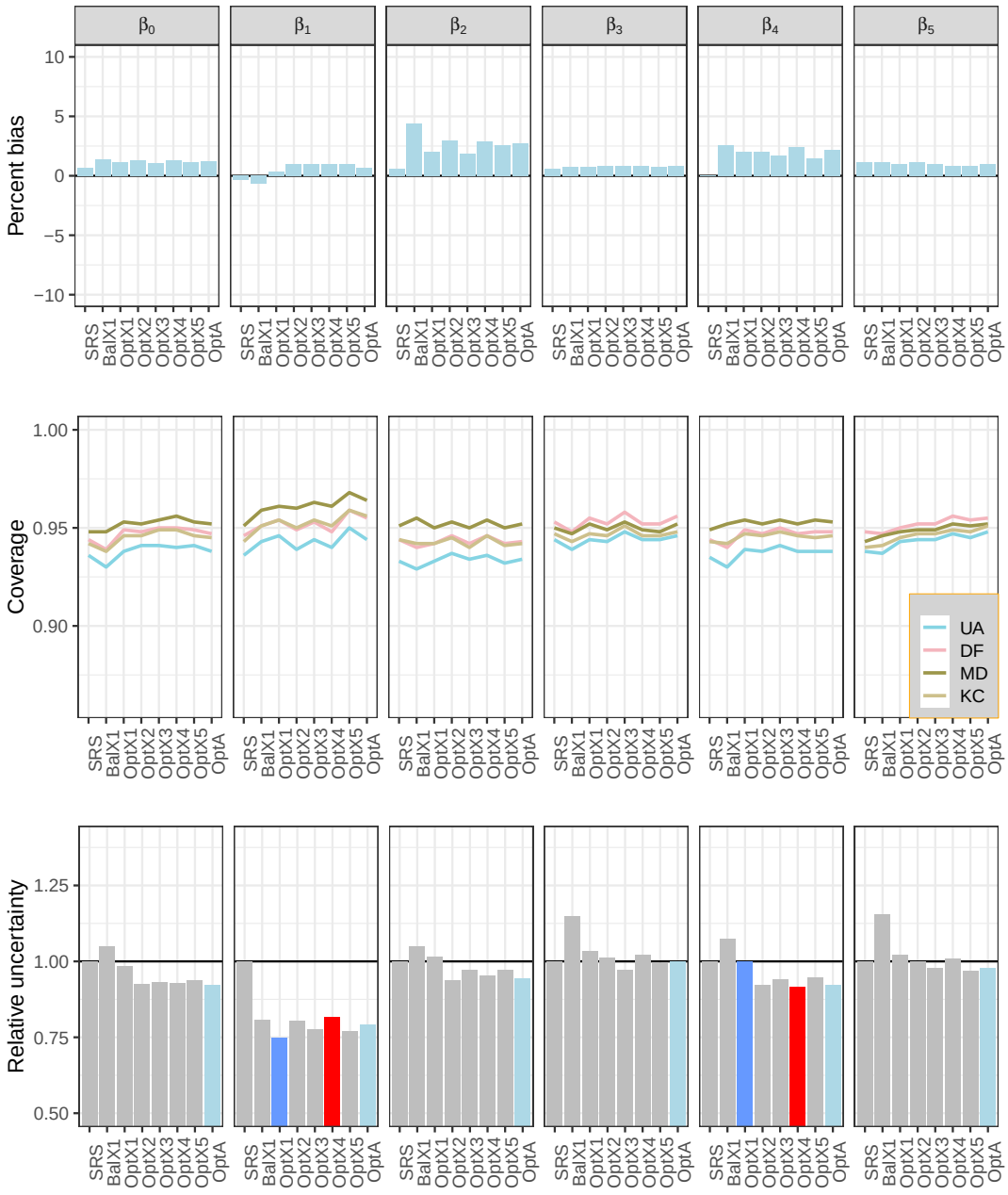

Figure B.28: Positive association X_1 and X_4: K=280, equal N_k=40, $K_s = 80$, $\sigma_V = 1$.

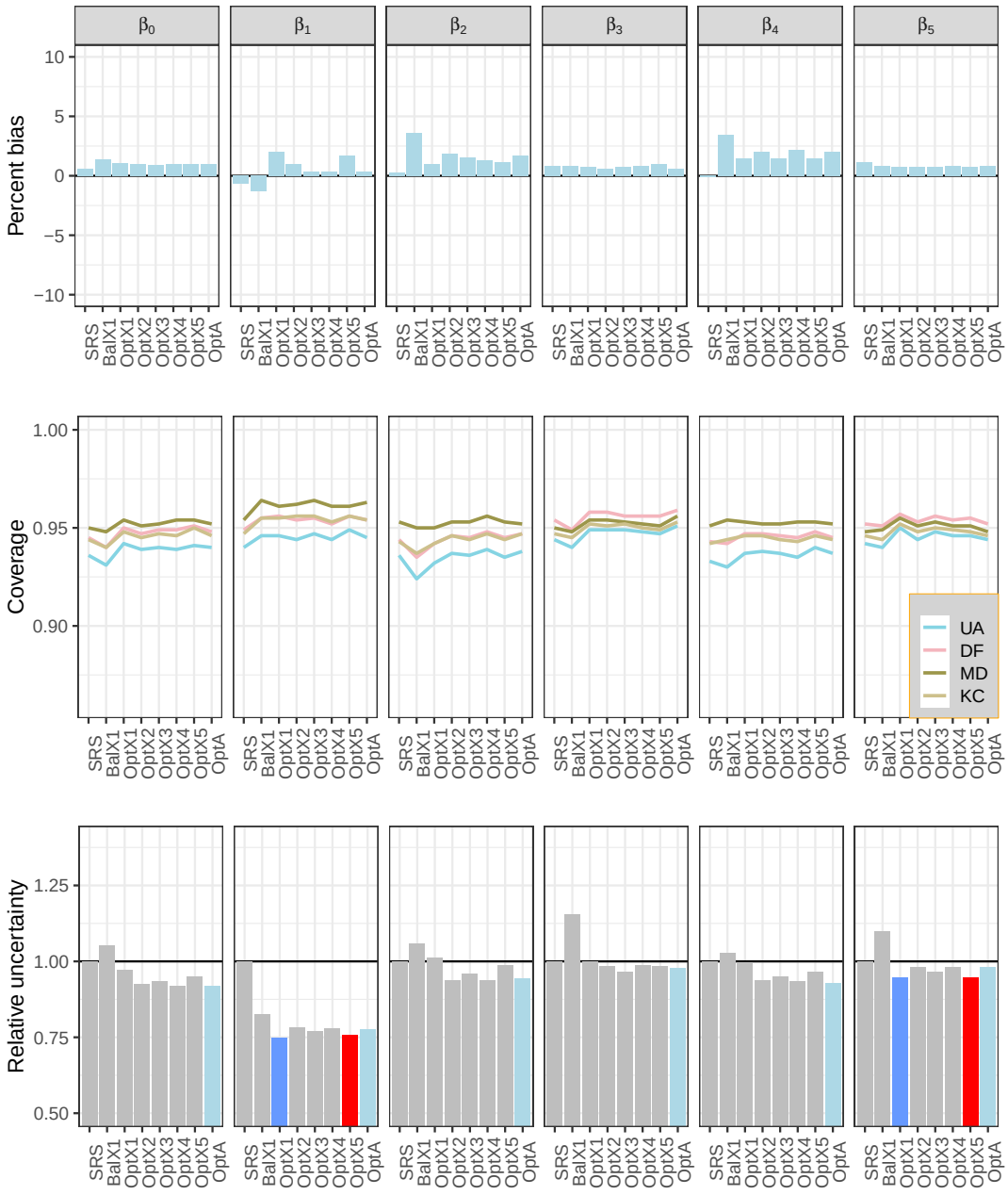

Figure B.29: Positive association X_1 and X_5: K=280, equal N_k=40, $K_s = 80$, $\sigma_V = 1$.

