Supplemental Materials for "Efficient and Intuitive Two-Phase Validation Across Multiple Models via Principal Components"

Sarah Lotspeich[1] and Cole Manschot[2]
[1]Department of Statistical Sciences, Wake Forest University
[2]Biostatistics and Research Decision Sciences, Merck & Co.

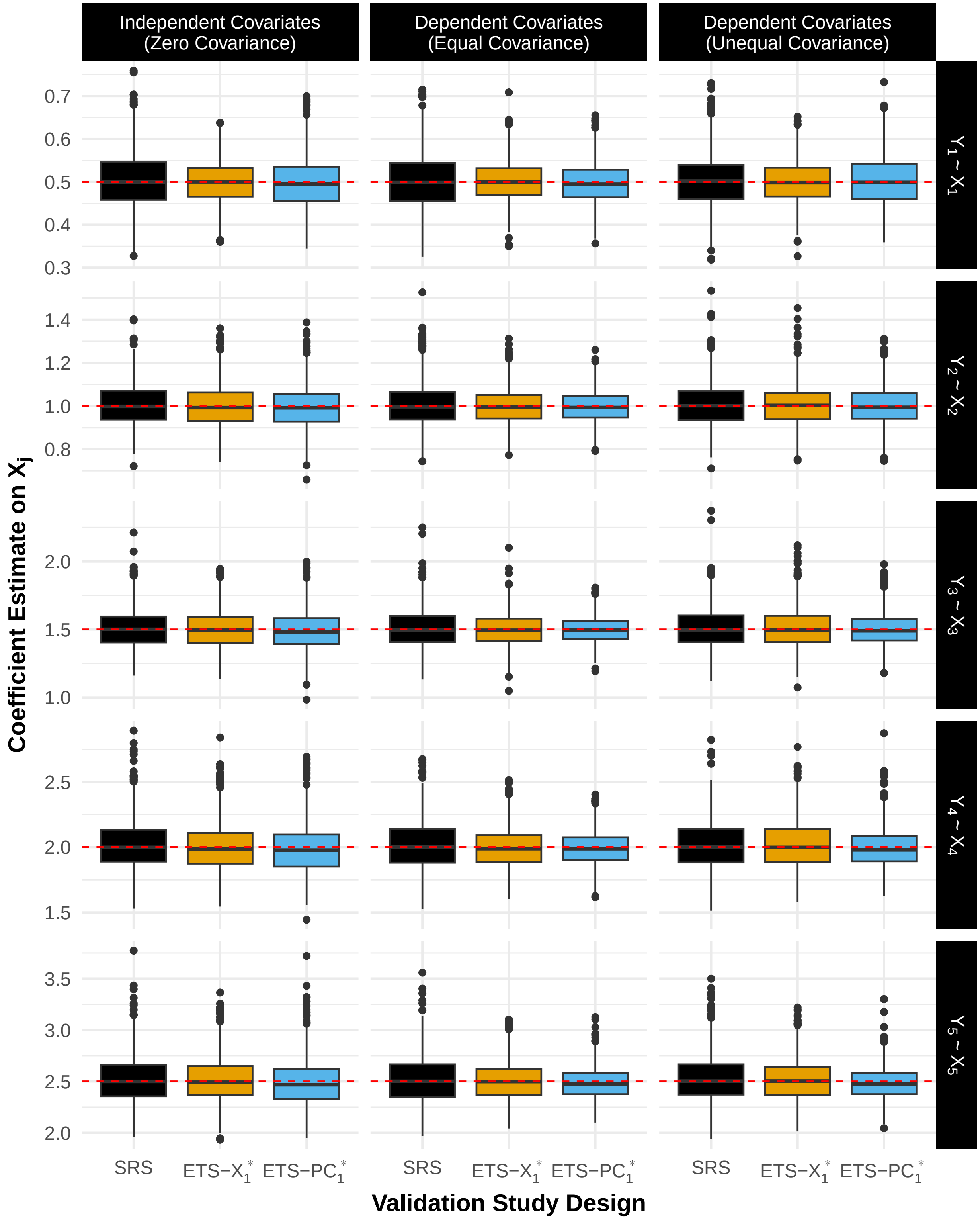


Figure S1: Simulation results comparing coefficient estimates under simple random sampling (SRS), extreme tail sampling on X_1^* (ETS-X_1^*), and extreme tail sampling on the first principal component (ETS-PC_1^*) validation study designs. Three different covariance structures for the five covariates $X_1, \ldots, X_5$ were considered.

