FCSM 2018: Bayesian State-level Estimates of Diabetes Prevalence in United States, 2006 – 2015

Diba Khan
Rong Wei
Yulei He
Hee-Choon Shin
Donald J. Malec
Centers for Disease Control and Prevention
National Center for Health Statistics
Division of Research and Methodology
Hyattsville, MD.

Objective

To explore geographic variation in diabetes prevalence at the state level for 4 domains from the National Health Interview Survey (NHIS).

Outline

1. **Data**
2. **Design based estimates**
3. **Within states covariances**
4. **Covariates**
5. **Models**
6. **Model Fit**
7. **Results**
8. **Simulations**
9. **Conclusions**

Pooling and Weighting

1. Due to the small sample sizes at the state level, the corresponding sample variance is too unstable
2. The data are pooled for the years $2006 - 2015$
   Note that even if data are pooled, the number of PSUs per state remain constant. A traditional variance of the S^2 form may still be unstable.
3. The data are weighted to adjust to population totals. Interim sampling weights are used in this analysis
4. The value of the interim sampling weight is roughly the number of people a given sample adult or sample child represents
5. Due to the changes in NHIS sample size between 2006 and 2015 the Interim sampling weights need to be adjusted for the pooling across all 10 years

Study population from NHIS (2006-2015)

Eligibles	Diabetics	Weighted (Eligibles)	Weighted (Diabetics)
296654	29426	1.6612E9	1.5646E8

Table 1: Pooled data from 2006 – 2015.

1. Weighted percent estimate 0.0942

4 age and gender based domains

No.	Domain
1	Males aged 18-64 years
2	Males aged 65 years and older
3	Females aged 18-64 years
4	Females aged 65 years and older

4 domains

Age	Gender
2	2

NHIS design based estimates using a SAS survey procedure

1. **proc surveylogistic**
2. **stratum STRATA**
3. **cluster PSU**
4. **by STATE**
5. **class DOMAIN**
6. **model DIABETIC(event='1') = DOMAIN/noint covb**
7. **weight** w_{ik}
8. w_{ik} **is the final survey weight adjusted by proportions for ith year and kth individual from 2006-2015**

Estimates from SAS/SURVEYLOGISTIC

1. 204 diabetes rates: 51 states/DC x 4 domains
2. 4x4 variance/covariance matrices for 4 domains within 51 states
3. US level variance/covariance matrix $\hat{\Omega}$

	Domain1	Domain2	Domain3	Domain4
Domain1	0.000275	5.52E-06	0.000046	0.000023
Domain2	5.52E-06	0.000365	1.53E-05	2.76E-05
Domain3	0.000046	1.53E-05	0.000277	2.44E-05
Domain4	0.000023	2.76E-05	2.44E-05	0.00035

4. US national level design effects for four domains
5. DF (degrees of freedom) for each of 51 states/DC

Adjustment on variance/covariance matrices for small states

1. Let $D_i = (\hat{\Sigma}_i[1,1], \hat{\Sigma}_i[2,2], \hat{\Sigma}_i[3,3], \hat{\Sigma}_i[4,4])$ as a 4×1 is a vector of the diagonal elements of $\hat{\Sigma}_i$ for state i, where $i = 1, ..., 51$
2. Let $U = (\hat{\Omega}[1,1], \hat{\Omega}[2,2], \hat{\Omega}[3,3], \hat{\Omega}[4,4])$ as a 4×1 is a vector of the diagonal elements of $\hat{\Omega}$
3. Let T be a 4×4 diagonal matrix whose elements are $sqrt(D_i/U)$. Thus the covariance matrix can be estimated as: $\hat{\Sigma}_i^* = T * \hat{\Omega} * T$, where, we pre- and post-multiply T with $\hat{\Omega}$ to obtain $\hat{\Sigma}_i^*$
4. $\hat{\Sigma}_i^*$ maintains the variance estimates for the corresponding state and four domains
5. $\hat{\Sigma}_i^*$ uses the correlation estimates from the national level covariance estimates matrix for the diabetes prevalence
6. The adjusted covariance matrix estimates $\hat{\Sigma}_i^*$ were used in the model estimation

Behavioral Risk Factor Surveillance System (BRFSS) diabetes prevalence estimates as covariates

1. Diabetes data from years 2006-2015 were downloaded from BRFSS website
2. Data were pulled and sample weights were adjusted across 10 years
3. 204 diabetes rates (51 states x 4 domains) were computed using SAS/SurveyFreq by incorporating BRFSS sampling design variables (stratum and PSU)

Modeling Weighted Diabetes Data

1. Hierarchical Bayes models
2. Auxiliary variables from Behavior Risk Factor Surveillance System (BRFSS) enhance the predictions (Raghunathan et.al Journal of the American Statistical Association June 2007, Vol. 102, No. 478.)
3. Additional covariates such as obesity, poverty, income and others were obtained from United States Department of Agriculture (USDA) and Area Resource File (ARF) datasets
4. Hierarchical Bayes estimates borrow strength over space (states) to produce SAEs

Problem formulation

1. $\hat{Y}_i$: diabetes prevalence estimate in state i on a logit scale where, $\hat{Y}_i = (\hat{y_{i1}}, \hat{y_{i2}}, \hat{y_{i3}}, \hat{y_{i4}})$, for $i = 1, \ldots, 51$ states and 4 domains
2. The multivariate FH model can be defined as:
   $\hat{Y}_i \sim MVN(\mu_i, \hat{\sum}_i)$, where,
   $\mu_i = (\mu_{i1}, \mu_{i2}, \mu_{i3}, \mu_{i4})$ denotes unknown diabetic prevalence that is to be estimated for state i and 4 domains
   $\hat{\sum}_i$ is the estimated design-based estimate of variance-covariance matrix of μ_i for state i and 4 domains on a logit scale

Proposed models

Model 1: Intercepts with random effects

$\mu_{ij} = \alpha_{0j} + v_{ij}$

1. α_{0j}: an intercept for each domain which is assigned a vague distribution: flat prior, $\alpha_0 \sim dflat()$
2. v_{ij}: random effects by state i (i=1,...,51) and domain j (j=1,2,3,4). Specifically, if $v_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4})$, then $v_i \sim MVN(\mathbf{0}, \boldsymbol{R})$
   Inverse of $\boldsymbol{R}$ is assigned a Wishart prior with I as the scale matrix and degrees of freedom = 4

$$\text{Inverse of } \boldsymbol{R} \sim Wish\left(\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, 4\right)$$

Proposed models

Model 2: Model 1 + BRFSS covariates

1. $\mu_{ij} = \alpha_{0j} + v_{ij} + X_{ij}{}'\boldsymbol{\beta_j}$

2. α_{0j}: Intercept for each domain assigned vague distribution: flat prior
   X_{ij} : is the i th row of the covariates matrix
3. v_{ij}: random effects by state i (i=1,...,51) and domain j (j=1,2,3,4). Specifically, if $v_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4})$, then $v_i \sim MVN(\mathbf{0}, \boldsymbol{R})$
4. Inverse of $\boldsymbol{R}$ is assigned a Wishart prior with I as the scale matrix and degrees of freedom = 4
5. Inverse of $\boldsymbol{R} \sim Wish\left(\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, 4\right)$
6. $\boldsymbol{\beta_j}$: vector of regression parameters. Normally distributed with uniform priors U(0,100) on the standard deviation

Proposed models

Model 3: Model 1 + ARF and USDA covariates

1. $\mu_{ij} = \alpha_{0j} + v_{ij} + Z_i{}'\gamma$

2. α_{0j}: Intercept for each domain assigned vague distribution: flat prior
3. v_{ij}: random effects by state i (i=1,...,51) and domain j (j=1,2,3,4). Specifically, if $v_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4})$, then $v_i \sim MVN(\mathbf{0}, \boldsymbol{R})$
   Inverse of $\boldsymbol{R}$ is assigned a Wishart prior with I as the scale matrix and degrees of freedom = 4
   Inverse of $\boldsymbol{R} \sim Wish\left(\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, 4\right)$
4. Z_i : is the i th row of the covariates matrix
5. γ: vector of regression parameters
   γ: assigned vague distribution: flat prior

Proposed models

Model 4

1. In model 4, the effect of ignoring the correlations between the different domains was investigated
2. The off diagonal elements of the matrices $\hat{\sum}_i$, the design-based estimate of variance-covariance matrix of $\hat{\mu}_i$ for state i and 4 domains were set to zero. Specifically, $\breve{\sum}_i = diag(\hat{\sum}_i)$. Hence, model 4 is defined as: $\mu_{ij} = \alpha_{0j} + v_{ij} + X_{ij}'\beta$, where, α_{0j}, v_{ij}, and $X_{ij}'\beta_{\boldsymbol{j}}$, are defined above as in model 2

Proposed models

Model 5

1. In model 5, the advantage of using ARF and USDA covariates and BRFSS diabetes prevalence estimates together in one model is further investigated
2. Model 5 is defined as:
   $\mu_{ij} = \alpha_{0j} + v_{ij} + X_{ij}'\boldsymbol{\beta_j} + Z_i'\boldsymbol{\gamma}$, where,
   α_{0j}, v_{ij}, $X_{ij}'\boldsymbol{\beta_j}$, and $Z_i'\boldsymbol{\gamma}$ are defined above as in model 2 and model 3

Assessment on the fit of proposed models

Table 2: DIC for the models.

Model	Dbar	Dhat	DIC	pD
Model 1	-299.4	-448.1	-150.7	148.7
Model 2	**-306.8**	**-448.9**	**-164.7**	**142.1**
Model 3	-279.5	-482.4	-76.59	202.9
Model 4	-258.7	-398.5	-118.9	139.8
Model 5	-279.8	-482.3	-77.22	202.5

Model results

Table 3: Model 2 results.

parameter	mean	sd	val2.5	median	val97.5
a0[1]	-3.424	0.02191	-4.066	-3.611	-0.0828
a0[2]	-2.149	0.01463	-3.059	-2.196	-0.1159
a0[3]	-3.462	0.01877	-4.063	-3.583	-0.08297
a0[4]	-2.656	0.01399	-3.297	-2.727	-0.315
β[1]	13.23	0.1955	-13.01	14.63	20.96
β[2]	3.867	0.04801	-0.06663	3.894	7.627
β[3]	13.28	0.1271	-0.8479	13.75	20.95
β**[4]**	6.489	0.09244	**1.685**	6.606	**9.635**