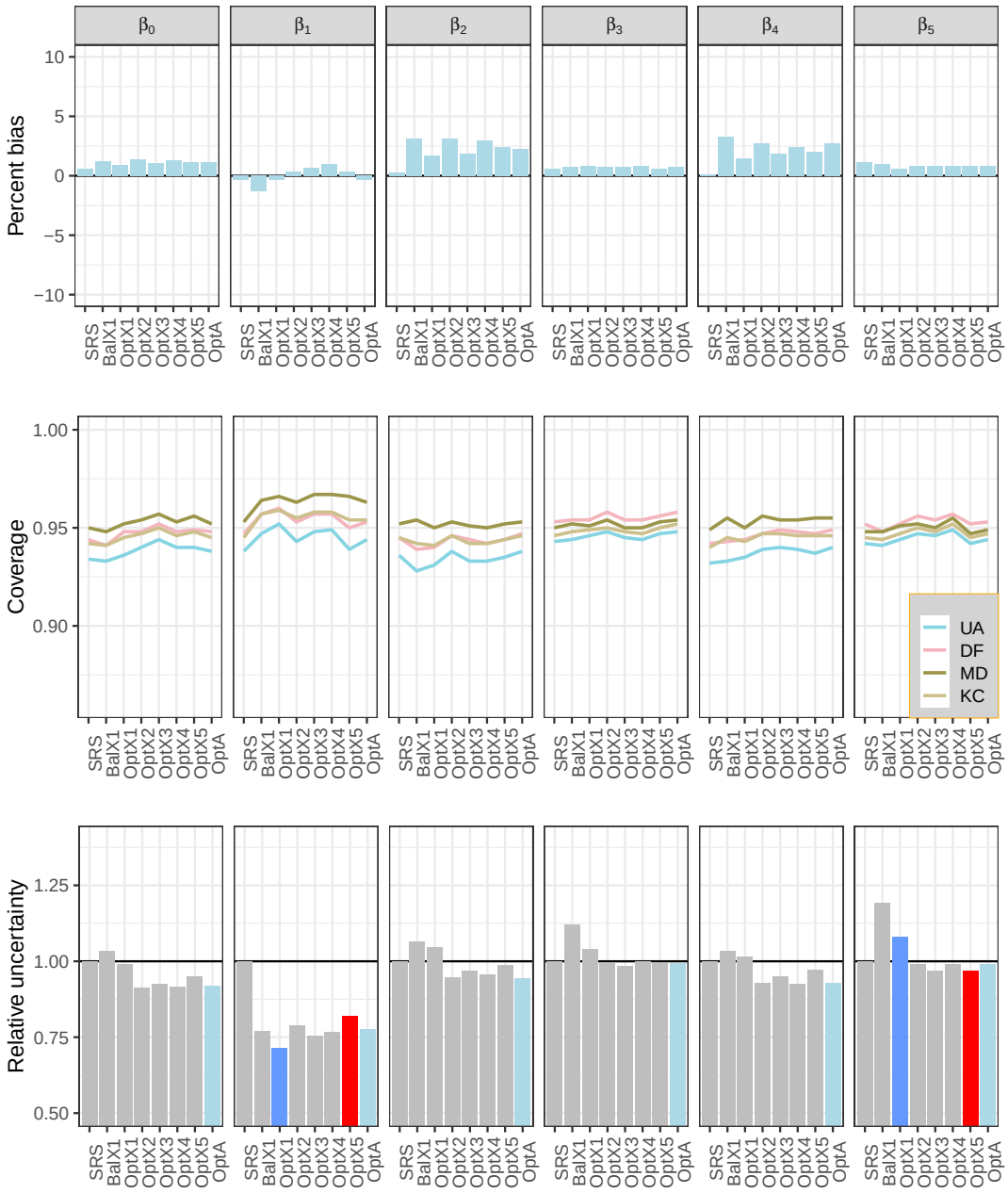

Figure B.30: Negative association X_1 and X_5: K=280, equal N_k=40, $K_s = 80, \sigma_V = 1$.

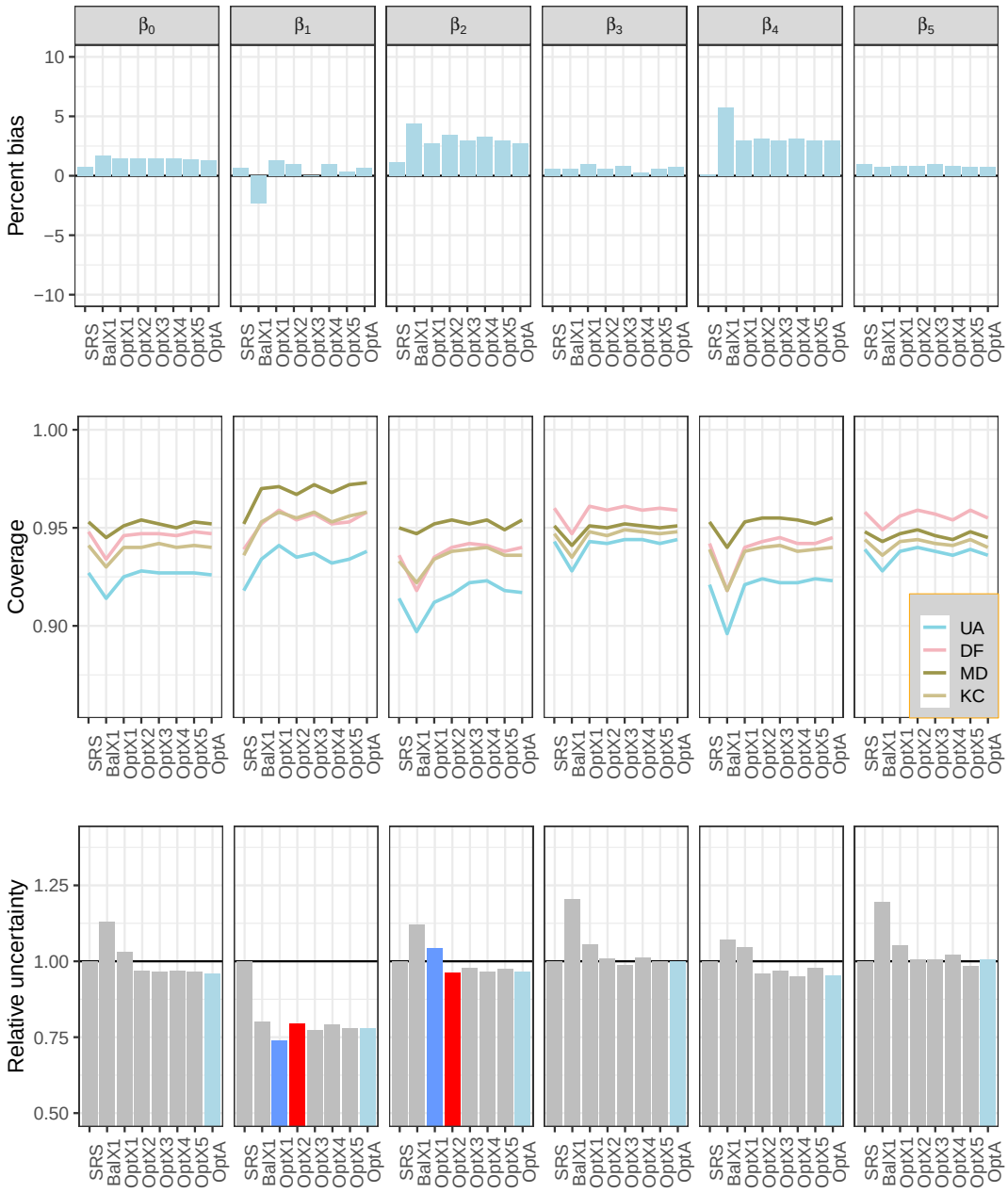

Figure B.31: Baseline scenario: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$.