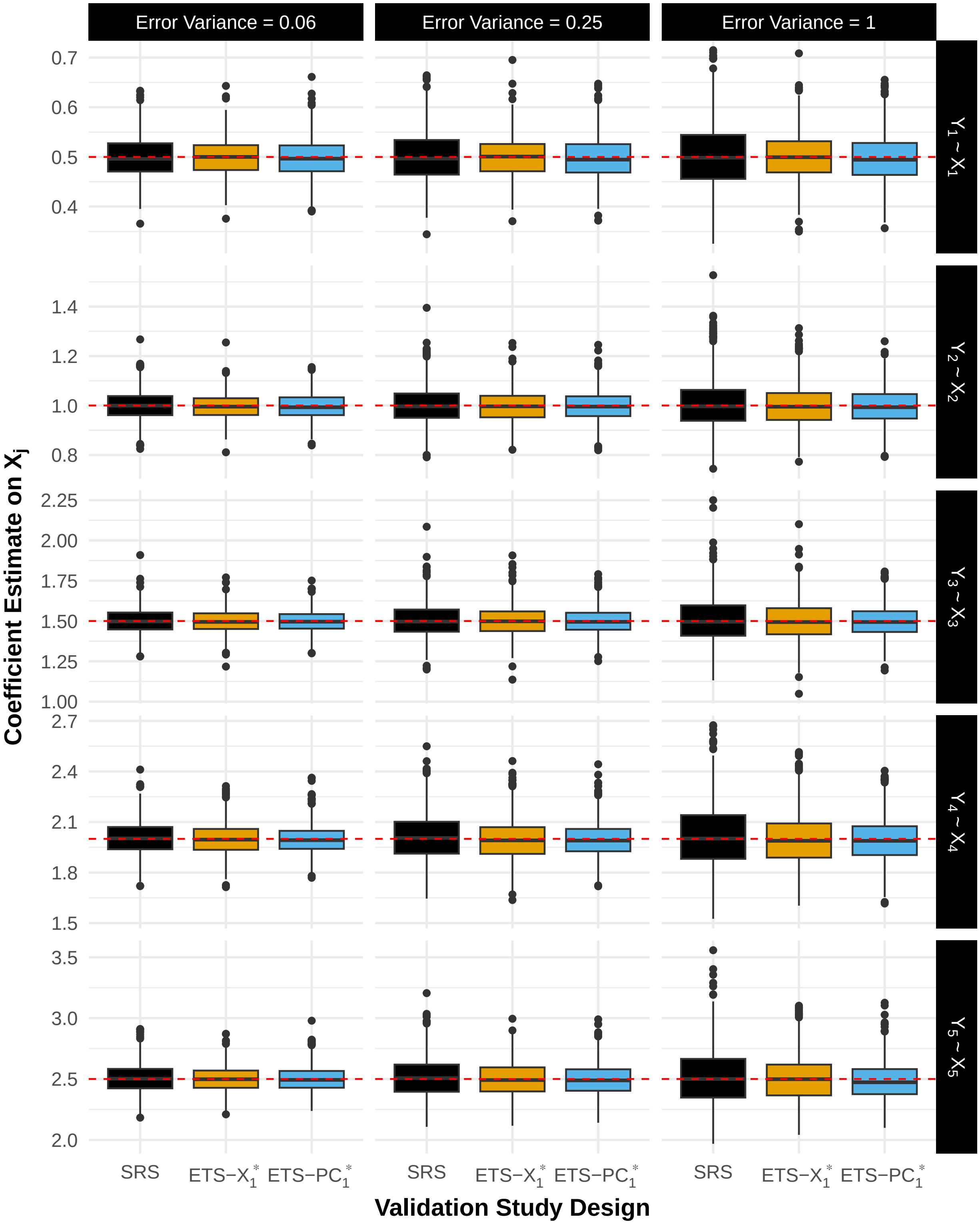


Figure S2: Simulation results comparing coefficient estimates under simple random sampling (SRS), extreme tail sampling on X_1^* (ETS-X_1^*), and extreme tail sampling on the first principal component (ETS-PC_1^*) validation study designs. Three different variances σ_U^2 for the additive measurement errors $U_1, \ldots, U_5$ in covariates $X_1, \ldots, X_5$ were considered.