Males in the agegroup 18-64

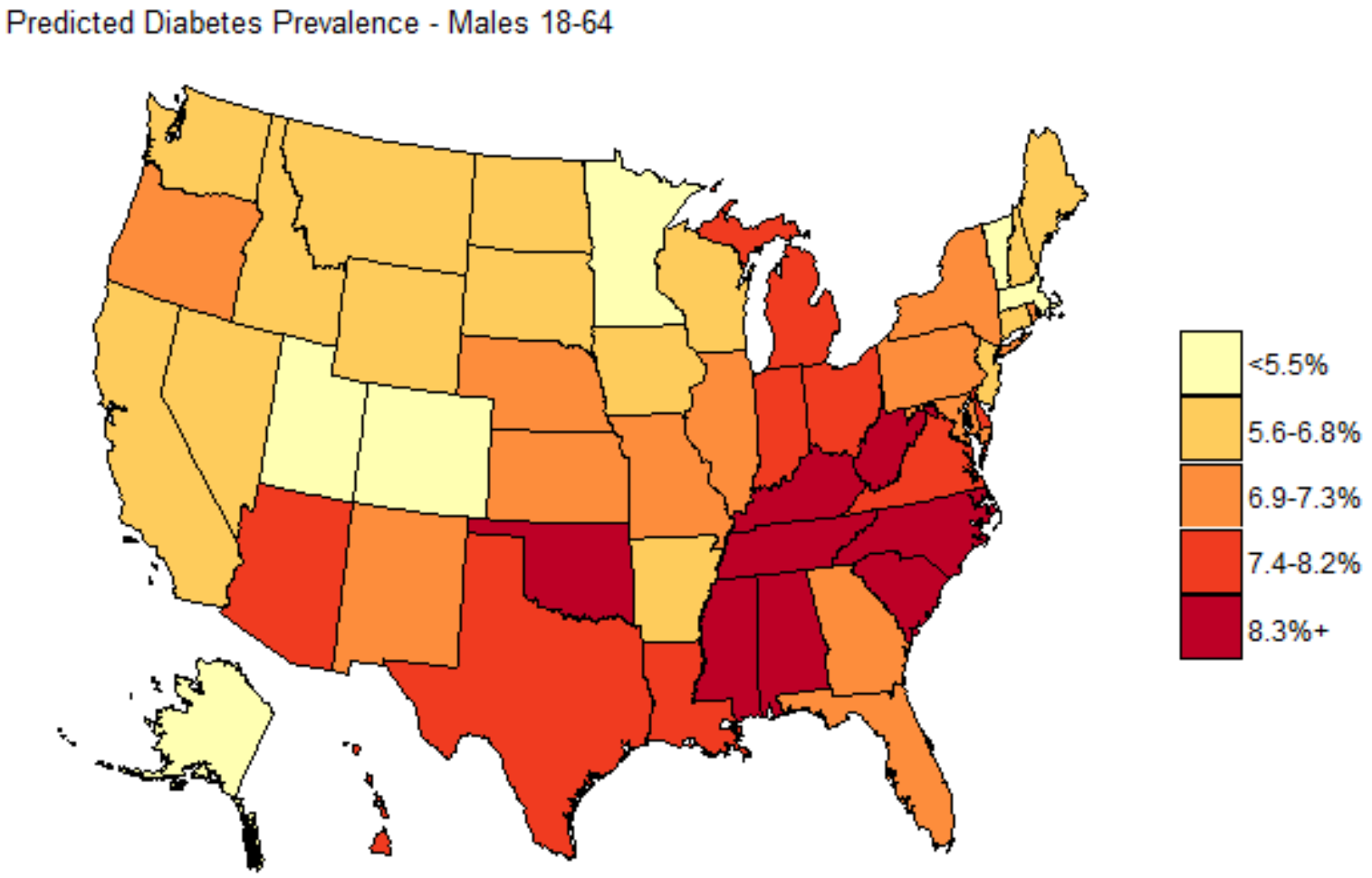


Figure 1: Estimates of Diabetics for domain 1: Males 18-64.

Males above the age 64

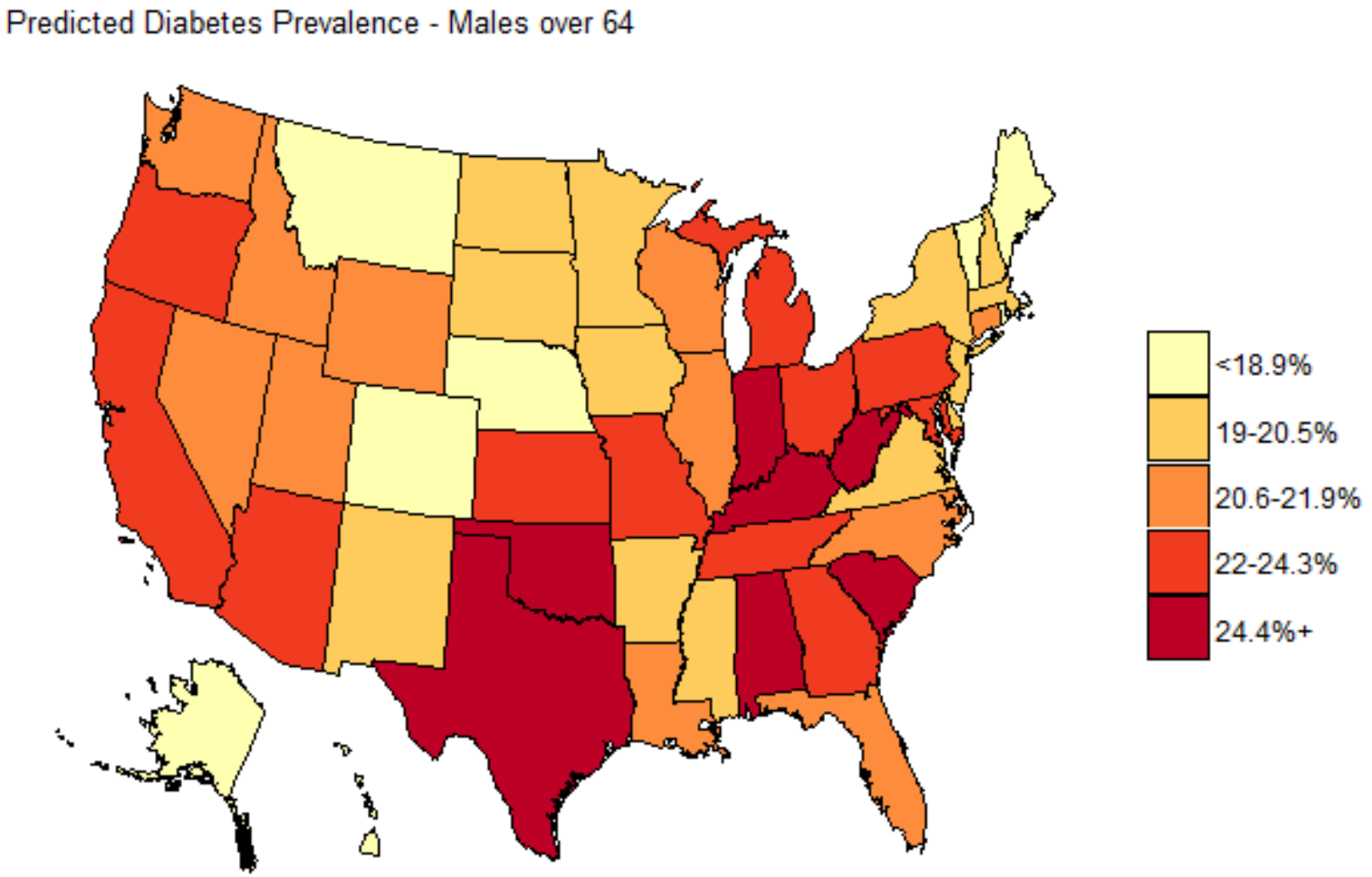


Figure 2: Estimates of Diabetics for domain 2: Males above 64.

Females in the agegroup 18-64

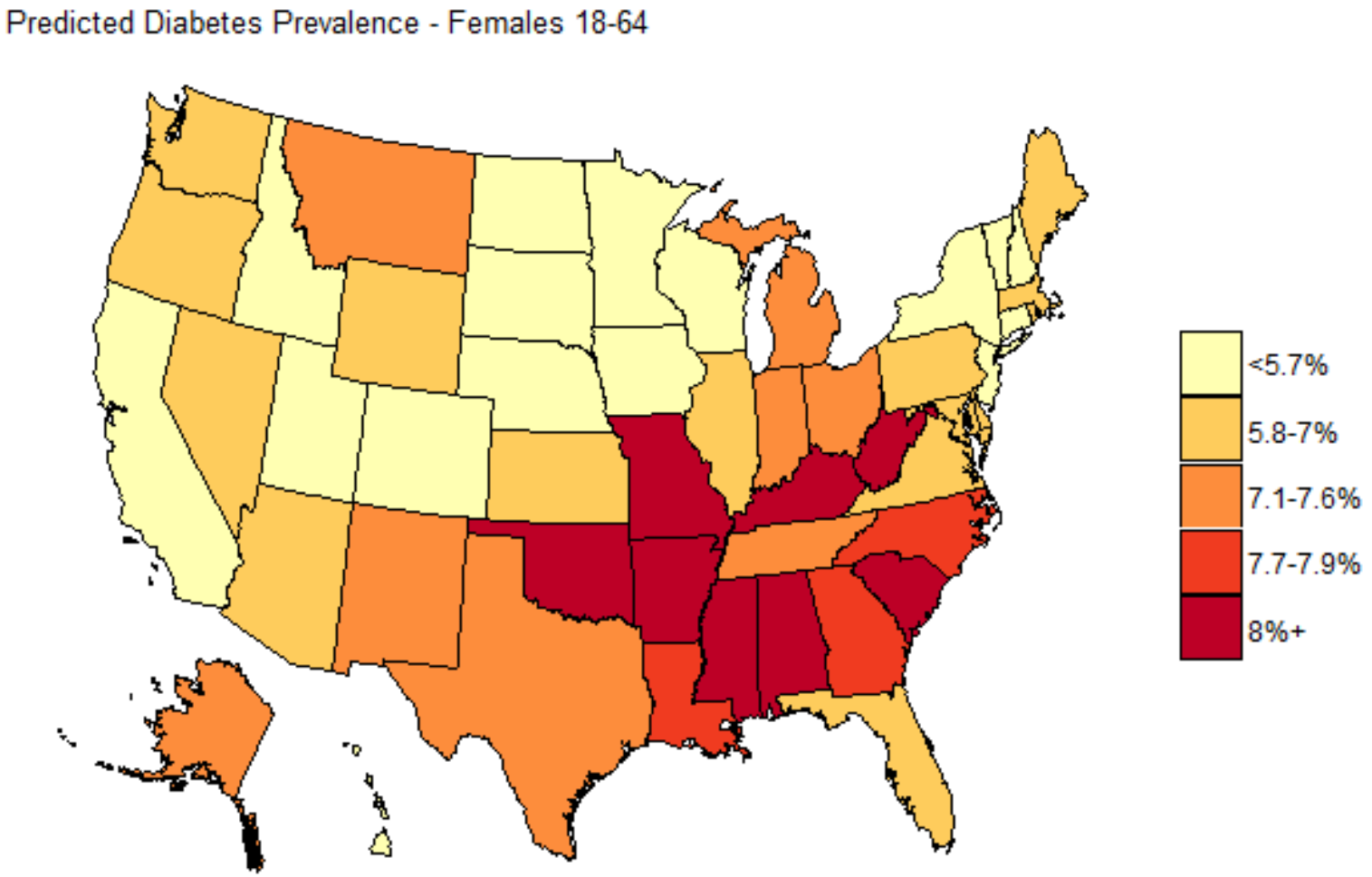


Figure 3: Estimates of Diabetics for domain 3: Females 18-64.