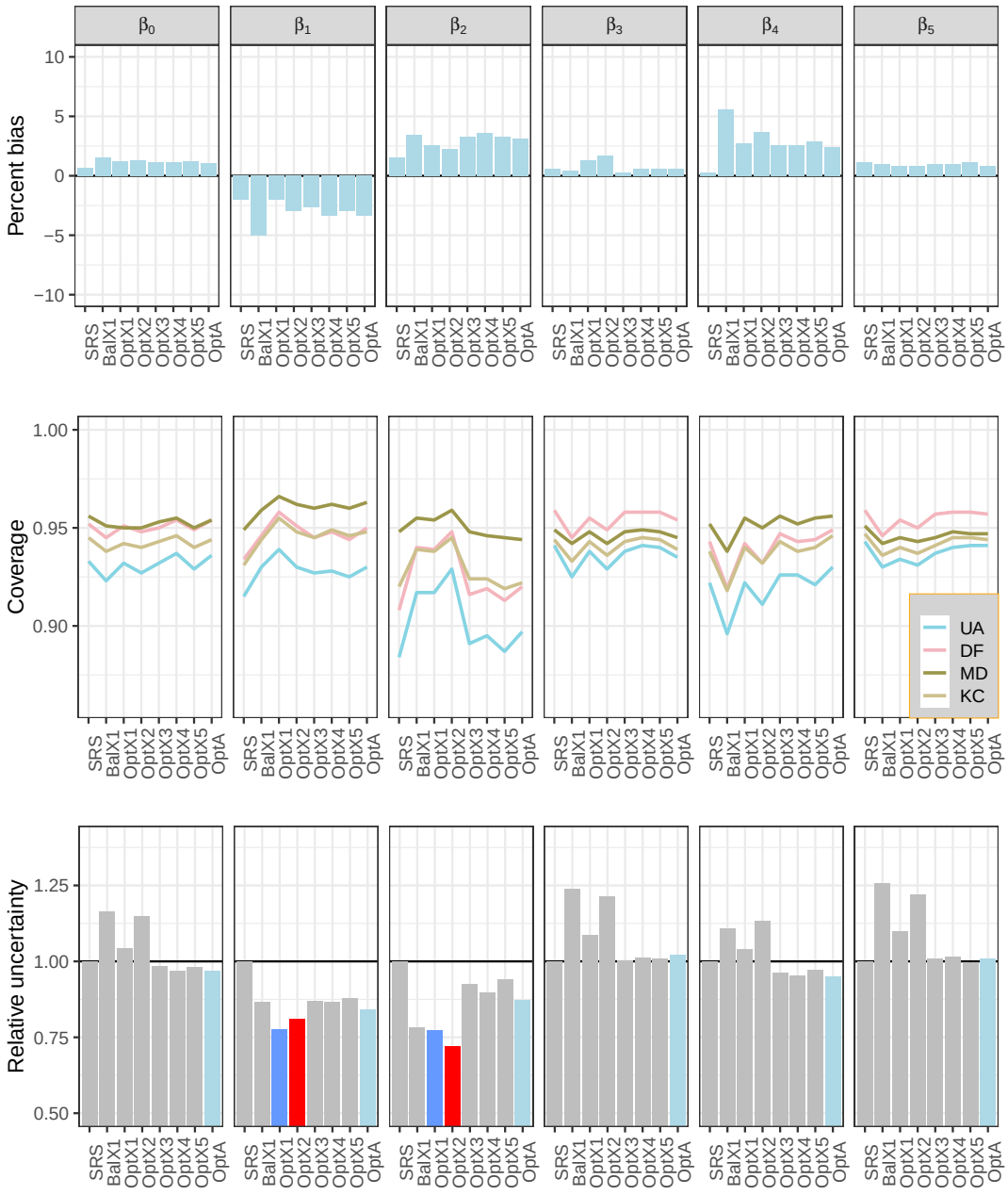

Figure B.32: Positive association X_1 and X_2: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$.

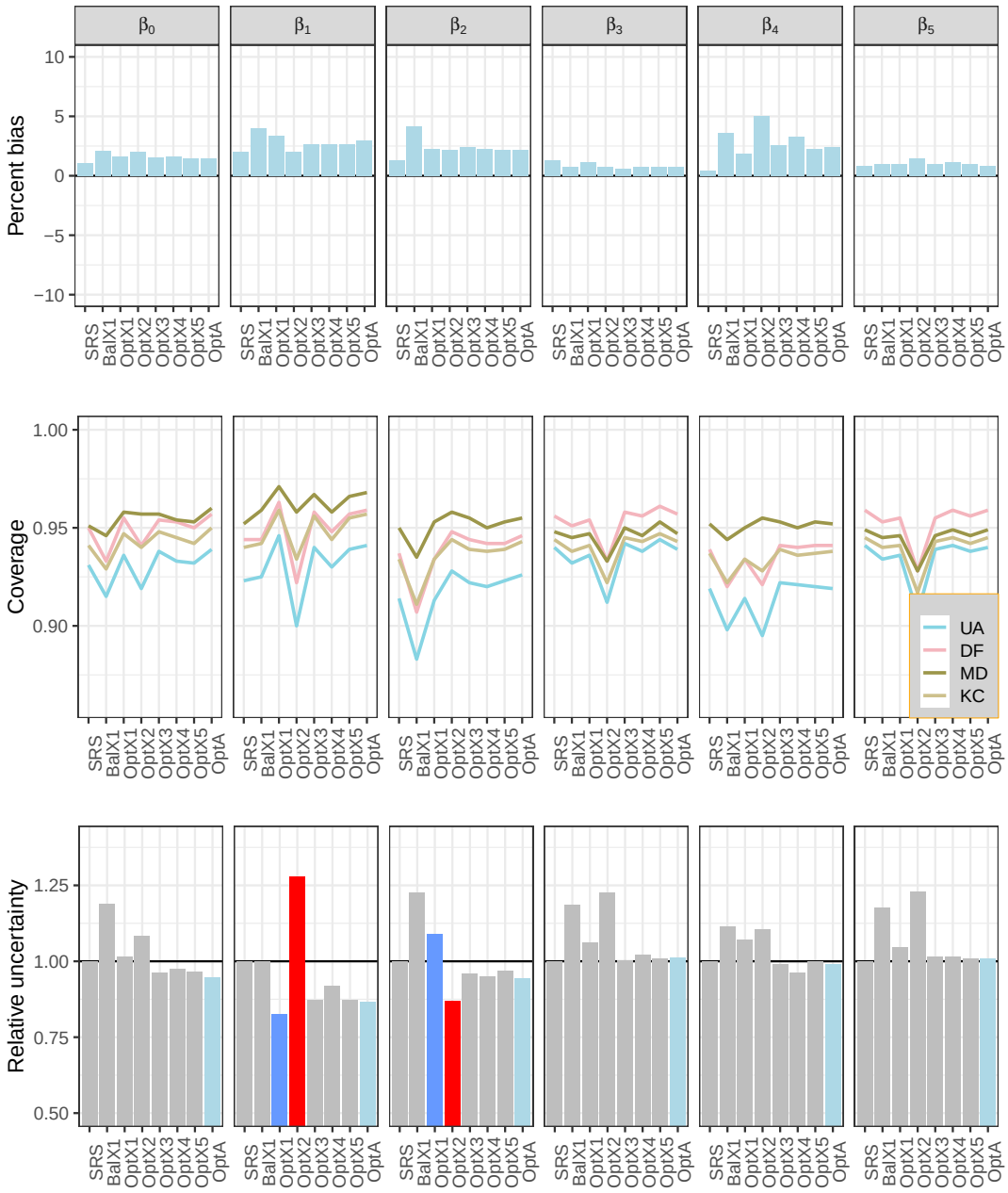

Figure B.33: Negative association X_1 and X_2: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$.