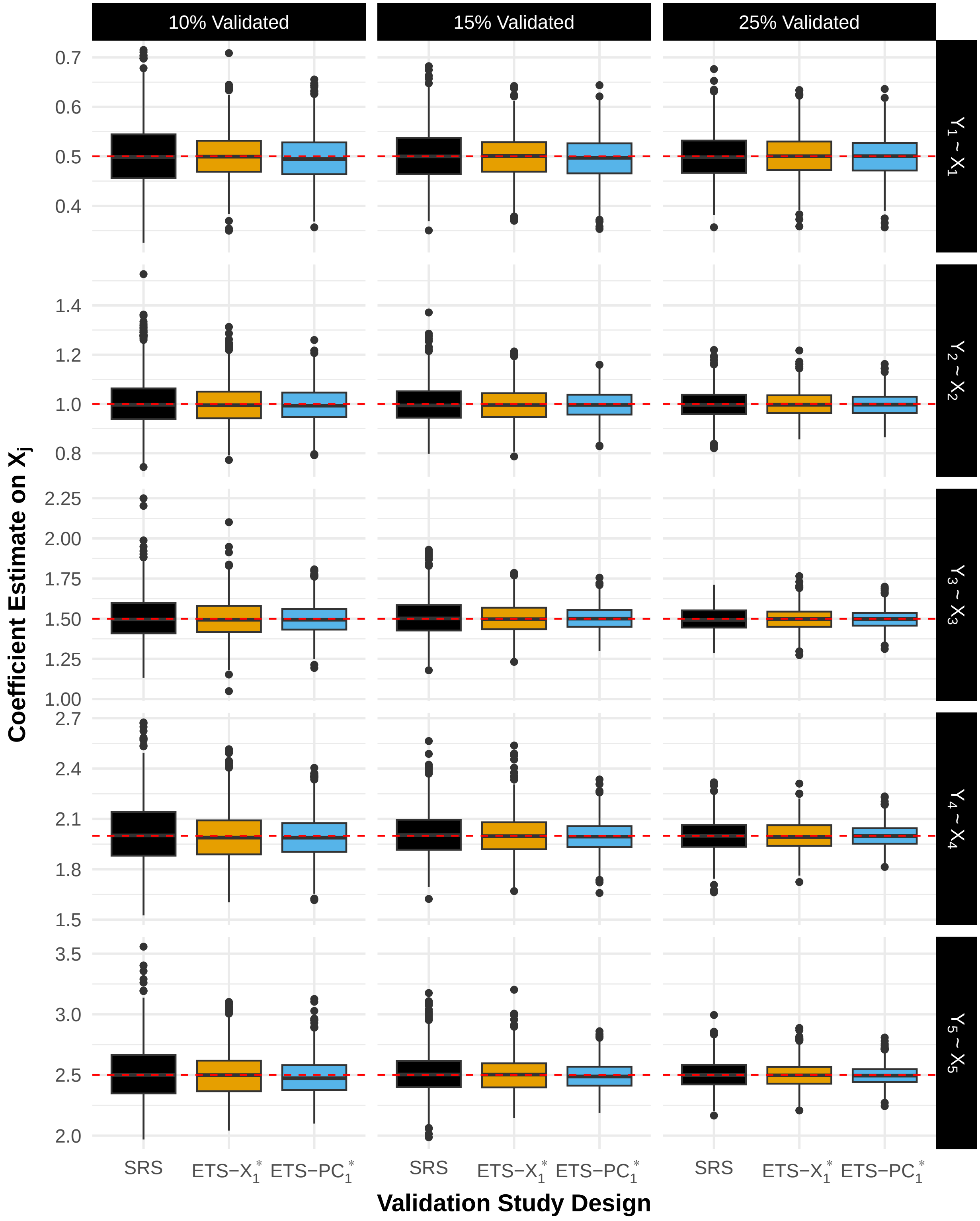


Figure S3: Simulation results comparing coefficient estimates under simple random sampling (SRS), extreme tail sampling on X_1^* (ETS-X_1^*), and extreme tail sampling on the first principal component (ETS-PC_1^*) validation study designs. Three different proportions of validated patients out of $N = 1000$ were considered.