Females above the age 64

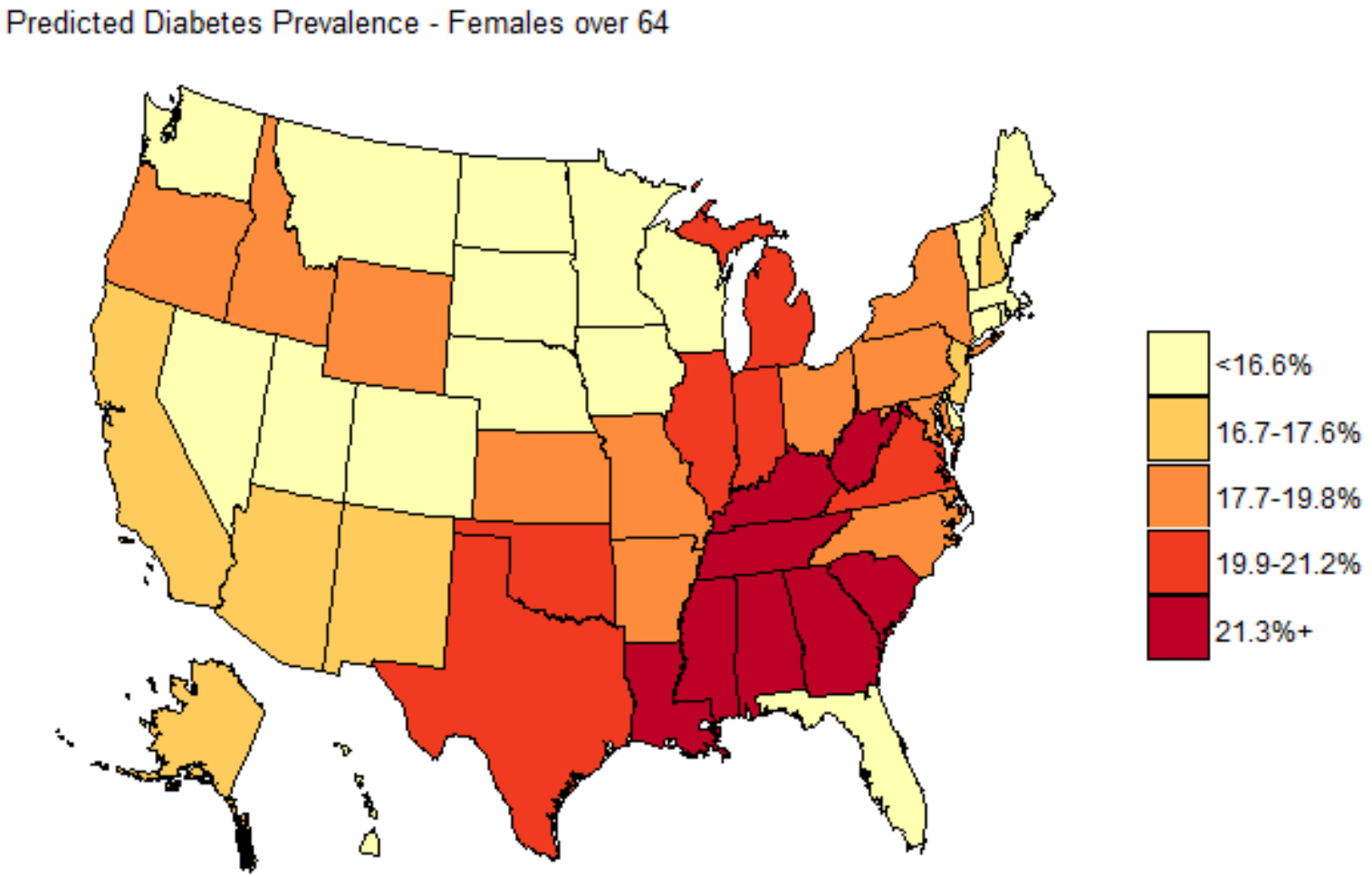


Figure 4: Estimates of Diabetics for domain 4: Females above 64.

Percentage difference between the NHIS design-based estimates and the model based SAEs.

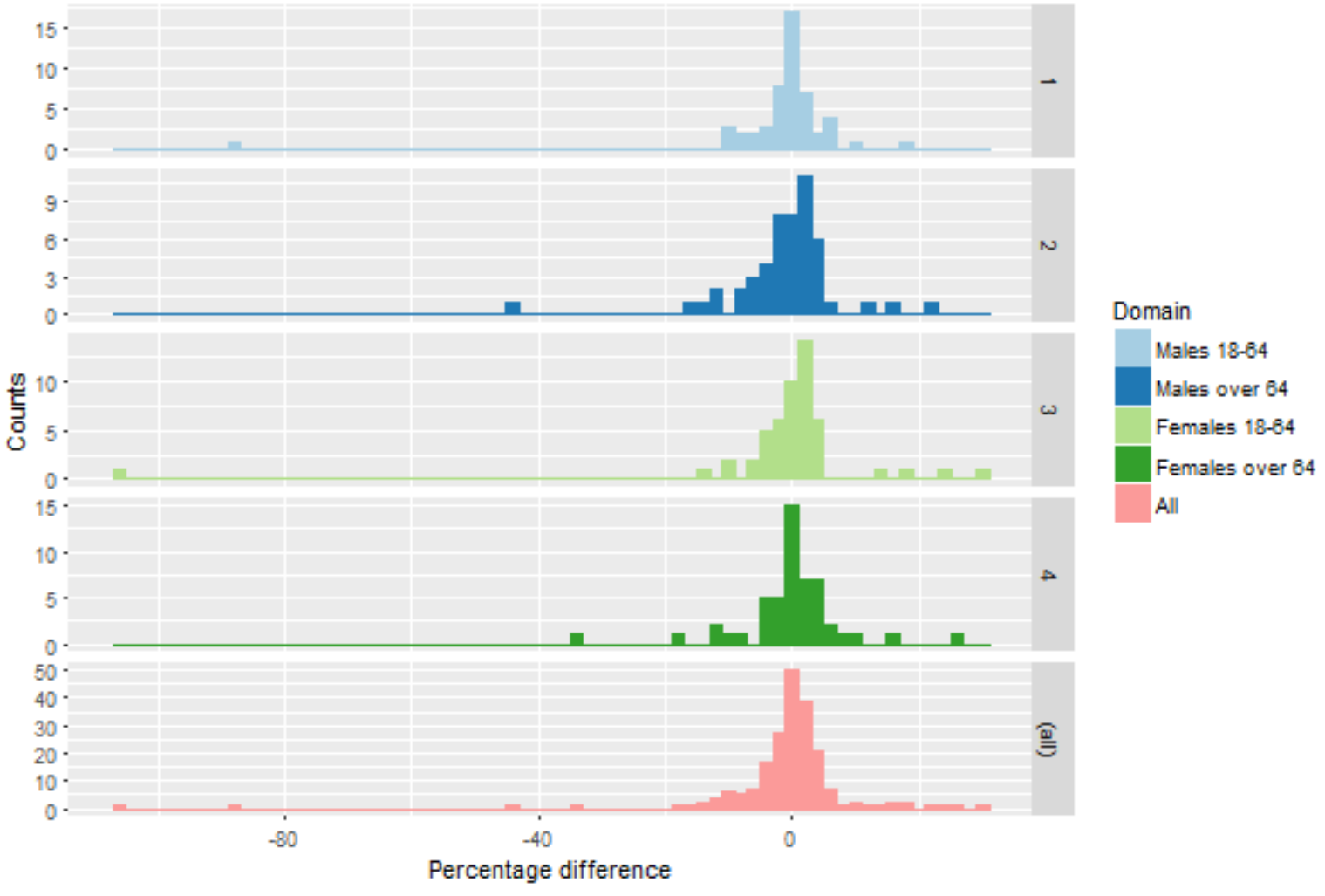


Figure 5: Percentage difference.

Residual analysis

Correlation

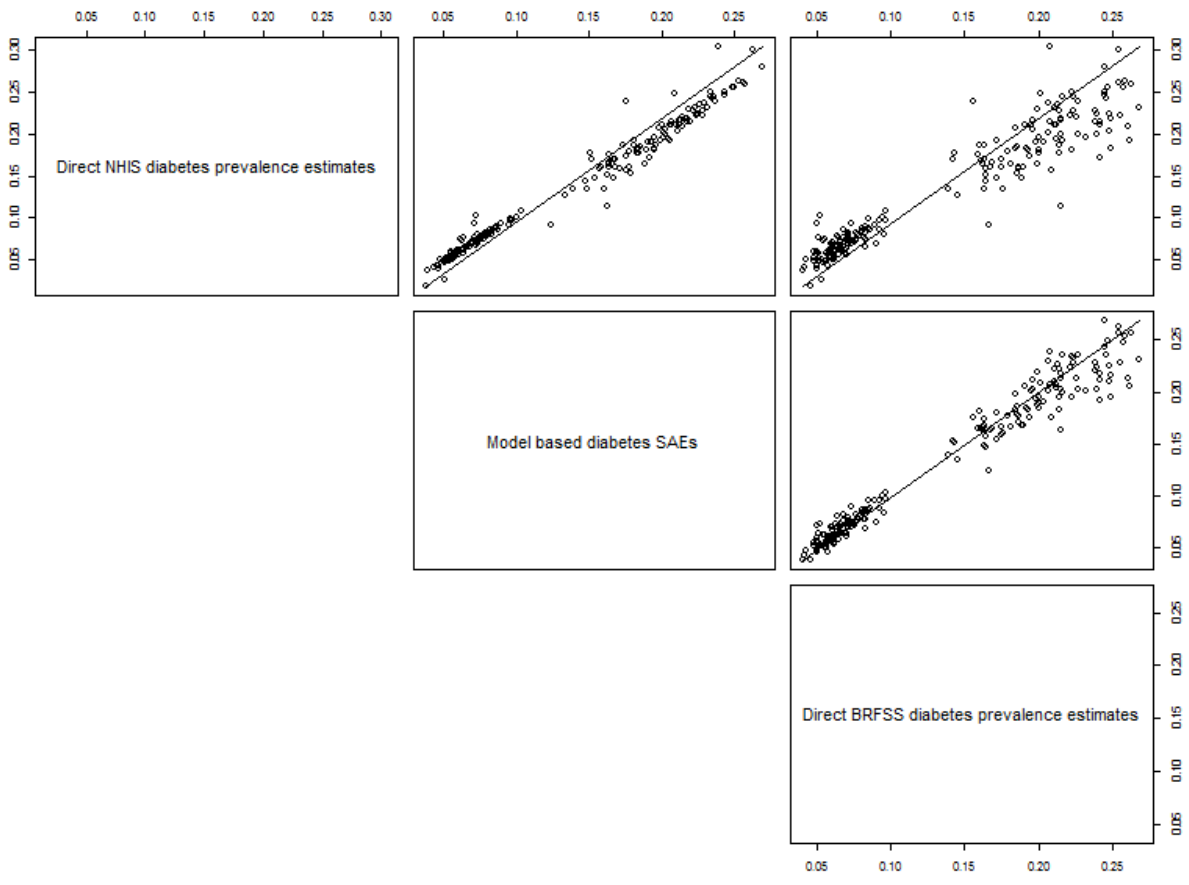


Figure 6: Correlations between the model based SAEs, NHIS design-based estimates and the BRFSS design based estimates.

Estimate Diagnostics

Table 4: Diagnostics with 2010 US Census population totals.

domain	NHIS design based National estimate	Model based national estimate aggregated by population totals
1	0.071278	0.0702
2	0.221015	0.2132
3	0.06516	0.065
4	0.186213	0.1822

Shrinkage

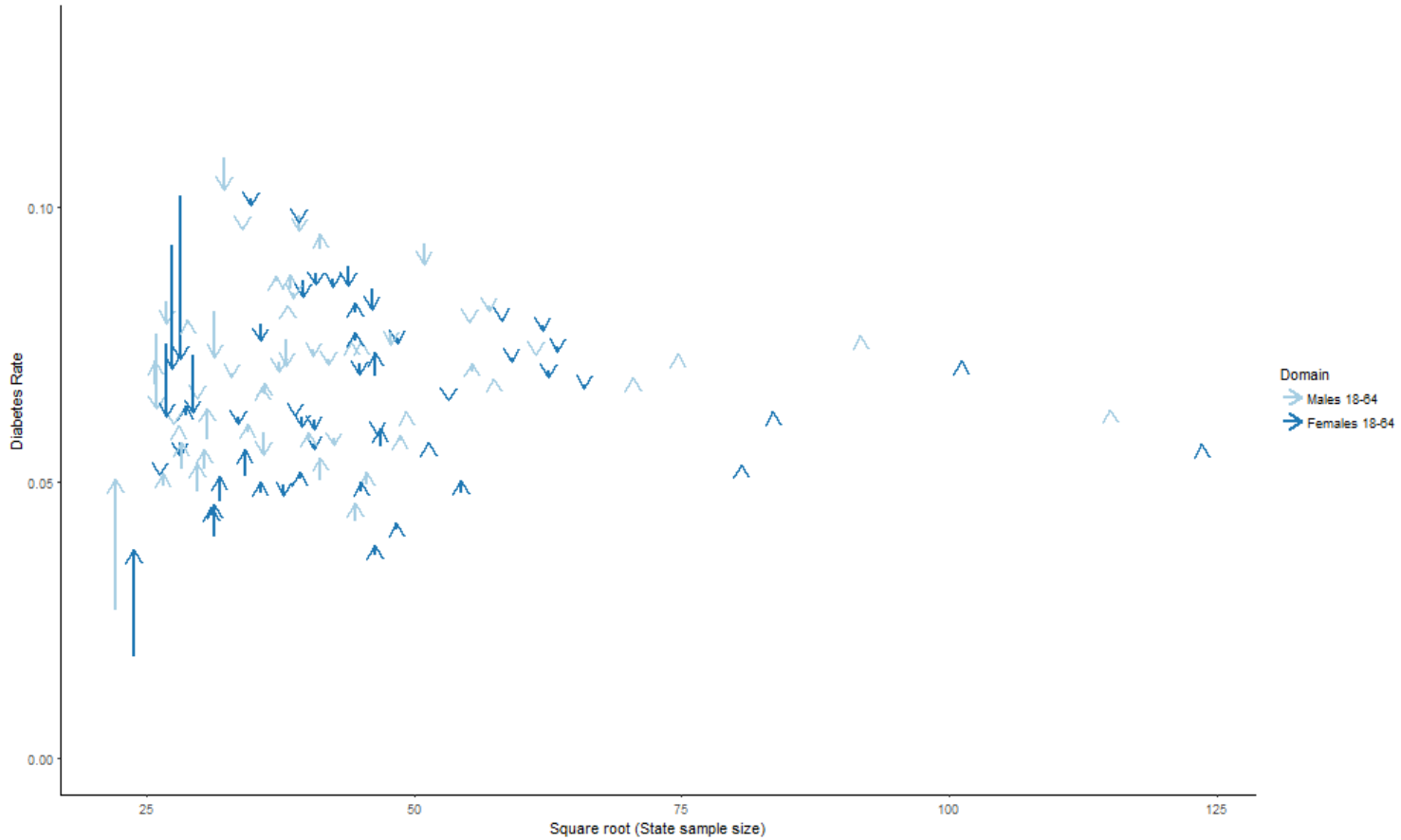


Figure 7: Shrinkage of the posterior estimates and the NHIS design-based estimates towards the mean. The start of the line is the NHIS design-based estimate and the end of the arrow is the model based SAE for the respective domain.