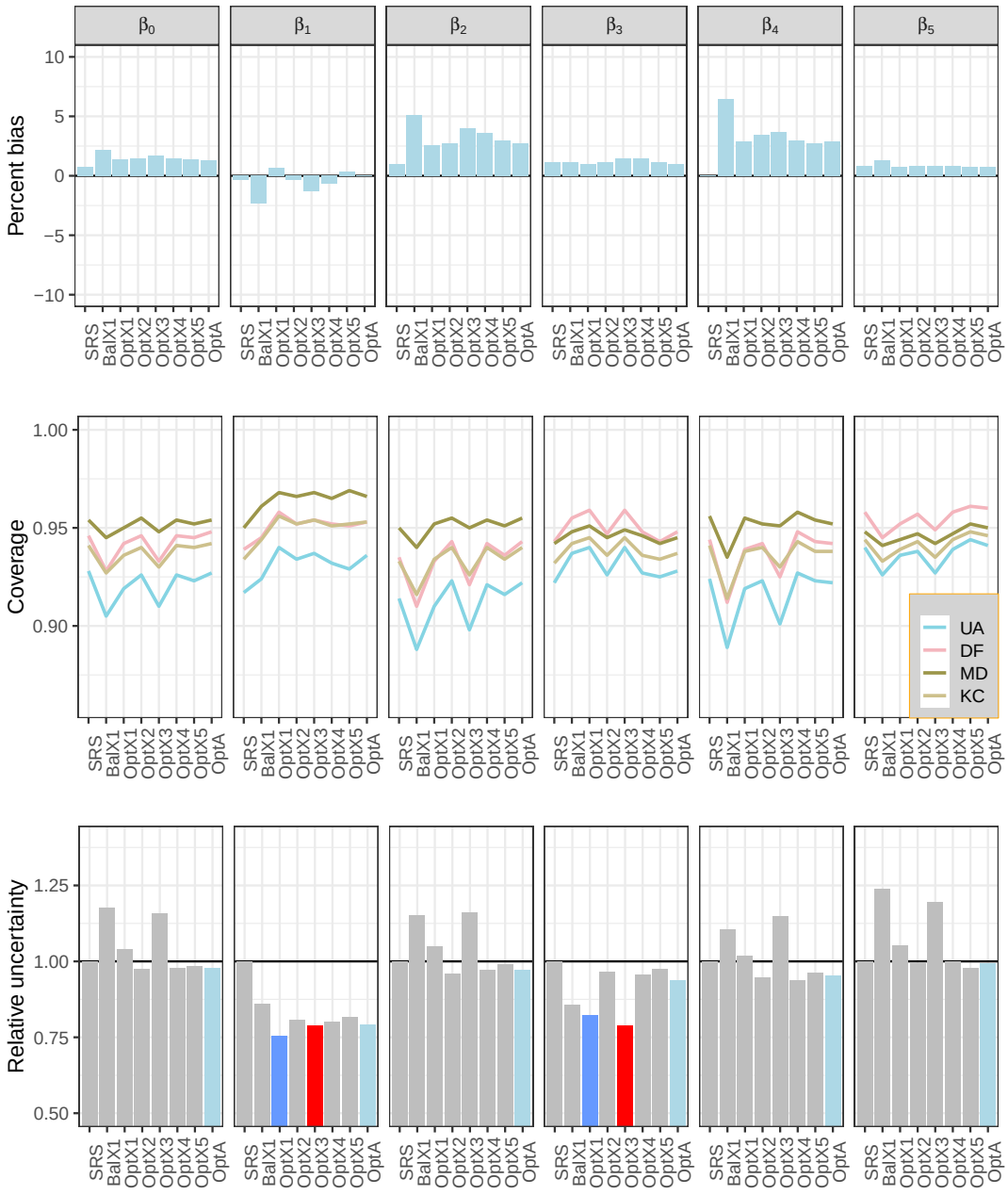

Figure B.34: Positive association X_1 and X_3: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$.

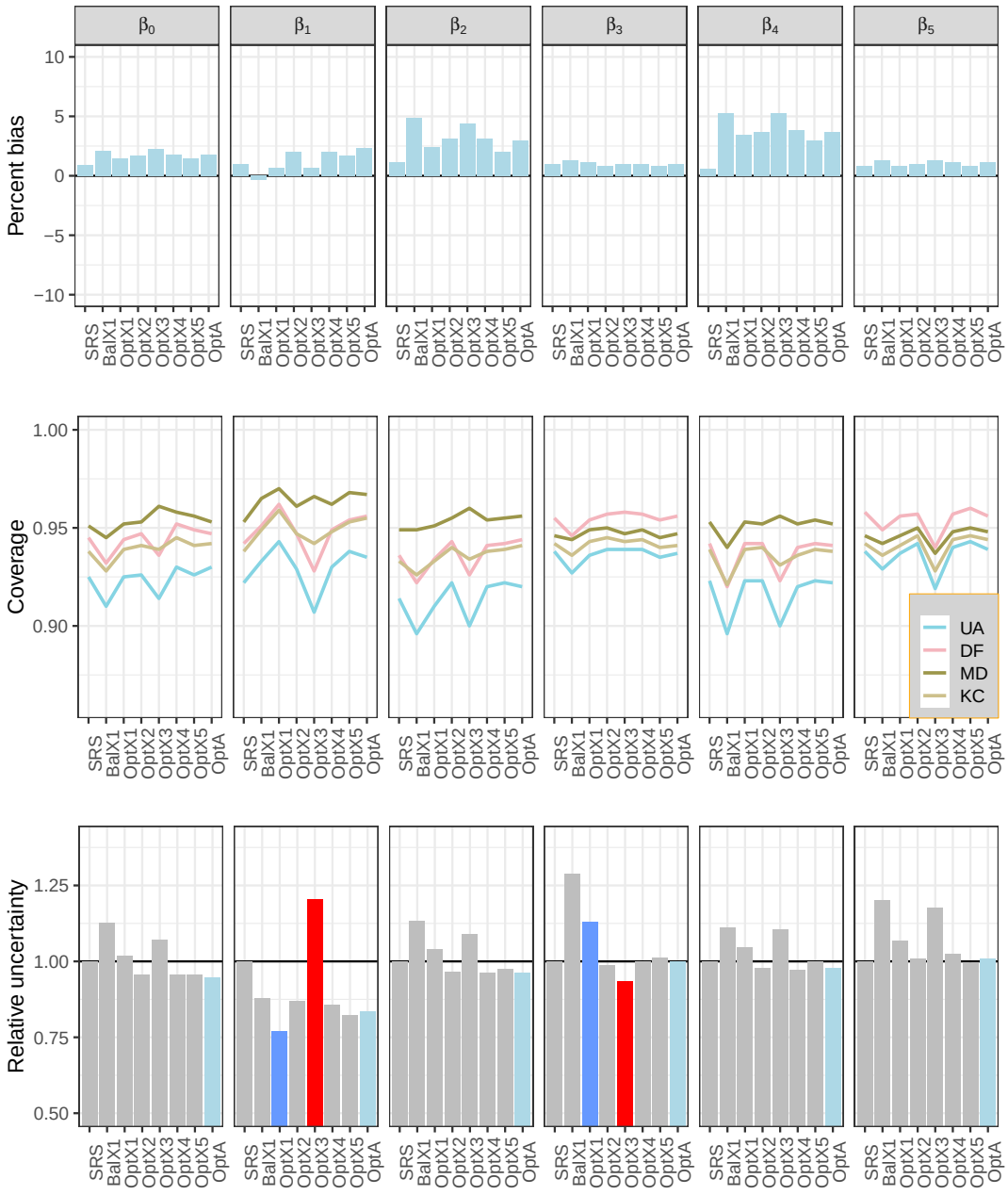

Figure B.35: Negative association X_1 and X_3: K=280, equal N_k=40, $K_s = 40, \sigma_V = 0.5$.