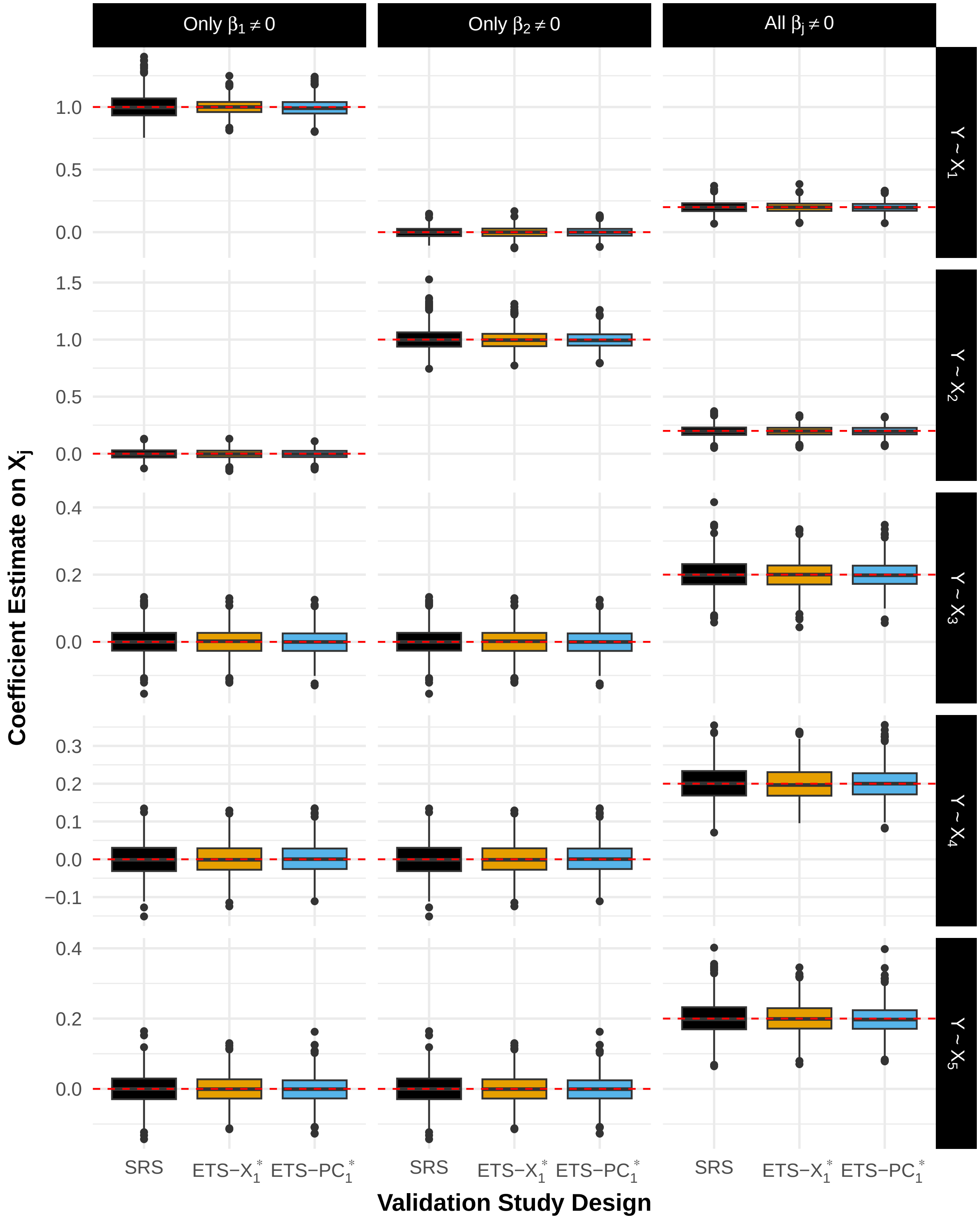


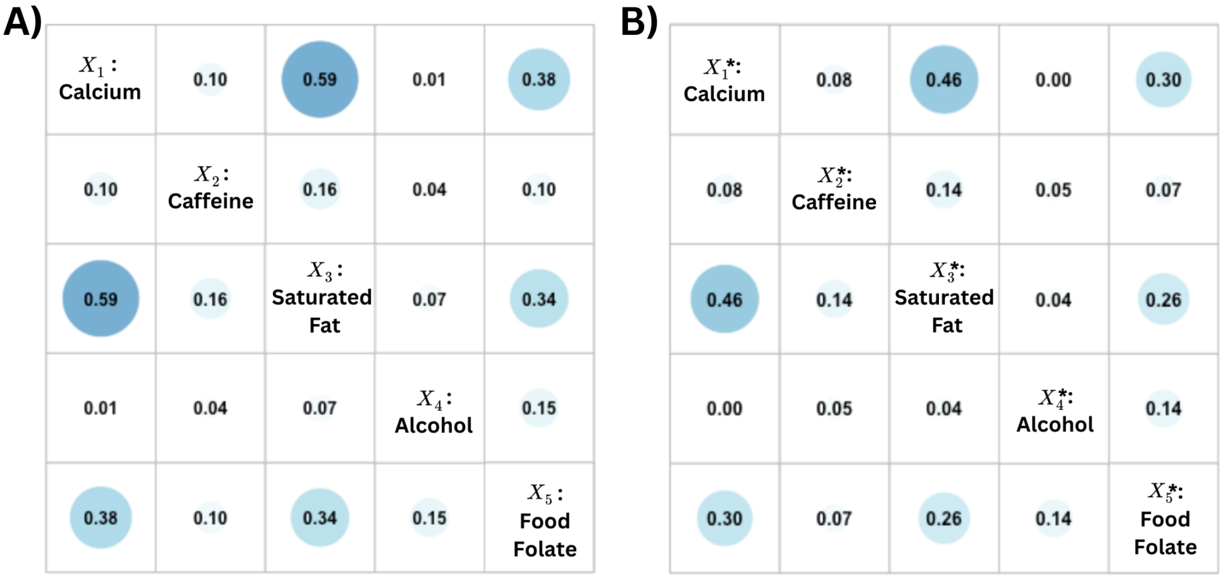
Figure S4: Simulation results comparing coefficient estimates under simple random sampling (SRS), extreme tail sampling on X_1^* (ETS-X_1^*), and extreme tail sampling on the first principal component (ETS-PC_1^*) validation study designs. There was a shared outcome Y, and it was generated from the covariates $X_1, \ldots, X_5$ under scenarios where only one covariate is associated (only $\beta_1 \neq 0$ or only $\beta_2 \neq 0$) versus all covariates are associated (all $\beta_j \neq 0$).