Shrinkage

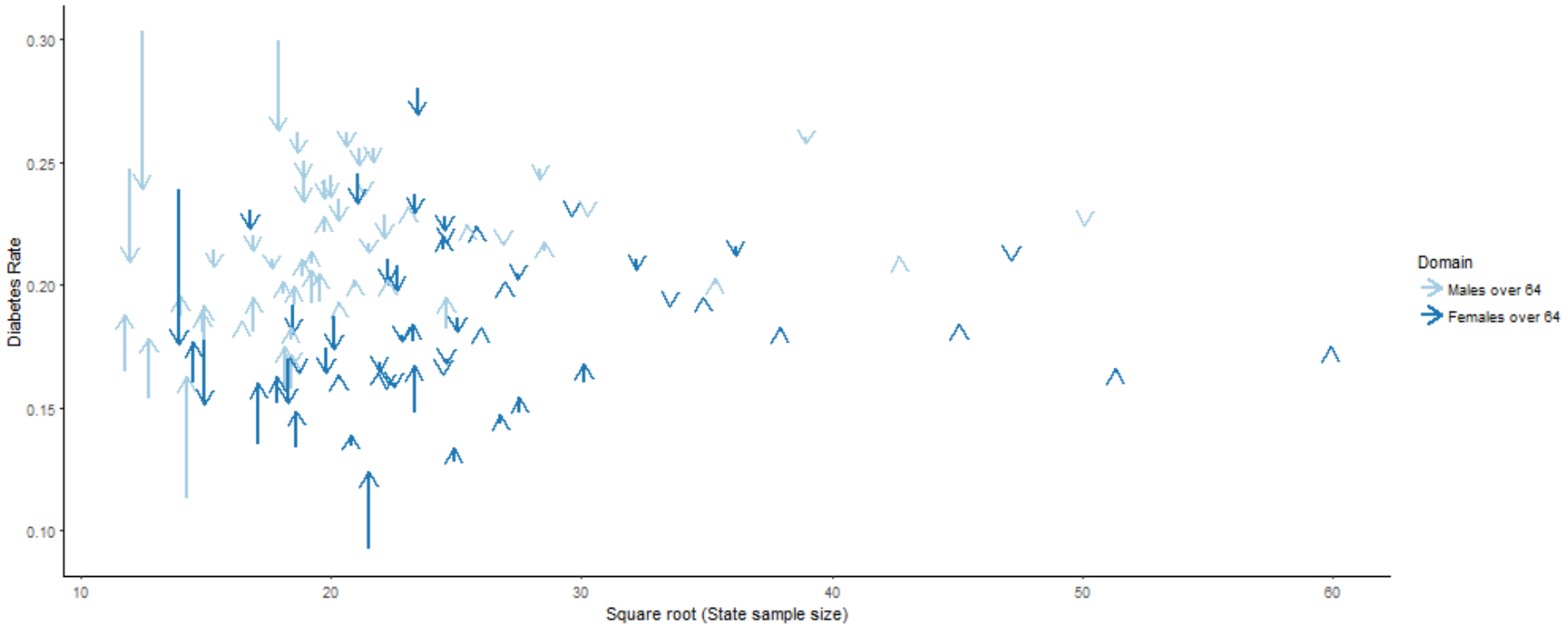


Figure 8: Shrinkage of the posterior estimates and the NHIS design-based estimates towards the mean. The start of the line is the NHIS design-based estimate and the end of the arrow is the model based SAE for the respective domain.

Lack of bias

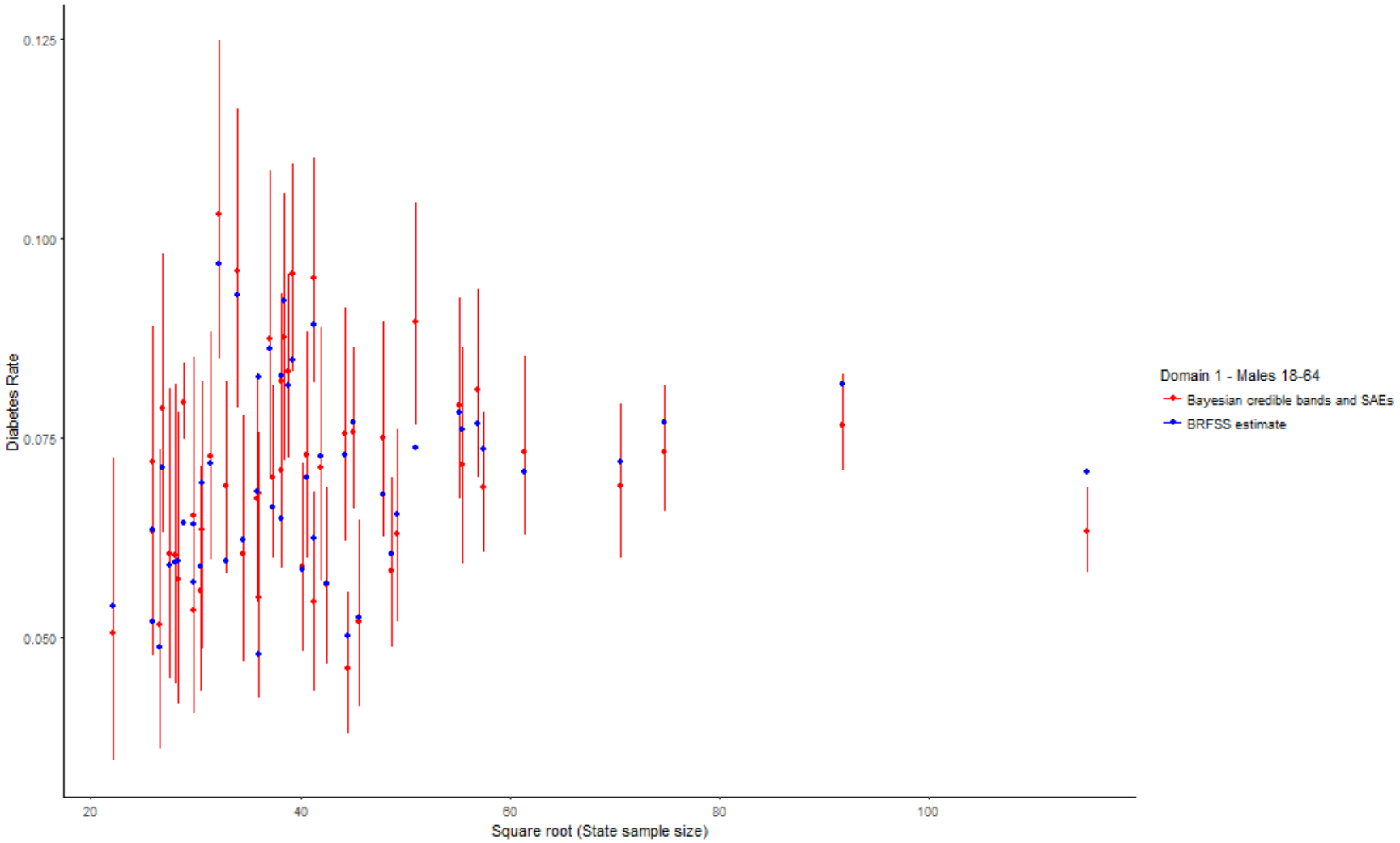


Figure 9: The model based SAEs (and the associated 95% Bayesian credible intervals) and the direct design based BRFSS estimates.

Lack of bias

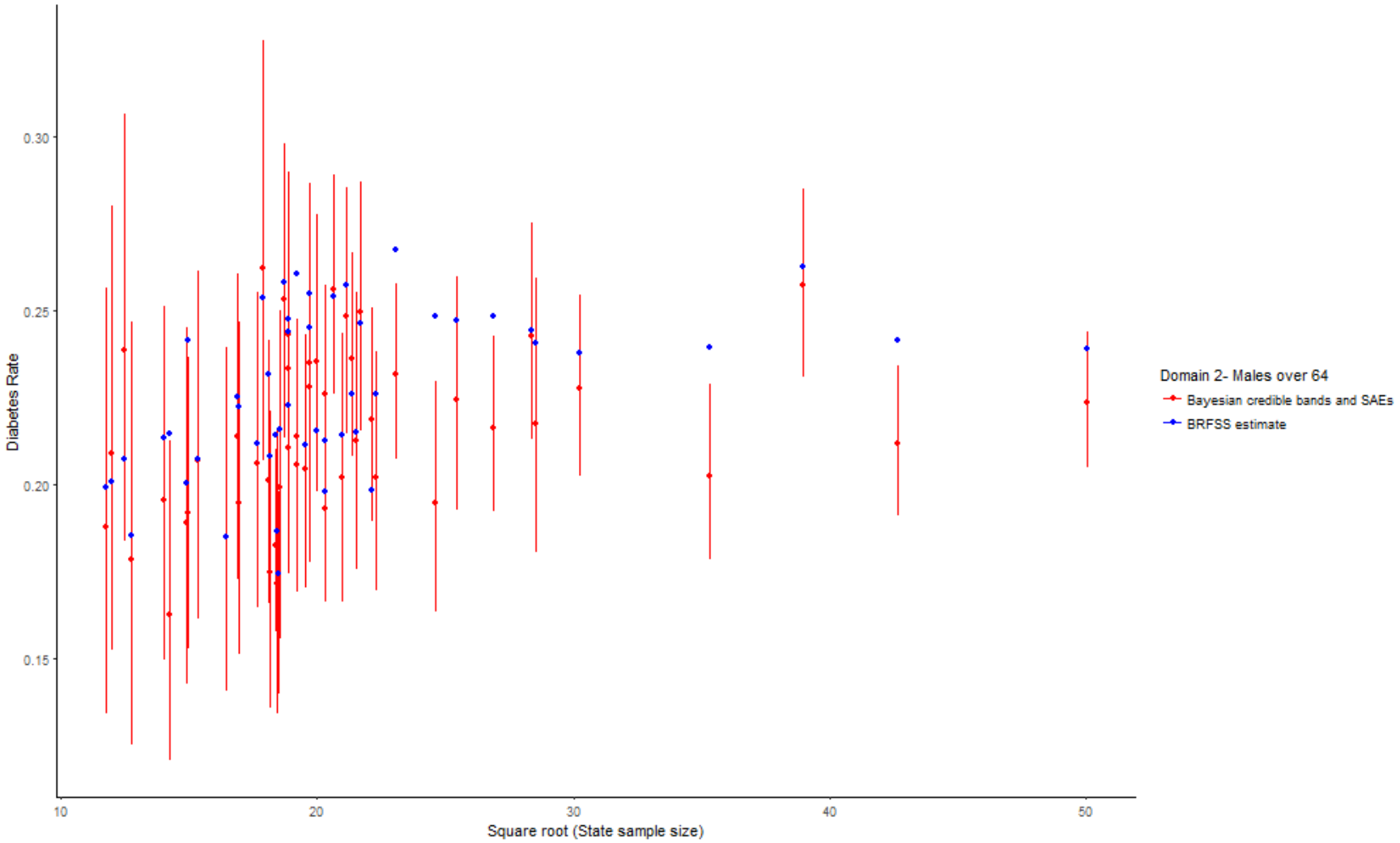


Figure 10: The model based SAEs (and the associated 95% Bayesian credible intervals) and the direct design based BRFSS estimates.