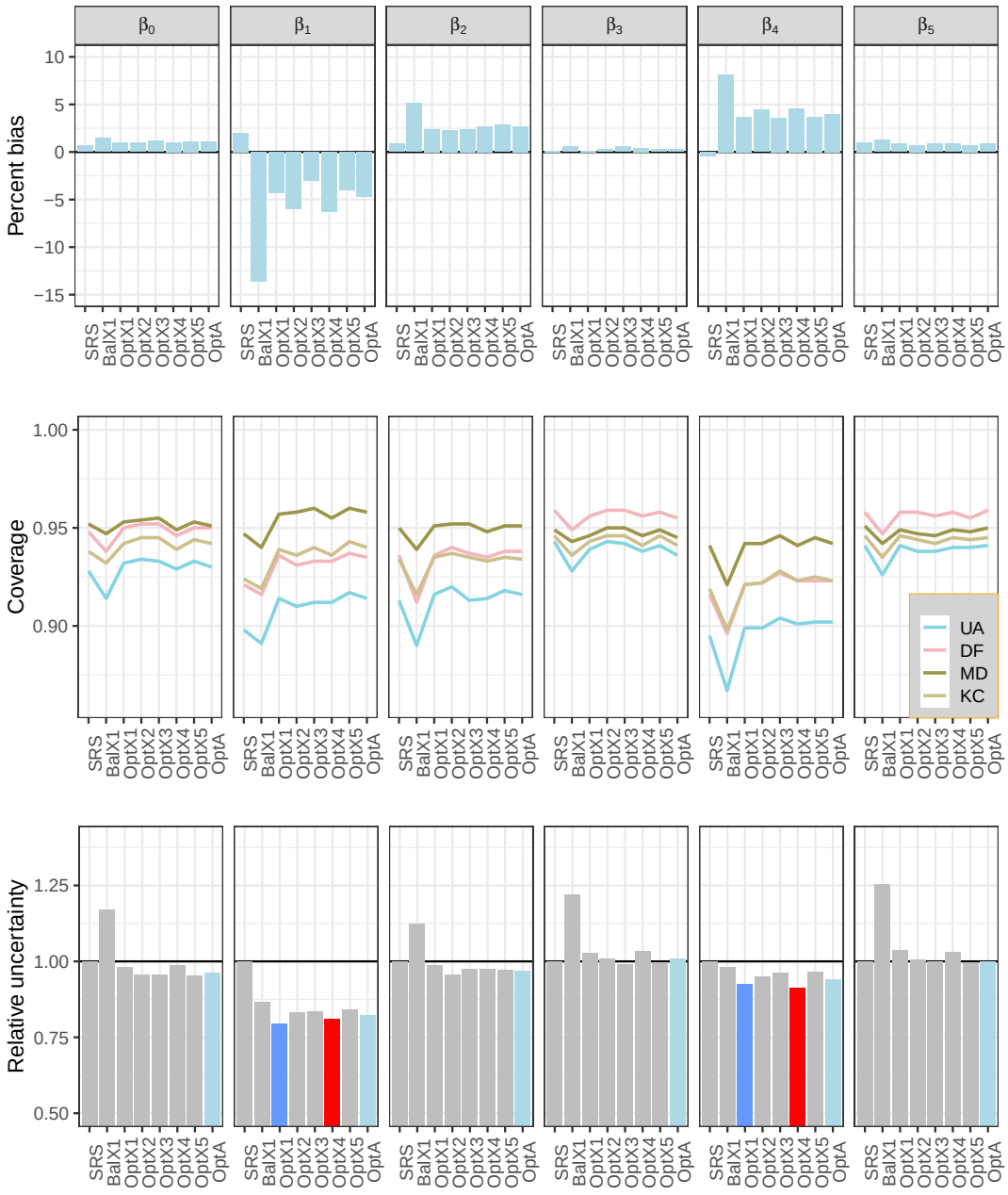

Figure B.36: Positive association X_1 and X_4: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$.

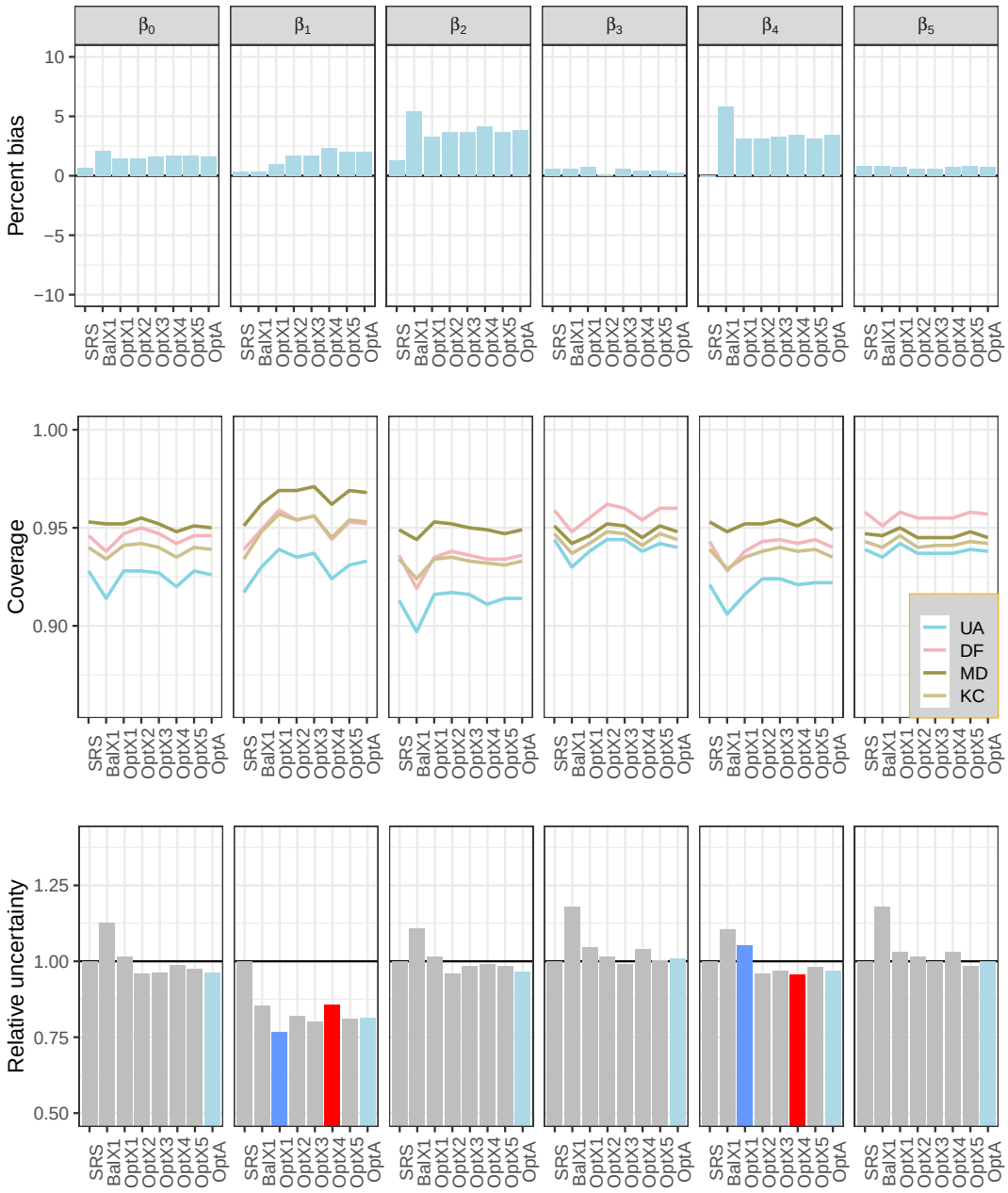

Figure B.37: Negative association X_1 and X_4: K=280, equal N_k=40, $K_s = 40, \sigma_V = 0.5$.