Outcome (Units)	Description	Variable	Source Data
Y_1 : Vitamin D (nmol/L)	25-hydroxyvitamin D2 + D3	`LBXVIDMS`	Laboratory values
Y_2 : Resting heart rate (bpm)	Pulse, first oscillometric reading	`BPXOPLS1`	Examination data
Y_3 : High-density lipoprotein (HDL) cholesterol (mg/dL)	Direct HDL-cholesterol	`LBDHDD`	Laboratory values
Y_4 : Insulin (uU/mL)	Serum insulin in plasma	`LBXIN`	Laboratory values
Y_5 : Folate (ng/mL)	Red blood cell folate	`LBDRFO`	Laboratory values

Table S1: Definition of outcomes for the models of interest fit to the National Health and Nutrition Examination Survey (NHANES), including the names of the variables in NHANES and tables from which they were sourced. Abbreviations of units: nanomoles per liter (nmol/L), beats per minute (bpm), milligrams per deciliter (mg/dL), microunits per milliliter (uU/mL), and nanograms per milliliter (ng/mL).

Covariate (Units)	Description	Variable	Source Data
X_1 : Calcium intake (mg)	24-hour cumulative intake	`DR1TCALC`	Dietary variables
X_2 : Caffeine intake (mg)	24-hour cumulative intake	`DR1TCAFF`	Examination data
X_3 : Total saturated fatty acids (gm)	24-hour cumulative intake	`DR1TSFAT`	Laboratory values
X_4 : Alcohol consumption (gm)	24-hour cumulative intake	`DR1TALCO`	Laboratory values
X_5 : Folate food (mcg)	24-hour cumulative intake	`DR1TFF`	Laboratory values

Table S2: Definition of nutrient intake covariates for the models of interest fit to the National Health and Nutrition Examination Survey (NHANES), including the names of the variables in NHANES and tables from which they were sourced. Abbreviations of units: milligram (mg), gram (gm), microgram(mcg).



X_1: Calcium	0.10	0.59	0.01	0.38
0.10	X_2: Caffeine	0.16	0.04	0.10
0.59	0.16	X_3: Saturated Fat	0.07	0.34
0.01	0.04	0.07	X_4: Alcohol	0.15
0.38	0.10	0.34	0.15	X_5: Food Folate

X_1^*: Calcium	0.08	0.46	0.00	0.30
0.08	X_2^*: Caffeine	0.14	0.05	0.07
0.46	0.14	X_3^*: Saturated Fat	0.04	0.26
0.00	0.05	0.04	X_4^*: Alcohol	0.14
0.30	0.07	0.26	0.14	X_5^*: Food Folate

Figure S5: Estimated correlation matrix between the **A)** error-free dietary intake exposures $X_1, \ldots, X_5$ (from the National Health and Nutrition Examination Survey [NHANES] dataset) and the **B)** error-prone dietary intake exposures $X_1^*, \ldots, X_5^*$ (simulated from the NHANES dataset).