Lack of bias

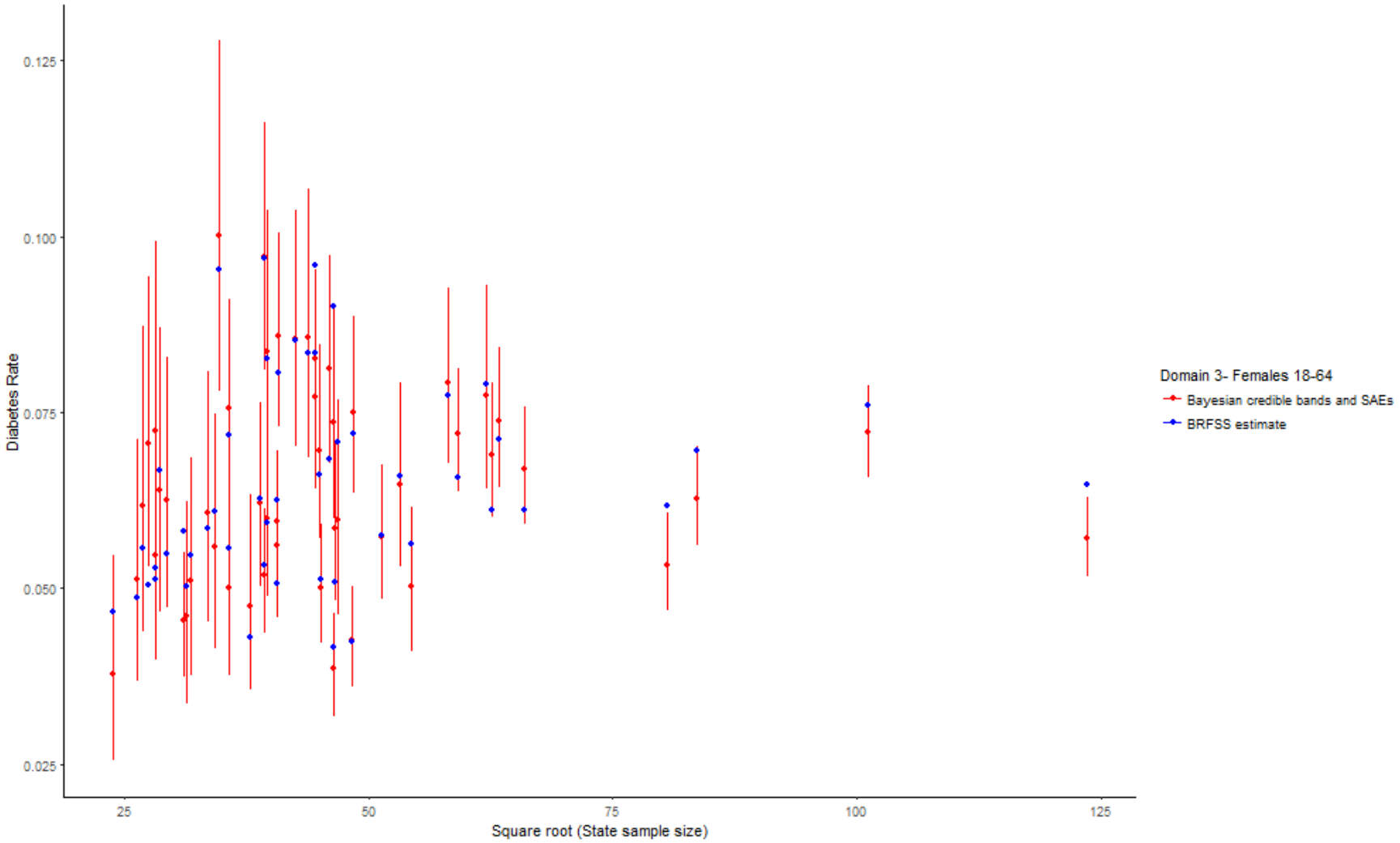


Figure 11: The model based SAEs (and the associated 95% Bayesian credible intervals) and the direct design based BRFSS estimates.

Lack of bias

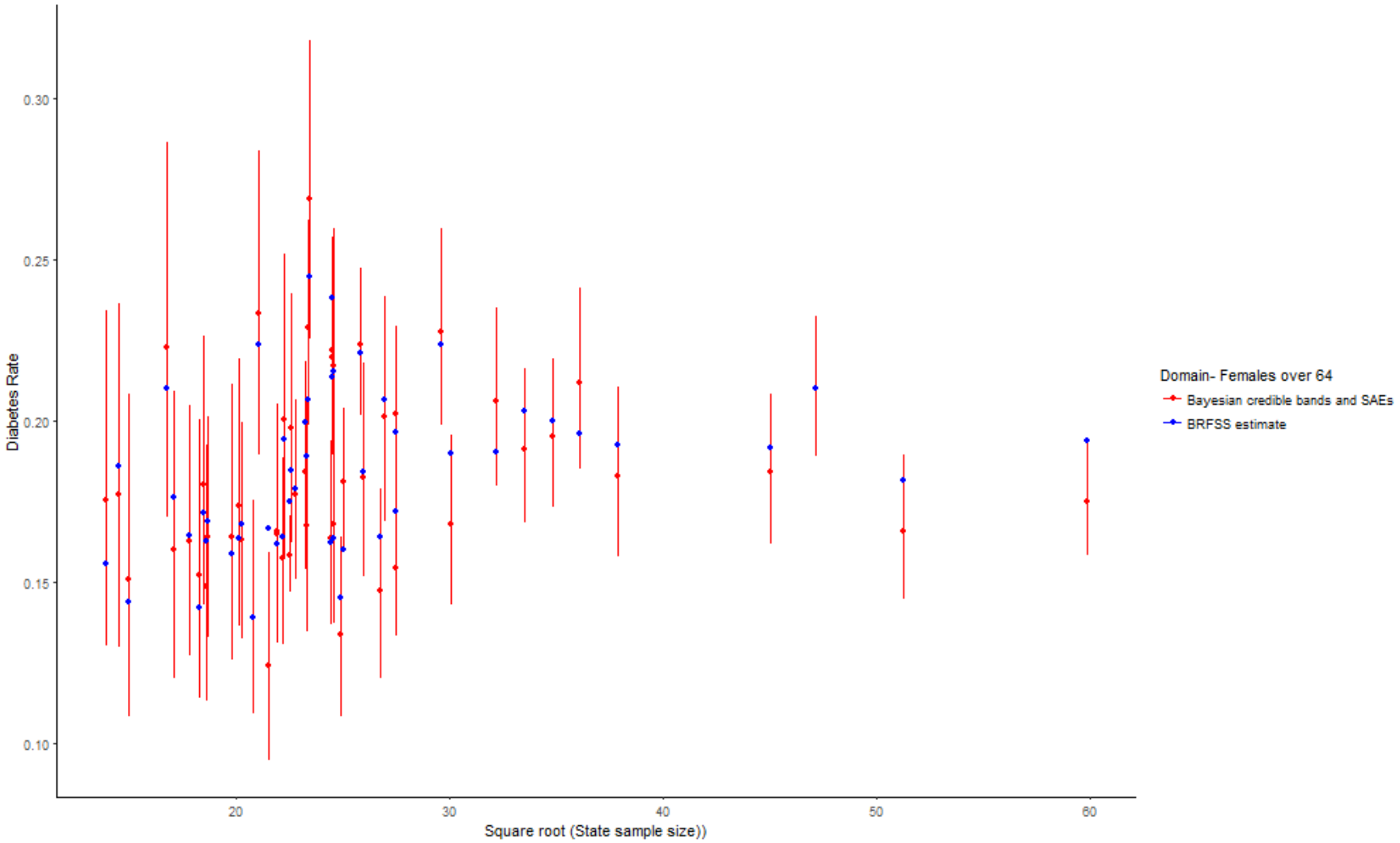


Figure 12: The model based SAEs (and the associated 95% Bayesian credible intervals) and the direct design based BRFSS estimates.

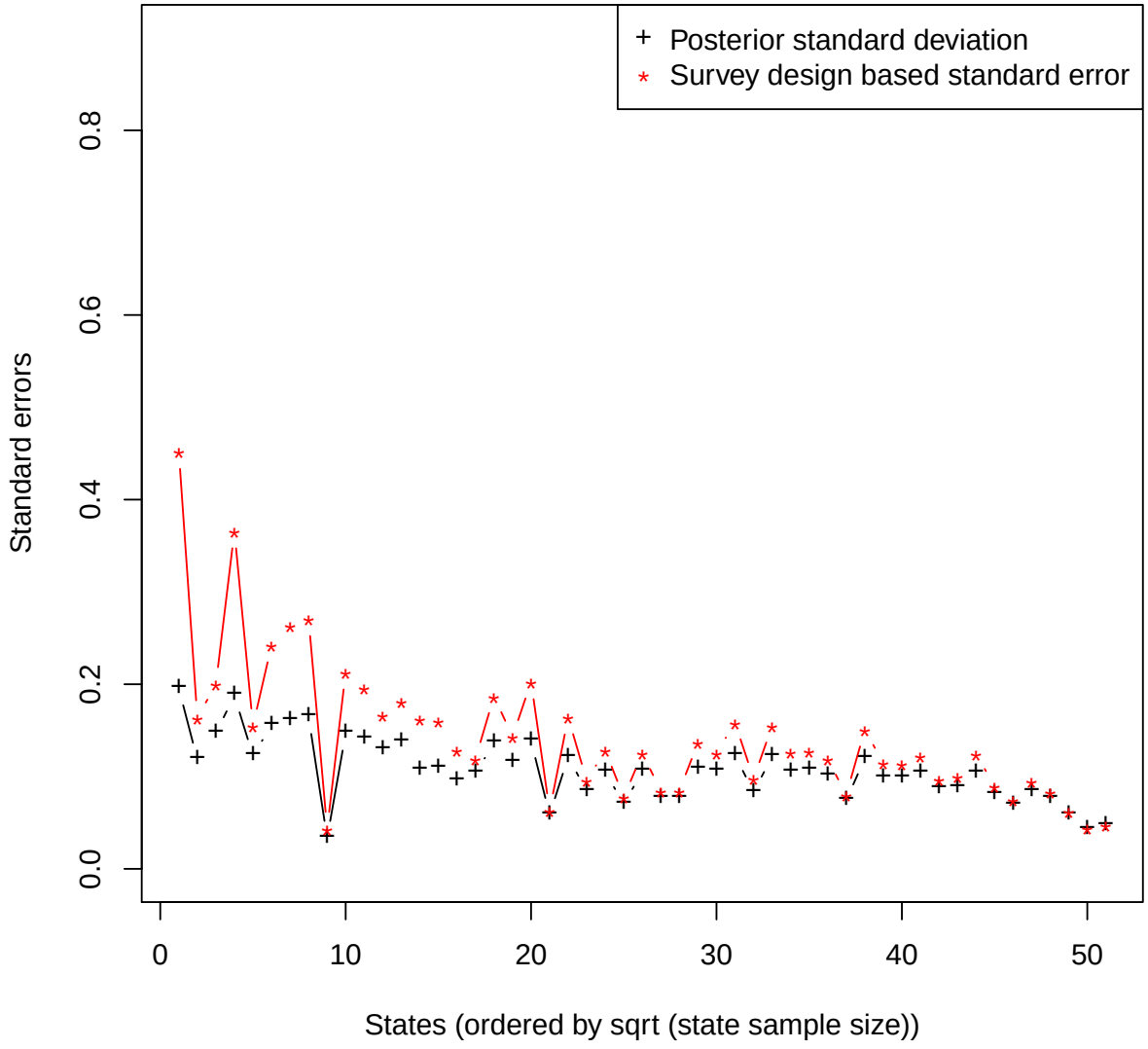


Figure 13: Standard errors for survey design based estimates and posterior standard deviations for model based SAEs for domain 1: Males 18-64.