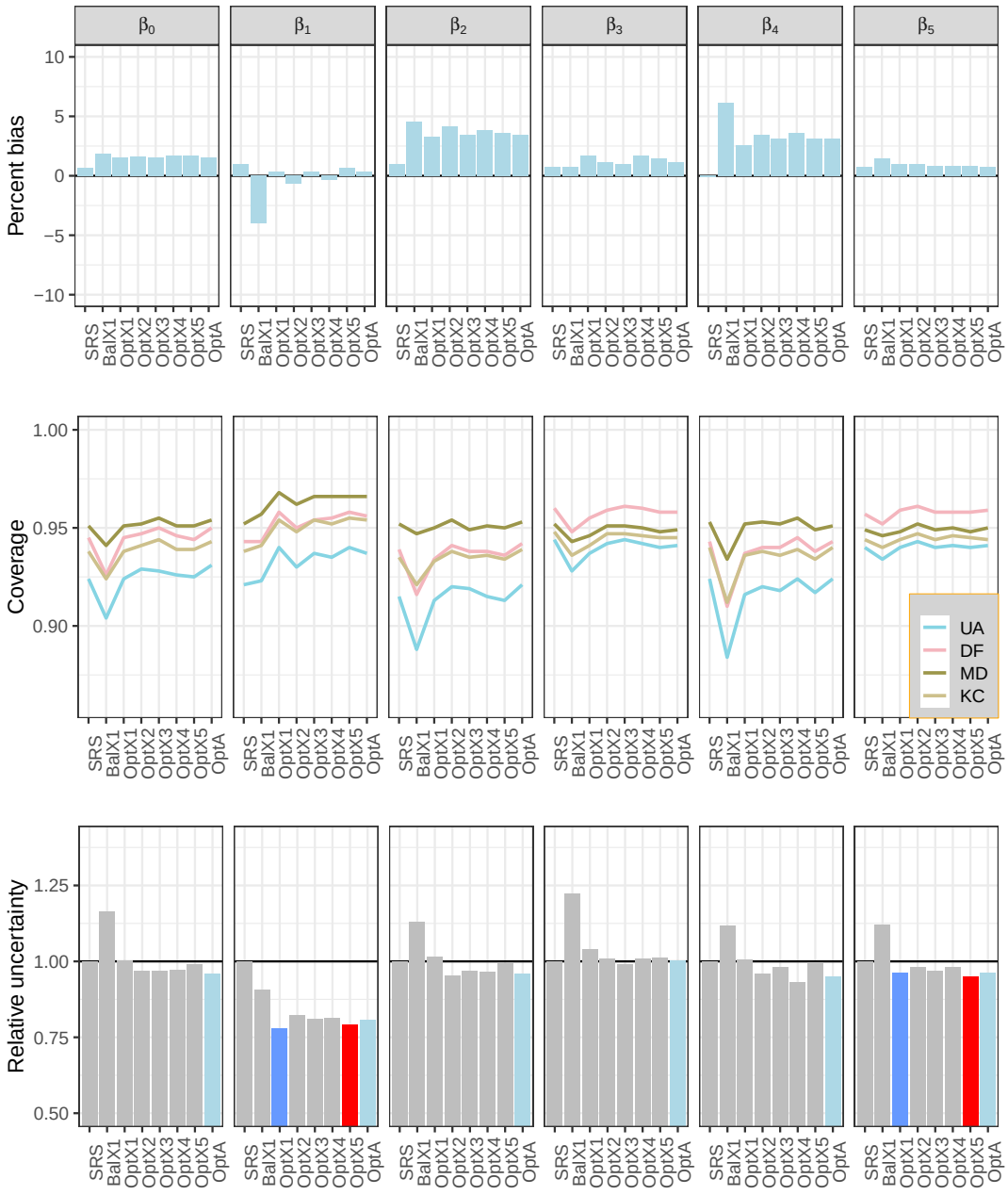

Figure B.38: Positive association X_1 and X_5: K=280, equal N_k=40, $K_s = 40$, $\sigma_V = 0.5$.

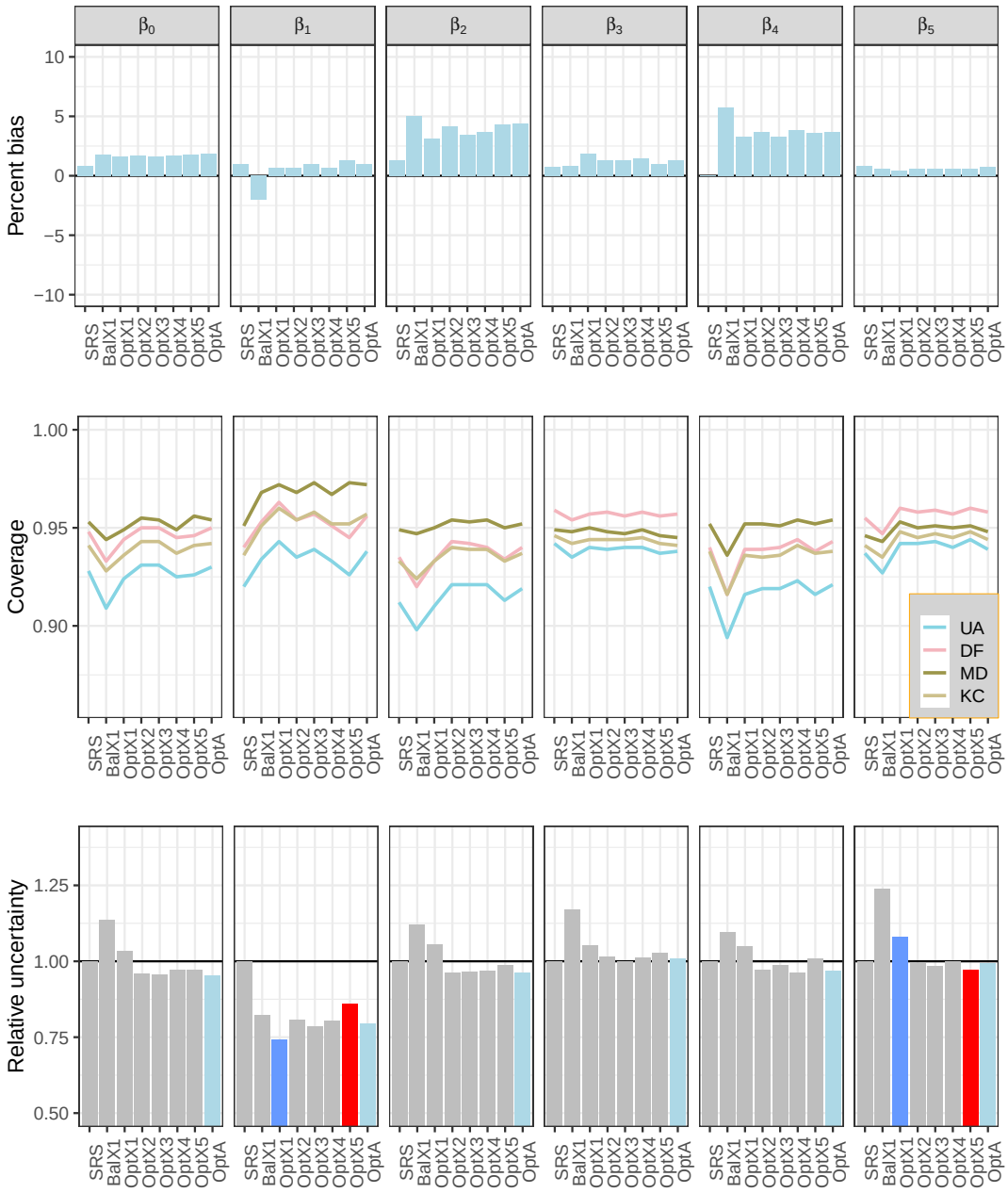

Figure B.39: Negative association X_1 and X_5: K=280, equal N_k=40, $K_s = 40, \sigma_V = 0.5$.

C

Supplementary material to accompany Chapter 3

C.1 Algorithm for Handling Edge Cases

Notation

- s_{max} is the maximum number of stages possible

- m_{sl} is the number of edge cases: either 1) $k_{j_2} < 0$ or 2) $k_{j_2} > K_j - k_{j_1}$
- τ is the threshold for the number of edge cases tolerated, m_{sl}
- K_{s_r} is the number of clusters that still need to be sampled after fixing the edge cases:

$$K_{s_r} = K_s - \sum_{j:k_{j_2}<0} k_{j_1} - \sum_{j:k_{j2}>K_j-k_{j1}} K_j$$

- NEC is the set of strata which did not yield edge cases

Algorithm

1. set $s = 1$ and sample K_{s_1} clusters
2. obtain $\widehat{\beta}_{s_1}$ by fitting the analysis model to Stage 1 data
3. compute the k_{j_2}
4. determine $m_{s_l} \in \{0, \ldots, J\}$
5. **while** $s < s_{max}$ and $m_{s_l} \neq 0$
   (a) **if** $m_{s_l} \geq \tau$ **or** $K_{s_r} > \sum_{j \in NEC} K_j$ (not enough clusters to achieve the desired sample size) **or** $K_{s_r} < \sum_{j \in NEC} k_{j_1}$ (k_{j_2} would necessarily be negative in order to end up with K_s clusters, which is not possible): sample more clusters at Stage 1
      i. set $s = s + 1$
      ii. sample $K_{s_{1inc}}$ more clusters
      iii. obtain updated $\widehat{\beta}_{s_1}$ by fitting the model to the data collected thus far
      iv. **if** s=s_{max}, set the k_{j_2} to 0 for all strata
      v. **else**, compute the updated k_{j_2} and update m_{s_l}
   (b) **else** set counter=0; **while** counter=0:

i. fix the existing edge cases (i.e. set $k_{j_2} = 0$ if $k_{j_2} < 0$, set $k_{j_2} = K_j - k_{j_1}$ if $k_{j_2} > K_j - k_{j_1}$)

ii. recalculate the k_{j_2} for the remaining strata using an updated constraint:

$$K_{s_r} = K_s - \sum_{j:k_{j_2}<0} k_{j_1} - \sum_{j:k_{j2}>K_j-k_{j1}} K_j$$

iii. determine the number of remaining cases: $m_{s_{l_1}}$

A. **if** $m_{s_{l_1}} > m_{s_l}$

- set $s = s + 1$
- sample $K_{s_{1inc}}$ more clusters
- obtain updated $\widehat{\beta}_{s_1}$ by fitting the model to the data collected thus far
- **if** $s = s_{max}$, set the k_{j_2} to 0 for all strata
- **else** compute the updated k_{j_2} and update m_{s_l}
- set counter=1, exiting while loop from (b) and going back to beginning of while loop in (a)

B. **else if** $m_{s_{l_1}}$=0

- set m_{s_l}=$m_{s_{l_1}}$
- set counter=1, exiting while loop from (b) and going to beginning of while loop in (a)

C. **else**

- update the indices
- set m_{s_l}=$m_{s_{l_1}}$ this then brings you back to beginning of while loop from (b), (i.e. fix edge cases again and recalculate the remaining k_{j_2})

References

[1] (2017). Using dhis 2 to strengthen health systems. https://www.measureevaluation.org/resources/publications/fs-17-212.

[2] (2018). Dhis2 factsheet. https://s3-eu-west-1.amazonaws.com/content.dhis2.org/general/dhis-factsheet.pdf.

[3] AbouZahr, C. & Boerma, T. (2005). Health information systems: the foundations of public health. *Bulletin of the World Health Organization*, 83, 578–583.

[4] Atkinson, A., Donev, A., Tobias, R., et al. (2007). *Optimum experimental designs, with SAS*, volume 34. Oxford University Press.

[5] Audigier, V., White, I. R., Jolani, S., Debray, T. P., Quartagno, M., Carpenter, J., Van Buuren, S., Resche-Rigon, M., et al. (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*, 33(2), 160–183.

[6] Bellhouse, D. R. (1984). A review of optimal designs in survey sampling. *Canadian Journal of Statistics*, 12(1), 53–65.

[7] Billingsley, P. (1995). Probability and measure. 1995. *John Wiley&Sons, New York*.

[8] Breidt, F. J., Opsomer, J. D., et al. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2), 190–205.

[9] Breslow, N. E. & Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4), 457–468.

[10] Breslow, N. E. & Day, N. E. (1980). *Statistical methods in cancer research. Vol. 1. The analysis of case-control studies.*, volume 1. Distributed for IARC by WHO, Geneva, Switzerland.

[11] Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., & Kulich, M. (2009). Improved horvitz–thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statistics in biosciences*, 1(1), 32–49.

[12] Cai, J., Qaqish, B., & Zhou, H. (2001). Marginal analysis for cluster-based case-control studies. *Sankhyā, Series B*, 63(3), 326–337.

[13] Cai, T. & Zheng, Y. (2013). Resampling procedures for making inference under nested case–control studies. *Journal of the American Statistical Association*, 108(504), 1532–1544.

[14] Chen, K., Hu, I., Ying, Z., et al. (1999). Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27(4), 1155–1163.

[15] Chen, T. & Lumley, T. (2020). Optimal multiwave sampling for regression modeling in two-phase designs. *Statistics in Medicine*, 39(30), 4912–4921.

[16] Fay, M. P. & Graubard, B. I. (2001). Small-sample adjustments for wald-type tests using sandwich estimators. *Biometrics*, 57(4), 1198–1206.

[17] Fedorov, V., Wu, Y., & Zhang, R. (2012). Optimal dose-finding designs with correlated continuous and discrete responses. *Statistics in medicine*, 31(3), 217–234.

[18] Fulcher, I., Hedt, K., Marealle, S., Tibaijuka, J., Abdalla, O., Hofmann, R., Layer, E., Mitchell, M., & Hedt-Gauthier, B. (2020a). Errors in estimated gestational ages reduce the likelihood of health facility deliveries: results from an observational cohort study in zanzibar. *BMC Health Services Research*, 20(1), 50.

[19] Fulcher, I. R., Nelson, A. R., Tibaijuka, J. I., Seif, S. S., Lilienfeld, S., Abdalla, O. A., Beckmann, N., Layer, E. H., Hedt-Gauthier, B., & Hofmann, R. L. (2020b). Improving health facility delivery rates in zanzibar, tanzania through a large-scale digital community health volunteer programme: a process evaluation. *Health Policy and Planning*.

[20] Fuller, W. A. (2011). *Sampling statistics*, volume 560. John Wiley & Sons.

[21] Han, K., Lumley, T., Shepherd, B. E., & Shaw, P. A. (2020). Two-phase analysis and study design for survival models with error-prone exposures. *Statistical Methods in Medical Research*, (pp. 0962280220978500).

[22] Haneuse, S., Hedt-Gauthier, B., Chimbwandira, F., Makombe, S., Tenthani, L., & Jahn, A. (2015). Strategies for monitoring and evaluation of resource-limited national antiretroviral therapy programs: the two-phase design. *BMC Medical Research Methodology*, 15(1), 31.

[23] Haneuse, S. & Rivera-Rodriguez, C. (2018). On the analysis of case–control studies in cluster-correlated data settings. *Epidemiology*, 29(1), 50–57.

[24] Haneuse, S., Schildcrout, J., & Gillen, D. (2012). A two-stage strategy to accommodate general patterns of confounding in the design of observational studies. *Biostatistics*, 13(2), 274–288.

[25] Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, 55(3), 688–698.

[26] Hug, L., Alexander, M., You, D., Alkema, L., & for Child, U. I.-a. G. (2019). National, regional, and global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: a systematic analysis. *The Lancet Global Health*, 7(6), e710–e720.

[27] John, R. S. & Draper, N. R. (1975). D-optimality for regression designs: a review. *Technometrics*, 17(1), 15–23.

[28] Jolani, S. (2018). Hierarchical imputation of systematically and sporadically missing data: An approximate bayesian approach using chained equations. *Biometrical Journal*, 60(2), 333–351.

[29] Jolani, S., Debray, T. P., Koffijberg, H., van Buuren, S., & Moons, K. G. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using mice. *Statistics in medicine*, 34(11), 1841–1863.

[30] Kauermann, G. & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *JASA*, 96(456), 1387–1396.

[31] Langholz, B. & Thomas, D. C. (1990). Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *AJE*, 131(1), 169–176.

[32] Li, Y. & Wang, Y.-G. (2008). Smooth bootstrap methods for analysis of longitudinal data. *Statistics in medicine*, 27(7), 937–953.

[33] Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.

[34] Lohr, S. L. (1990). Accurate multivariate estimation using triple sampling. *The Annals of Statistics*, (pp. 1615–1633).

[35] Lubchenco, L. O., Searls, D., & Brazie, J. (1972). Neonatal mortality rate: relationship to birth weight and gestational age. *The Journal of pediatrics*, 81(4), 814–822.

[36] Lunardon, N. & Scharfstein, D. (2017). Comment on 'small sample gee estimation of regression parameters for longitudinal data'. *Statistics in medicine*, 36(22), 3596–3600.

[37] MacKinnon, J. G. & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3), 305–325.