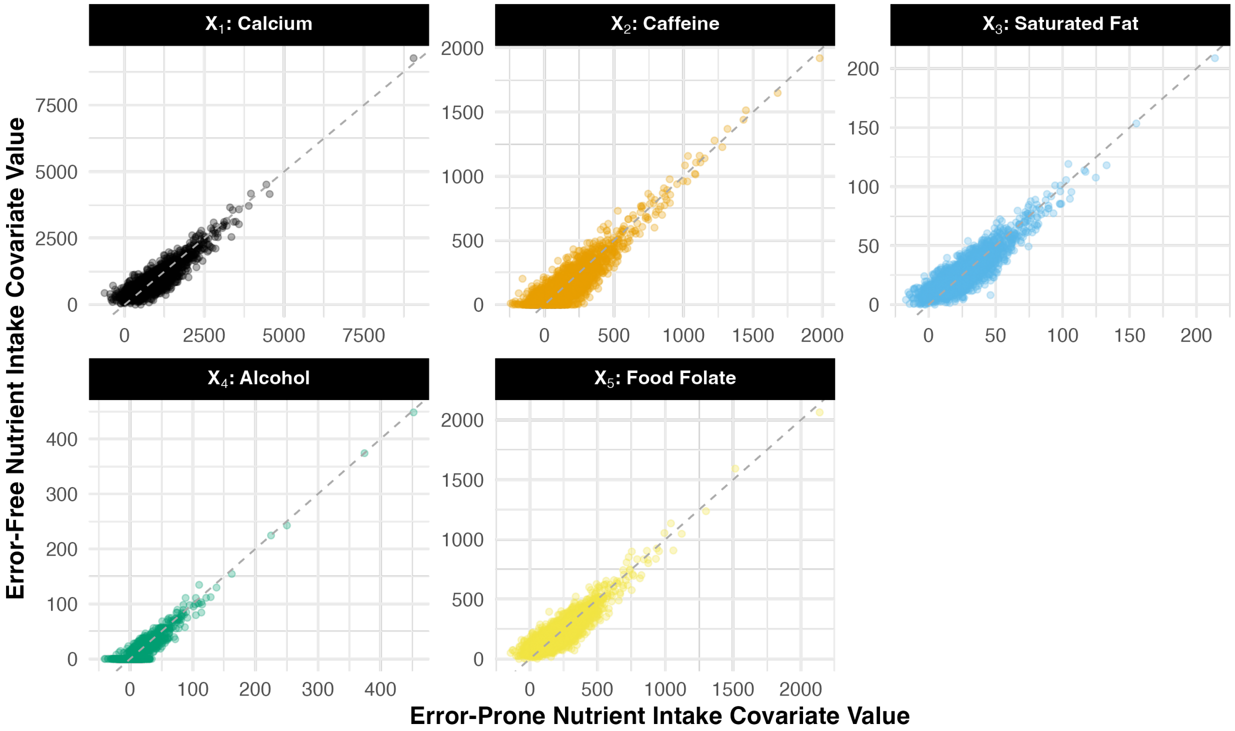


Figure S6: Comparison of error-free nutrient intake exposures $X_1, \dots, X_5$ (from the National Health and Nutrition Examination Survey [NHANES] dataset) with the simulated error-prone versions $X_1^*, \dots, X_5^*$. The dashed line denotes the line of equality (i.e., $X_j = X_j^*$).