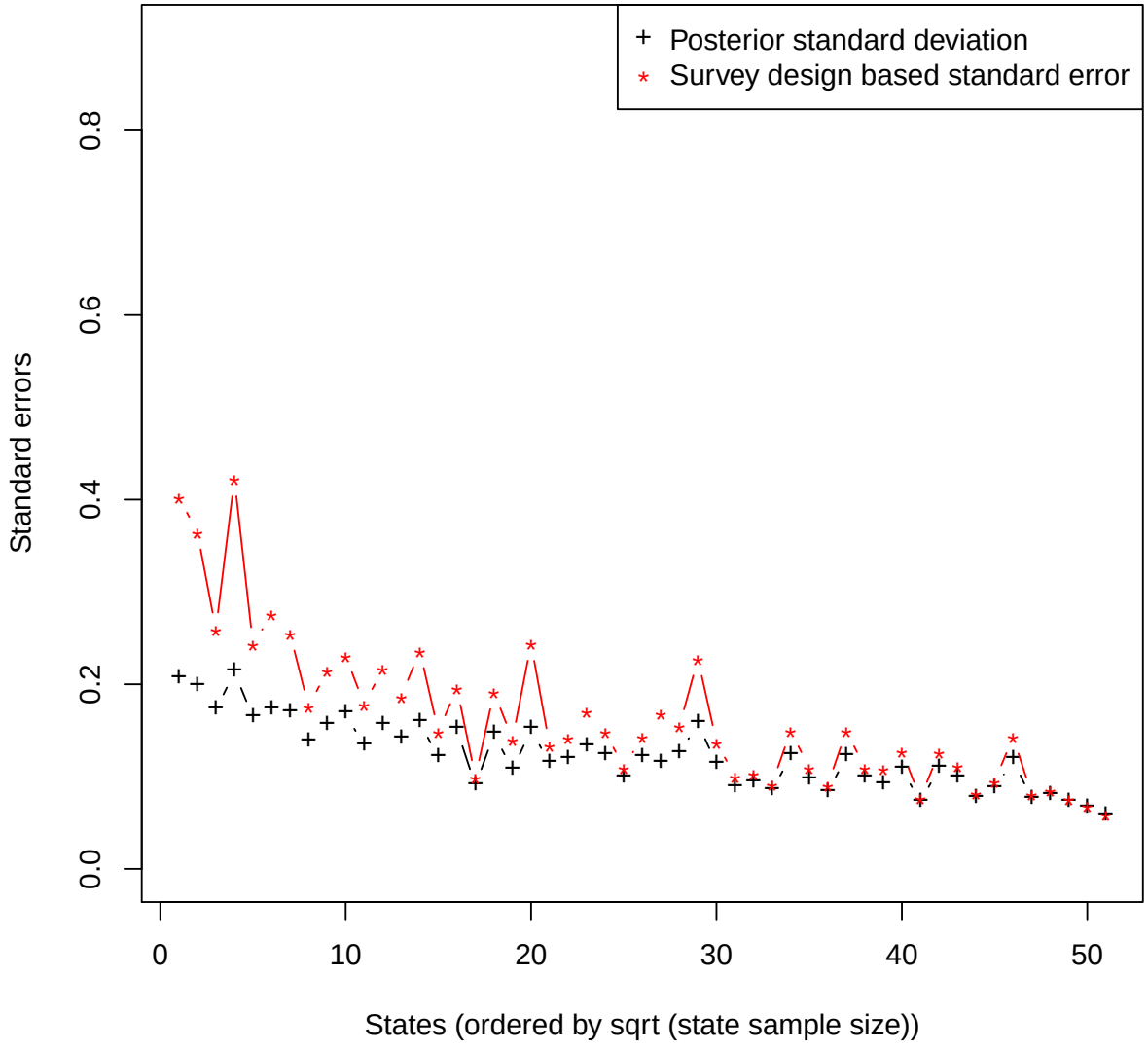


Figure 14: Standard errors for survey design based estimates and posterior standard deviations for model based SAEs for domain 2: Males above 64.

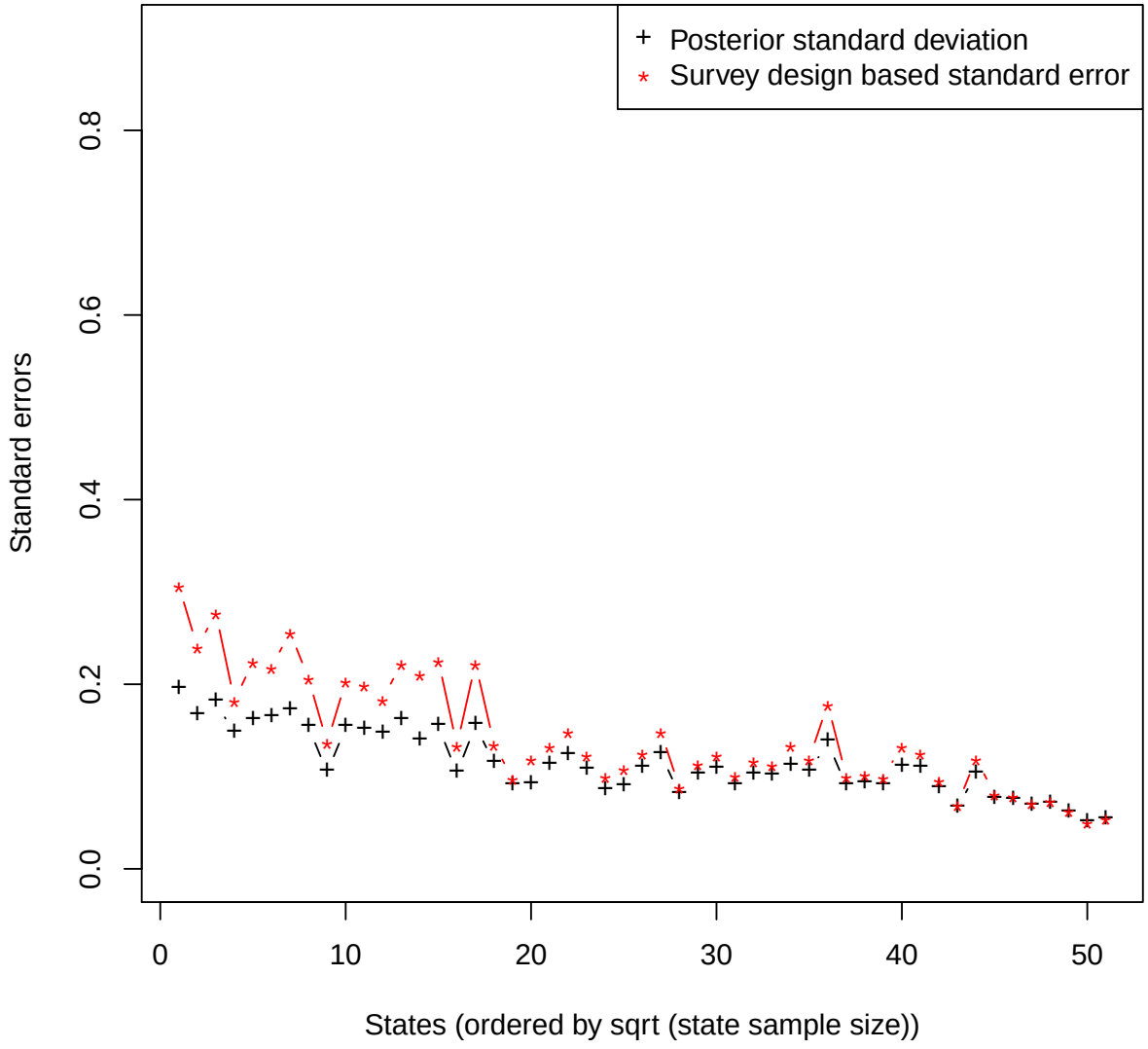


Figure 15: Standard errors for survey design based estimates and posterior standard deviations for model based SAEs for domain 3: Females 18-64.

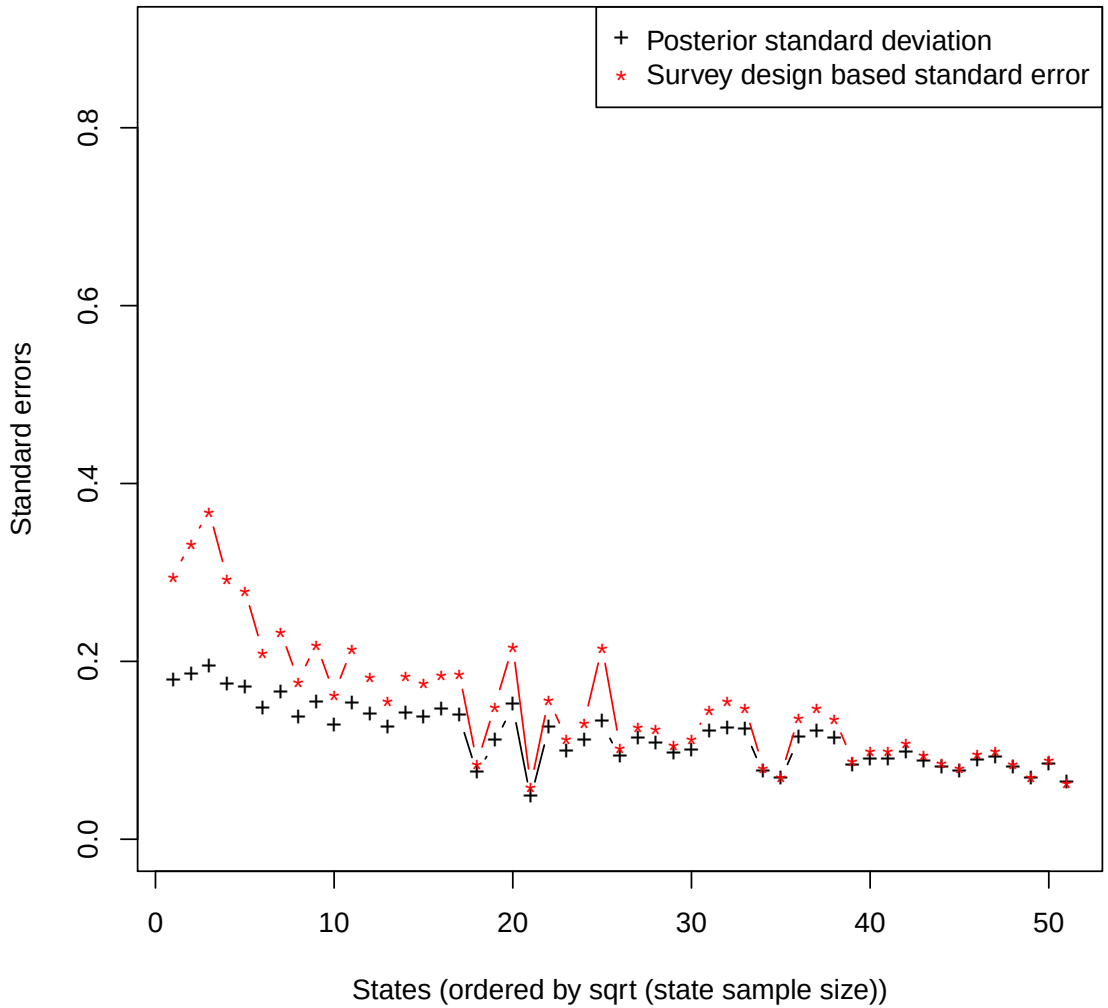


Figure 16: Standard errors for survey design based estimates and posterior standard deviations for model based SAEs for domain 4: Females above 64.