[38] Mancl, L. A. & DeRouen, T. A. (2001). A covariance estimator for gee with improved small-sample properties. *Biometrics*, 57(1), 126–134.

[39] Maokola, W., Willey, B., Shirima, K., Chemba, M., Armstrong Schellenberg, J., Mshinda, H., Alonso, P., Tanner, M., & Schellenberg, D. (2011). Enhancing the routine health information system in rural southern tanzania: successes, challenges and lessons learned. *Tropical medicine & international health*, 16(6), 721–730.

[40] McIsaac, M. A. & Cook, R. J. (2013). Response-dependent sampling with clustered and longitudinal data. In *ISS-2012 proceedings volume on longitudinal data analysis subject to measurement errors, missing values, and/or outliers* (pp. 157–181). Springer.

[41] McIsaac, M. A. & Cook, R. J. (2014). Response-dependent two-phase sampling designs for biomarker studies. *Canadian Journal of Statistics*, 42(2), 268–284.

[42] McIsaac, M. A. & Cook, R. J. (2015). Adaptive sampling in two-phase designs: a biomarker study for progression in arthritis. *Statistics in medicine*, 34(21), 2899–2912.

[43] Morel, J. G., Bokossa, M., & Neerchal, N. K. (2003). Small sample correction for the variance of gee estimators. *Biometrical Journal*, 45(4), 395–409.

[44] Neuhaus, J., Scott, A., & Wild, C. (2002). The analysis of retrospective family studies. *Biometrika*, 89(1), 23–37.

[45] Neuhaus, J. M., Scott, A. J., & Wild, C. (2006). Family-specific approaches to the analysis of case–control family data. *Biometrics*, 62(2), 488–494.

[46] Neuhaus, J. M., Scott, A. J., Wild, C. J., Jiang, Y., McCulloch, C. E., & Boylan, R. (2014). Likelihood-based analysis of longitudinal data from outcome-related sampling designs. *Biometrics*, 70(1), 44–52.

[47] Nisingizwe, M. P., Iyer, H. S., Gashayija, M., Hirschhorn, L. R., Amoroso, C., Wilson, R., Rubyutsa, E., Gaju, E., Basinga, P., Muhire, A., et al. (2014). Toward utilization of data for program management and evaluation: quality assessment of five years of health management information system data in rwanda. *Global health action*, 7(1), 25829.

[48] Nshimyiryo, A., Kirk, C. M., Sauer, S. M., Ntawuyirusha, E., Muhire, A., Sayinzoga, F., & Hedt-Gauthier, B. (2020). Health management information system (hmis) data verification: A case study in four districts in rwanda. *PloS one*, 15(7), e0235823.

[49] Nyamtema, A. S. (2010). Bridging the gaps in the health management information system in the context of a changing health sector. *BMC medical informatics and decision making*, 10(1), 36.

[50] Pan, W. & Wall, M. M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in medicine*, 21(10), 1429–1441.

[51] Paul, S. & Zhang, X. (2014). Small sample gee estimation of regression parameters for longitudinal data. *Statistics in medicine*, 33(22), 3869–3881.

[52] Pepe, M. & Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics*, 23(4), 939–951.

[53] Prentice, R. L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3), 403–411.

[54] Qaqish, B. F., Zhou, H., & Cai, J. (1997). On case-control sampling of clustered data. *Biometrika*, 84(4), 983–986.

[55] Reilly, M. (1996). Optimal sampling strategies for two-stage studies. *American journal of epidemiology*, 143(1), 92–100.

[56] Resche-Rigon, M., White, I. R., Bartlett, J. W., Peters, S. A., Thompson, S. G., & Group, P.-I. S. (2013). Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in medicine*, 32(28), 4890–4905.

[57] Rivera-Rodriguez, C., Spiegelman, D., & Haneuse, S. (2019). On the analysis of two-phase designs in cluster-correlated data settings. *Statistics in medicine*, 38(23), 4611–4624.

[58] Rwanda DHS (2015). National Institute of Statistics of Rwanda (NISR) [Rwanda], Ministry of Health (MOH) [Rwanda], and ICF International. Rwanda Demographic and Health Survey 2014-15.

[59] Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

[60] Sauer, S., Hedt-Gauthier, B., Rivera-Rodriguez, C., & Haneuse, S. (2021). Small-sample inference for cluster-based outcome-dependent sampling schemes in resource-limited settings: Investigating low birthweight in rwanda. *Biometrics*.

[61] Schildcrout, J. S., Garbett, S. P., & Heagerty, P. J. (2013). Outcome vector dependent sampling with longitudinal continuous response data: stratified sampling based on summary statistics. *Biometrics*, 69(2), 405–416.

[62] Schildcrout, J. S., Haneuse, S., Tao, R., Zelnick, L. R., Schisterman, E. F., Garbett, S. P., Mercaldo, N. D., Rathouz, P. J., & Heagerty, P. J. (2020). Two-phase, generalized case-control designs for the study of quantitative longitudinal outcomes. *American Journal of Epidemiology*, 189(2), 81–90.

[63] Schildcrout, J. S. & Heagerty, P. J. (2008). On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics*, 9(4), 735–749.

[64] Schildcrout, J. S. & Heagerty, P. J. (2011). Outcome-dependent sampling from existing cohorts with longitudinal binary response data: Study planning and analysis. *Biometrics*, 67(4), 1583–1593.

[65] Schildcrout, J. S. & Rathouz, P. J. (2010). Longitudinal studies of binary response data following case–control and stratified case–control sampling: Design and analysis. *Biometrics*, 66(2), 365–373.

[66] Schildcrout, J. S., Rathouz, P. J., Zelnick, L. R., Garbett, S. P., & Heagerty, P. J. (2015). Biased sampling designs to improve research efficiency: factors influencing pulmonary function over time in children with asthma. *The annals of applied statistics*, 9(2), 731.

[67] Schildcrout, J. S., Schisterman, E. F., Mercaldo, N. D., Rathouz, P. J., & Heagerty, P. J. (2018). Extending the case–control design to longitudinal data: stratified sampling based on repeated binary outcomes. *Epidemiology (Cambridge, Mass.)*, 29(1), 67.

[68] Seaman, S. R. & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3), 278–295.

[69] Tao, R., Mercaldo, N. D., Haneuse, S., Maronge, J. M., Rathouz, P. J., Heagerty, P. J., & Schildcrout, J. S. (2021). Two-wave two-phase outcome-dependent sampling designs, with applications to longitudinal binary data. *Statistics in Medicine*.

[70] Tao, R., Zeng, D., & Lin, D.-Y. (2020). Optimal designs of two-phase studies. *Journal of the American Statistical Association*, 115(532), 1946–1959.

[71] Tong, C. & Thomas, L. (2020). Optimal multi-wave sampling for regression modelling in two-phase designs.

[72] Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3), 219–242.

[73] Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

[74] Wacholder, S. (1991). Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology*, (pp. 155–158).

[75] White, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *AJE*, 115(1), 119–128.

[76] WHO (2019). Newborns: Reducing mortality. https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality. Accessed: 2019-09-30.

[77] Wickremasinghe, D., Hashmi, I. E., Schellenberg, J., & Avan, B. I. (2016). District decision-making for health in low-income settings: a systematic literature review. *Health Policy and Planning*, 31(suppl_2), ii12–ii24.

[78] Wittes, J. & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in medicine*, 9(1-2), 65–72.

[79] Woods, D. C. & Van de Ven, P. (2011). Blocked designs for experiments with correlated non-normal response. *Technometrics*, 53(2), 173–182.

[80] Wright, T. (2017). Exact optimal sample allocation: More efficient than neyman. *Statistics & Probability Letters*, 129, 50–57.

[81] Xie, M. & Yang, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *The Annals of Statistics*, 31(1), 310–347.

[82] Zaslavsky, A. M., Zheng, H., & Adams, J. (2008). Optimal sample allocation for design-consistent regression in a cancer services survey when design variables are known for aggregates. *Survey methodology*, 34(1), 65.

[83] Zhong, Y. & Cook, R. J. (2021). Selection models for efficient two-phase design of family studies. *Statistics in Medicine*, 40(2), 254–270.