					Race and Ethnicity (Reference = Mexican American)				Education Level (Reference = Less than 9th Grade			
	(Intercept)	Nutrient Intake	Female	Age	Other Hispanic	Non-Hispanic White	Non-Hispanic Black	Other Race (Incl. Multi-Racial)	9–11th Grade	High School Grad GED or Equiv.	Some College or AA Degree	College Graduate or Above
Model 1												
Gold Standard	21.4 (11.9, 30.9)	0.0 (0.0, 0.0)	12.3 (9.5, 15.1)	0.7 (0.6, 0.8)	4.0 (−2.7, 10.8)	14.6 (8.8, 20.4)	−4.2 (−11.1, 2.7)	8.0 (1.0, 15.0)	−7.9 (−16.8, 1.1)	0.09 (−7.72, 7.9)	4.7 (−3.1, 12.4)	7.4 (−0.3, 15.0)
SRS	23.3 (12.9, 33.7)	0.0 (0.0, 0.0)	11.9 (8.9, 15.0)	0.7 (0.6, 0.8)	3.9 (−2.9, 10.7)	14.8 (8.9, 20.6)	−4.4 (−11.3, 2.6)	7.9 (0.9, 15.0)	−7.8 (−16.8, 1.1)	0.1 (−7.7, 7.9)	4.9 (−2.9, 12.6)	7.5 (−0.1, 15.2)
ETS-X_1^*	18.6 (7.9, 29.3)	0.0 (0.0, 0.01)	12.6 (9.7, 15.5)	0.7 (0.7, 0.8)	4.3 (−2.5, 11.1)	14.5 (8.7, 20.3)	−4.3 (−11.3, 2.6)	7.9 (0.8, 14.9)	−8.1 (−17.1, 0.8)	0.2 (−7.7, 8.0)	4.6 (−3.1, 12.3)	7.3 (−0.3, 15.0)
ETS-PC_1^*	20.2 (10.2, 30.2)	0.0 (0.0, 0.01)	12.3 (9.5, 15.1)	0.7 (0.6, 0.8)	4.5 (−2.4, 11.3)	14.6 (8.8, 20.4)	−4.2 (−11.1, 2.7)	7.8 (0.7, 14.8)	−7.8 (−16.7, 1.2)	0.2 (−7.6, 8.0)	4.6 (−3.1, 12.3)	7.3 (−0.3, 14.9)
Model 2												
Gold Standard	72.05 (68.83, 75.27)	0.0 (0.0, 0.0)	3.13 (2.16, 4.09)	−0.12 (−0.15, −0.09)	3.22 (0.86, 5.59)	2.83 (0.79, 4.86)	1.81 (−0.61, 4.23)	3.47 (1.01, 5.93)	1.96 (−1.17, 5.08)	1.29 (−1.44, 4.02)	1.45 (−1.24, 4.14)	−1.27 (−3.93, 1.39)
SRS	71.59 (68.36, 74.83)	0.0 (0.0, 0.01)	3.33 (2.36, 4.31)	−0.12 (−0.15, −0.09)	3.3 (0.93, 5.67)	2.45 (0.39, 4.5)	2.16 (−0.28, 4.59)	3.17 (0.7, 5.64)	1.42 (−1.76, 4.6)	1.13 (−1.62, 3.87)	1.01 (−1.71, 3.72)	−1.58 (−4.26, 1.1)
ETS-X_1^*	71.96 (68.73, 75.18)	0.0 (−0.01, 0.0)	3.17 (2.2, 4.14)	−0.12 (−0.15, −0.09)	3.23 (0.86, 5.59)	2.81 (0.75, 4.86)	1.86 (−0.56, 4.28)	3.49 (1.02, 5.96)	1.96 (−1.17, 5.09)	1.3 (−1.43, 4.03)	1.43 (−1.28, 4.13)	−1.28 (−3.94, 1.39)
ETS-PC_1^*	71.9 (68.67, 75.13)	0.0 (0.0, 0.0)	3.19 (2.22, 4.15)	−0.12 (−0.15, −0.09)	3.26 (0.89, 5.63)	2.75 (0.7, 4.81)	1.88 (−0.54, 4.3)	3.45 (0.98, 5.91)	1.94 (−1.19, 5.07)	1.29 (−1.45, 4.02)	1.39 (−1.31, 4.09)	−1.3 (−3.96, 1.36)
Model 3												
Gold Standard	36.54 (32.61, 40.48)	0.01 (−0.03, 0.04)	8.82 (7.66, 9.98)	0.11 (0.08, 0.15)	1.56 (−1.25, 4.36)	3.62 (1.21, 6.03)	6.38 (3.51, 9.24)	3.45 (0.54, 6.36)	1.43 (−2.28, 5.14)	2.95 (−0.29, 6.19)	3.13 (−0.06, 6.33)	6.42 (3.27, 9.58)
SRS	36.26 (32.09, 40.42)	0.02 (−0.04, 0.08)	8.89 (7.67, 10.11)	0.11 (0.08, 0.15)	1.52 (−1.28, 4.33)	3.51 (1.03, 5.98)	6.31 (3.42, 9.19)	3.32 (0.35, 6.29)	1.36 (−2.38, 5.09)	2.95 (−0.29, 6.18)	3.12 (−0.08, 6.31)	6.42 (3.27, 9.58)
ETS-X_1^*	36.41 (32.21, 40.61)	0.01 (−0.05, 0.07)	8.83 (7.65, 10.01)	0.11 (0.08, 0.15)	1.58 (−1.23, 4.39)	3.6 (1.18, 6.02)	6.37 (3.5, 9.24)	3.43 (0.51, 6.35)	1.42 (−2.29, 5.13)	2.94 (−0.3, 6.18)	3.11 (−0.09, 6.31)	6.41 (3.25, 9.57)