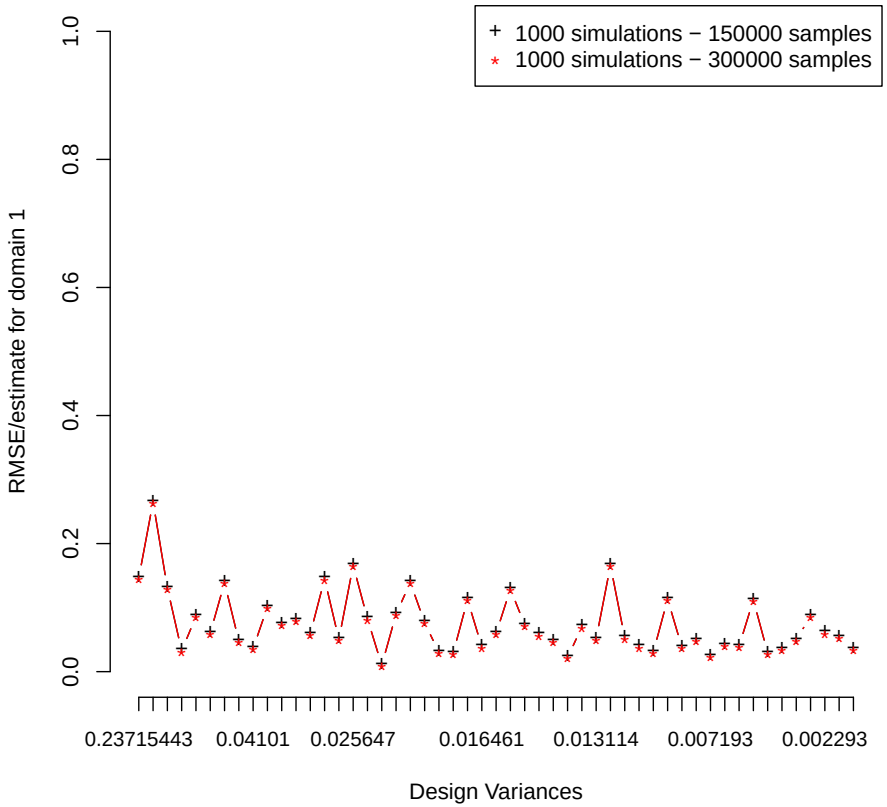
Simulation study - Evaluation through simulation

1. Used the direct design based diabetes prevalence NHIS estimates and the corresponding design based covariance matrix estimates as truth to generate artificial populations for 1000 cases
2. Specifically, in the simulation, for state i, where $i = (1, ..., 51)$, it is assumed that the population values are: $\mu_i^* = \hat{Y}_i$
3. Assume the sample covariance for state i is: $\sum_i^* = \hat{\sum}_i$
4. If $\hat{\sum}_i$ is non-singular then simulate samples from the population according to the sample design. Specifically, for state i where, $\hat{\sum}_i$ is non-singular,
   $\hat{Y_i^{(f)}} \sim MVN(\mu_i^*, \hat{\sum}_i^*)$, and
   $Df_i * \hat{\sum}_i^f \sim Wish(Df_i, \hat{\sum}_{\mathbf{i}}^*)$, where Df_i= degrees of freedom (number of PSUs - number of strata) for state i

Simulation study - Evaluation through simulation

1. For state i, where, $\hat{\sum}_i$ is singular, use the matrices $\hat{\sum}_i^{**}$ proposed in slide 9
2. Fitted the Bayesian model to 1000 simulated datasets to determine simulated diabetes prevalence estimates estimates μ_i^* for $i = 1, ..., 51$ states
3. Evaluated the accuracy of estimates via the coefficient of variations (CVs) (Root Mean Squared Error (RMSE)/ design based estimate)
4. RMSE is defined as:
   RMSE=Sqrt (Mean Squared error (NHIS design-based estimate, Simulated SAEs))
   $RMSE = sqrt(\frac{1}{G}\sum_{g=1}^{G}(\mu_i - \hat{\mu}_i))$, where G denotes the total number of simulated cases

Simulation study - 150000 vs 300000 samples



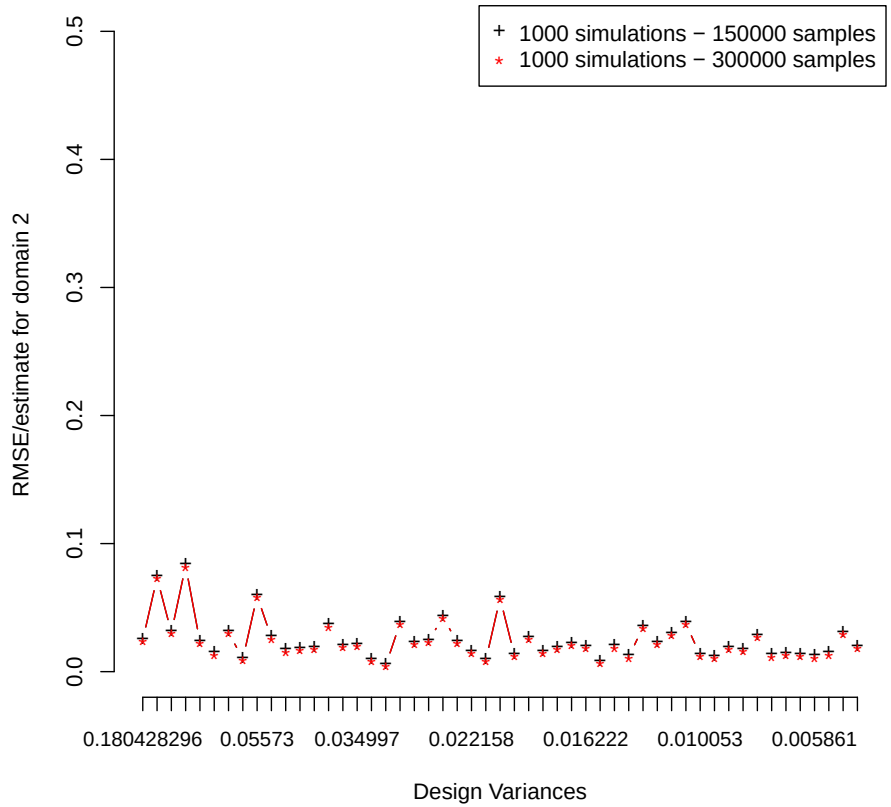


Figure 17: RMSE/NHIS design-based value vs design based variance (in decreasing order) for 150000 vs 300000 samples

Simulation study - 150000 vs 300000 samples

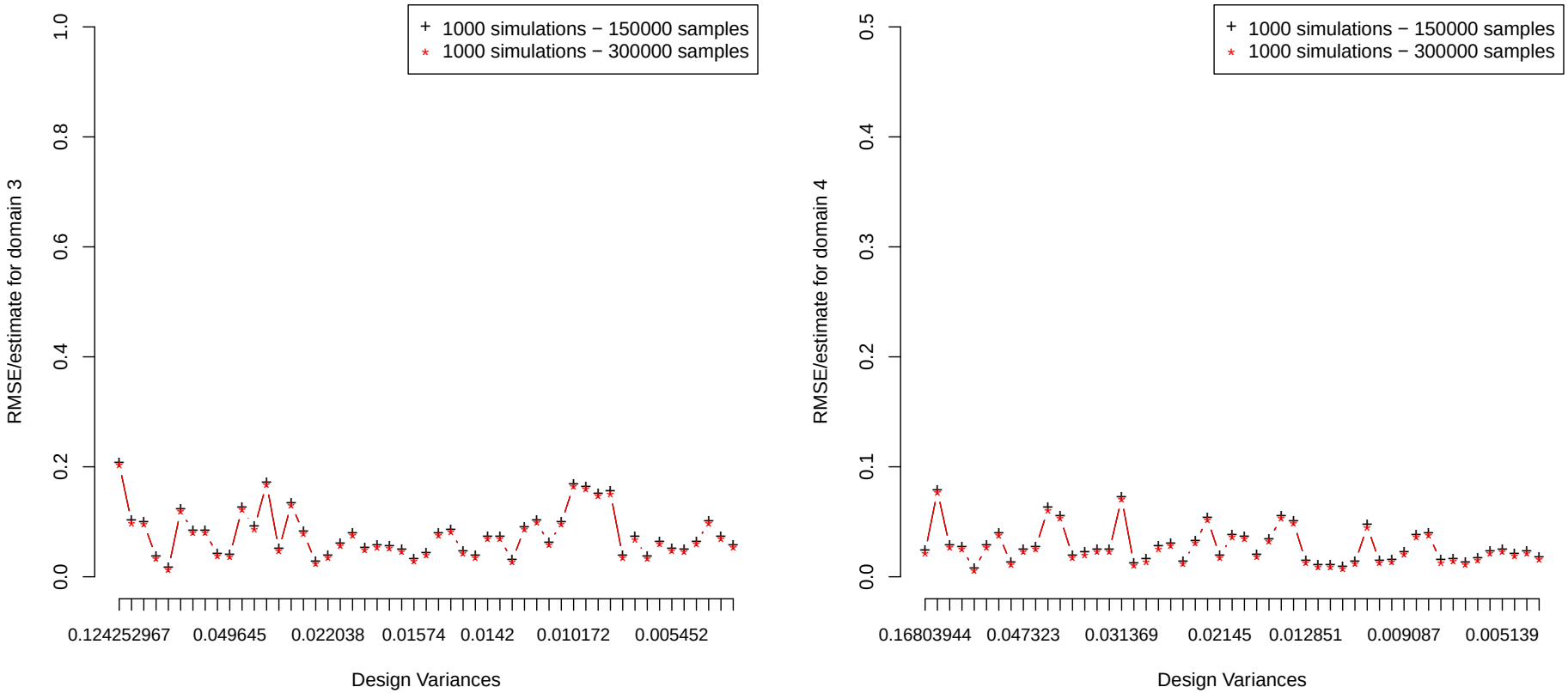


Figure 18: RMSE/NHIS design-based value vs design based variance (in decreasing order) for 150000 vs 300000 samples

Simulation study - Domain 1

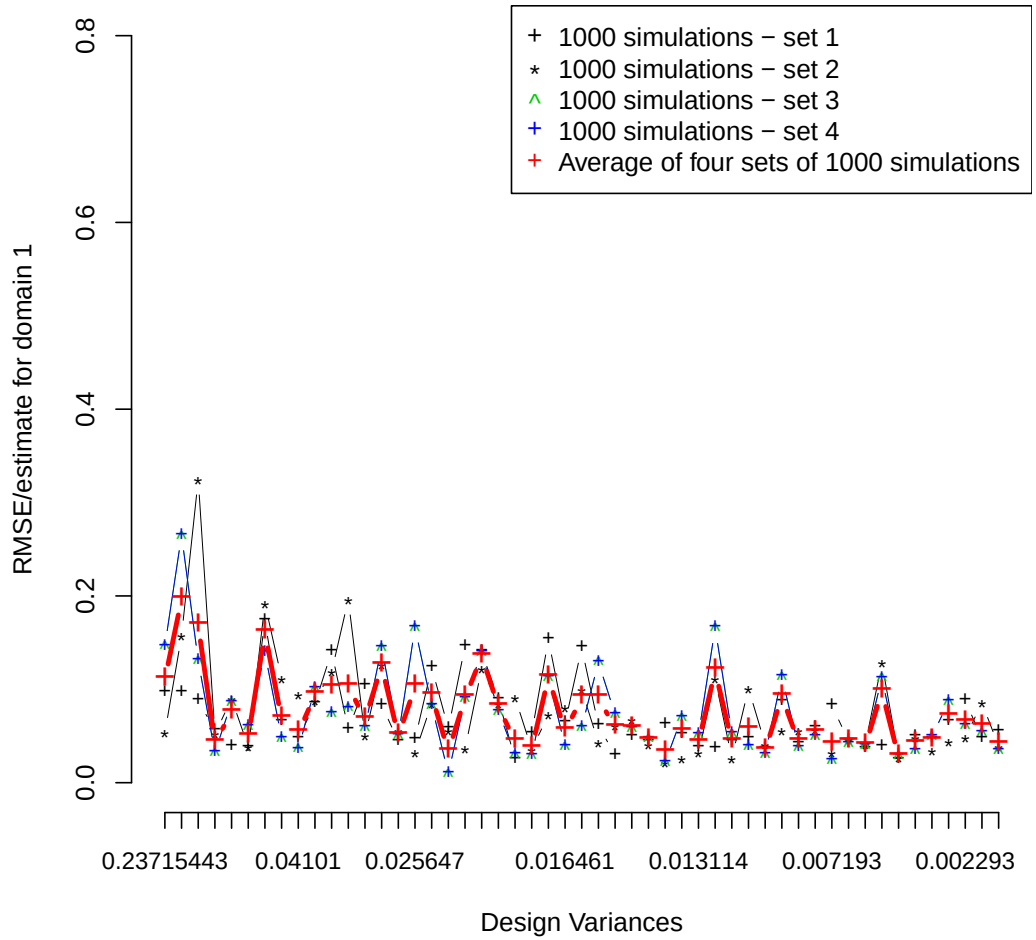


Figure 19: RMSE/NHIS design-based value vs design based variance (in decreasing order)