ETS-PC_1^*	35.8 (31.63, 39.97)	0.03 (−0.02, 0.08)	8.95 (7.76, 10.13)	0.11 (0.08, 0.15)	1.67 (−1.14, 4.48)	3.5 (1.07, 5.92)	6.35 (3.49, 9.22)	3.33 (0.4, 6.25)	1.42 (−2.28, 5.13)	2.93 (−0.3, 6.17)	3.07 (−0.12, 6.26)	6.41 (3.26, 9.57)
Model 4												
Gold Standard	18.5 (12.88, 24.11)	−0.03 (−0.07, 0.01)	−2.57 (−4.27, −0.87)	−0.01 (−0.06, 0.04)	−0.2 (−4.34, 3.93)	−3.4 (−6.95, 0.16)	−2.95 (−7.19, 1.28)	−4.37 (−8.66, −0.07)	1.72 (−3.75, 7.19)	1.78 (−3, 6.55)	0.89 (−3.81, 5.59)	−1.29 (−5.94, 3.37)
SRS	18.62 (12.96, 24.28)	−0.06 (−0.15, 0.03)	−2.67 (−4.4, −0.94)	−0.01 (−0.06, 0.05)	−0.26 (−4.42, 3.9)	−3.34 (−6.94, 0.25)	−2.74 (−7.04, 1.56)	−4.4 (−8.74, −0.07)	1.9 (−3.63, 7.42)	1.77 (−3.03, 6.56)	0.99 (−3.73, 5.72)	−1.01 (−5.71, 3.7)
ETS-X_1^*	18.6 (12.96, 24.24)	−0.08 (−0.18, 0.02)	−2.81 (−4.58, −1.05)	−0.01 (−0.06, 0.04)	−0.07 (−4.22, 4.08)	−3.24 (−6.82, 0.33)	−2.77 (−7.03, 1.49)	−4.3 (−8.61, 0.01)	2.32 (−3.24, 7.89)	2.01 (−2.78, 6.81)	1.12 (−3.6, 5.84)	−1.09 (−5.76, 3.59)
ETS-PC_1^*	18.94 (13.28, 24.6)	−0.07 (−0.12, −0.02)	−2.87 (−4.6, −1.15)	−0.01 (−0.06, 0.04)	−0.3 (−4.45, 3.85)	−3.34 (−6.91, 0.22)	−2.85 (−7.09, 1.4)	−4.43 (−8.75, −0.12)	2.15 (−3.35, 7.65)	1.75 (−3.04, 6.53)	1.06 (−3.65, 5.78)	−1.05 (−5.71, 3.62)
Model 5												
Gold Standard	312.23 (246.93, 377.53)	−0.01 (−0.08, 0.05)	26 (6.86, 45.14)	3.46 (2.88, 4.04)	−15.22 (−61.65, 31.2)	48.04 (8.19, 87.89)	−56.95 (−104.42, −9.49)	−7.6 (−55.76, 40.56)	−25.77 (−87.08, 35.54)	1.24 (−52.28, 54.77)	25.22 (−27.5, 77.95)	19.47 (−32.82, 71.76)
SRS	279.61 (208.5, 350.72)	0.11 (−0.01, 0.23)	32.43 (12.37, 52.48)	3.51 (2.93, 4.09)	−7.93 (−55.04, 39.19)	52.1 (11.96, 92.24)	−51.17 (−99.11, −3.22)	−4.37 (−52.84, 44.11)	−29.5 (−91.15, 32.15)	−0.19 (−53.81, 53.44)	24.63 (−28.18, 77.45)	10.81 (−42.33, 63.95)
ETS-X_1^*	334.49 (262.3, 406.69)	−0.09 (−0.2, 0.03)	22.2 (2.37, 42.04)	3.42 (2.84, 4)	−18.08 (−64.7, 28.54)	45.24 (5.2, 85.28)	−58.92 (−106.46, −11.37)	−5.32 (−53.71, 43.08)	−25.98 (−87.27, 35.32)	−1.84 (−55.54, 51.86)	22.18 (−30.7, 75.05)	20.77 (−31.46, 72.99)
ETS-PC_1^*	319.1 (249.43, 388.77)	−0.04 (−0.14, 0.07)	24.7 (5, 44.4)	3.45 (2.86, 4.03)	−15.53 (−62, 30.93)	47.23 (7.27, 87.2)	−57.19 (−104.66, −9.71)	−6.6 (−54.92, 41.72)	−25.56 (−86.89, 35.76)	1.09 (−52.46, 54.65)	25.25 (−27.47, 77.98)	20.16 (−32.18, 72.5)

Table S3: Estimates from all fitted models from the application to error-prone dietary intake exposures in the National Health and Nutrition Examination Survey (NHANES) data. The Gold Standard estimates used only the error-free $\boldsymbol{X}$ from NHANES for all individuals. All others used $\boldsymbol{X}$ from a subset of $n = 250$ individuals and imputed them from $\boldsymbol{X}^*$ and $\boldsymbol{Z}$ for the rest. Three different validation study designs were considered: simple random sampling (SRS), extreme tail sampling on X_1^* (ETS-X_1^*), and extreme tail sampling on the first principal component (ETS-PC_1^*).