Simulation study - Domain 2

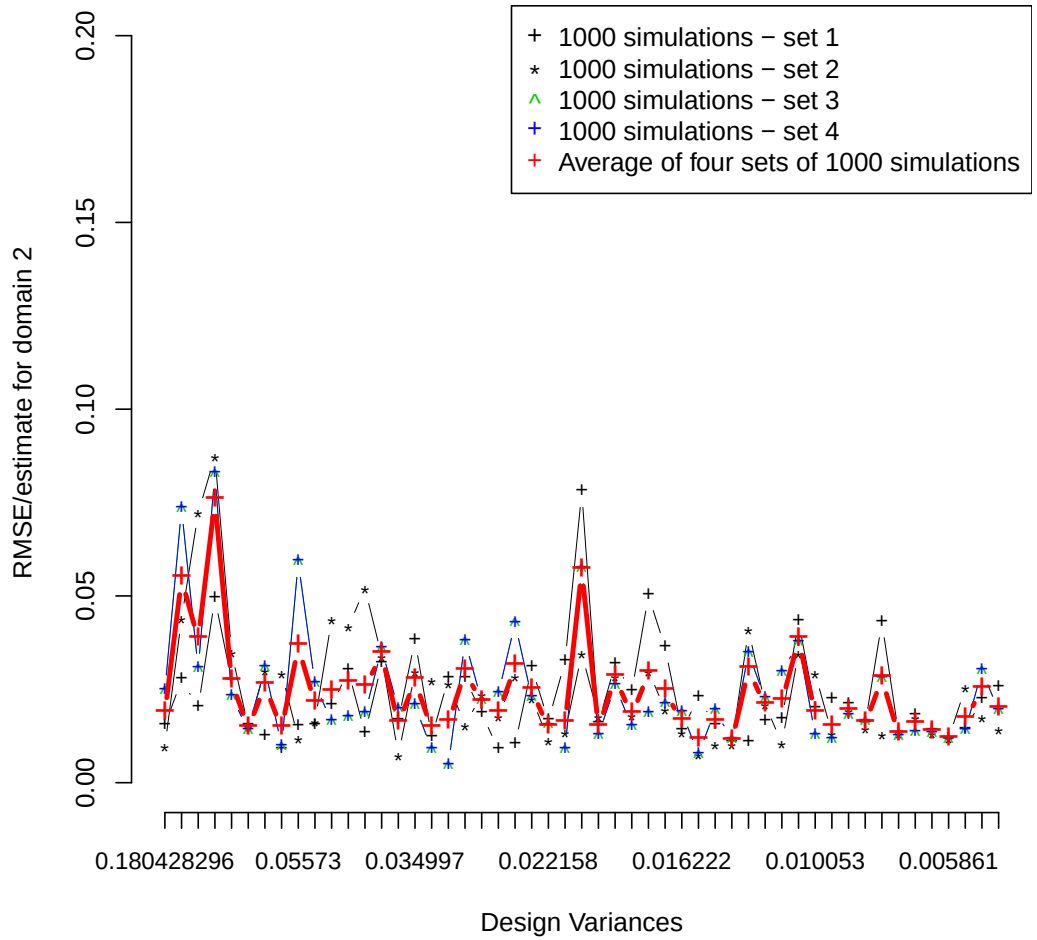


Figure 20: RMSE/NHIS design-based value vs design based variance (in decreasing order)

Simulation study - Domain 3

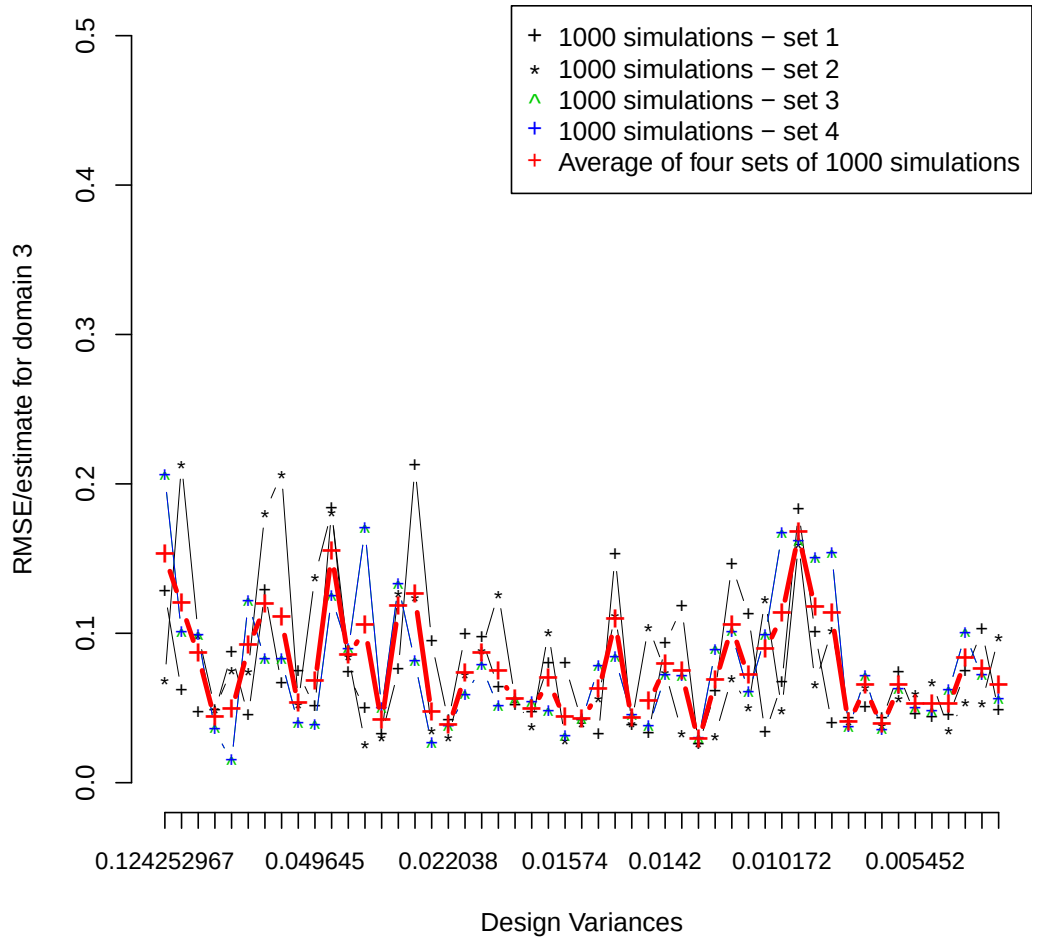


Figure 21: RMSE/NHIS design-based value vs design based variance (in decreasing order)

Simulation study - Domain 4

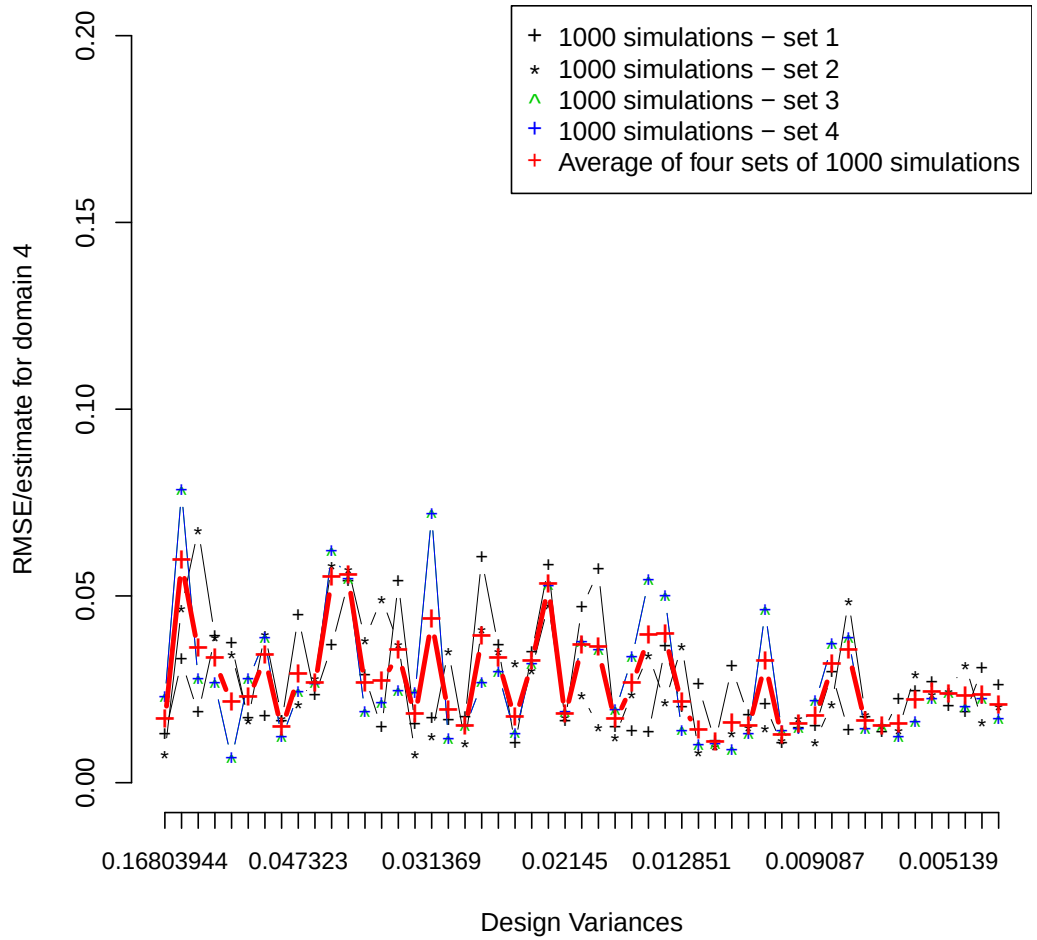


Figure 22: RMSE/NHIS design-based value vs design based variance (in decreasing order)

Conclusions

1. There might be some error in the covariate estimates
2. All ARF variables were non-significant
3. BRFSS estimates as covariates:one domain was highly correlated with design based NHIS estimates
4. High Diabetes prevalence was found to be in: Oklahoma, West Virginia, Ohio, Kentucky, Indiana, South Carolina, Missouri, Mississippi, Tennessee, Virginia, North Carolina, Texas, Alabama and some pockets in the western U.S.
5. This method can provide improved estimates (coefficient of variation is less than or equal to 0.3)
6. The simulation study checks for model failures that are important to inference

The End

1. More details in Khan D., Wei R., He, Y., Shin, H., Malec, D.J. Bayesian State-level Estimates of Diabetes Prevalence in United States, 2006 – 2015. Working manuscript.
2. Thanks..
3. Questions..

Contact information

Diba Khan *ild*1@*cdc.gov* or 301-458-4474