

# Endogenous Estrogen Metabolites and Gastric Cancer Risk Among Postmenopausal Women

## Statistical Analysis and Results

*Isaac Zhao*

*01/17/19 Version 4*

# Materials and Methods

## Study Design

Table 1: Demographics Table

	Cohort								Case Control				Overall	
	Iran		Korea (KMCC)		Germany		Korea (SNU)		Japan					
	Case	Control	Case	Control	Case	Control	Case	Control	Case	Control	Case	Control	Case	Control
Sample Size (n)	43	81	29	54	10	11	146	137	15	15	243	298		
Age (mean (sd))	62.3 (8.2)	62.5 (8)	63.8 (6.3)	61.7 (6.8)	63.7 (5.4)	65.7 (5.5)	64.5 (8)	57.3 (5.9)	72.1 (6.1)	71.9 (5.7)	64.5 (7.9)	60.6 (7.6)		
BMI (mean (sd))	27.7 (5.7)	27.3 (5.5)	24.5 (3)	24.5 (3.6)	30 (5.5)	29.7 (4.6)	23.9 (3.2)	24.9 (3.3)	22.9 (3.4)	21 (2.4)	24.7 (4.3)	25.4 (4.4)		
Ever smoked (n (%))	3 (7)	2 (2.5)	3 (10.3)	4 (7.4)	0	2 (18.2)	8 (5.5)	6 (4.4)	2 (13.3)	1 (6.7)	16 (6.6)	15 (5)		
Ever drank alcohol (n (%))	0	0	6 (20.7)	11 (20.4)	2 (20)	5 (45.5)	30 (20.5)	47 (34.3)	2 (13.3)	3 (20)	40 (16.5)	66 (22.1)		
Any educational degree (n (%))	1 (2.3)	10 (12.3)	19 (65.5)	33 (61.1)	10 (100)	10 (90.9)	130 (89)	122 (89.1)	NA	NA	160 (65.8)	175 (58.7)		
Relative with gastric cancer (n (%))	1 (2.3)	1 (1.2)	NA	NA	0	1 (9.1)	28 (19.2)	15 (10.9)	1 (6.7)	5 (33.3)	30 (12.3)	22 (7.4)		

Note:

NA = Data not available

Incident gastric cancer and two case-control studies of early-stage cancer were used for analysis. For the incident gastric cancer set, pre-diagnostic urine samples from all available postmenopausal (or age 60+ years) women diagnosed with gastric cancer and incidence-density matched controls from three prospective cohort studies (Golestan Cohort (Iran), Korean Multicenter Cancer Cohort, and ESTHER Cohort (Germany)) were tested. Iran and Korea (KMCC) were approximately 1:2 ratio of cases vs. controls. For the early-stage case-control gastric cancer set, urine samples from postmenopausal (or age 60+ years) women diagnosed with early-stage gastric cancer (AJCC clinical stages 1A [T1, N0, M0] or 1B [T1, N1, M0 or T2, N0, M0]) and 1:1 age-matched (+/- 5 years) controls from established case-control studies in Japan and Korea (Seoul Gastric Cancer Study) were tested.

Postmenopausal women with gastric cancer were matched by age to gastric cancer-free controls. Women who ever used post-menopausal hormone replacement were excluded since we were specifically interested in the effects of endogenous estrogens. Premenopausal women were also excluded since estrogen levels vary over the menstrual cycle, greatly complicating interpretation of measurements; in any case, gastric cancer is rare prior to age 50 years. On the other hand, restriction of the case-control set to stage 1 gastric cancer will limit the risk of reverse causality.

Urine specimens were collected at enrollment in prospective studies and pre-treatment in case-control studies and continuously cryopreserved at -70/-80 degrees Celcius until analysis.

Variables considered for covariate adjustment to estimate odds ratio effect sizes in the analysis portion were age, BMI, smoking, alcohol, education, and if the subject's relative had gastric cancer. Categories were simplified into never or ever to account for differences in how each variable was recorded for each study. This avoided ambiguous temporality and count of substance use for variables such as smoking and alcohol. Dividing into only two categories also reduced extreme sample size sparsity problems such as educational degree.

## Laboratory Assay

Stable isotope dilution liquid chromatography-tandem mass spectrometry (LC-TMS) was used at the NCI Laboratory of Proteomics and Analytical Technologies, MD to simultaneously measure the total concentration of 2 parent estrogens (estrone and estradiol) and 13 estrogen metabolites (2-hydroxyestrone, 2-methoxyestrone, 2-hydroxyestradiol, 2-methoxyestradiol, 2-hydroxyestrone-3-methyl ether, 4-hydroxyestrone, 4-methoxyestrone, 4-methoxyestradiol, 16 $\alpha$ -hydroxyestrone, 16-ketoestradiol, estriol, 17-epiestriol, and 16-epiestriol) in an aliquot of 500  $\mu$ L urine assay for each participant. In urine, parent estrogens and their metabolites are present primarily in conjugated form. Estrogen concentrations in spot urine samples were normalized to creatinine levels in order to adjust for variation in urinary volume.

# Results

## Covarying Relationships Between Metabolites

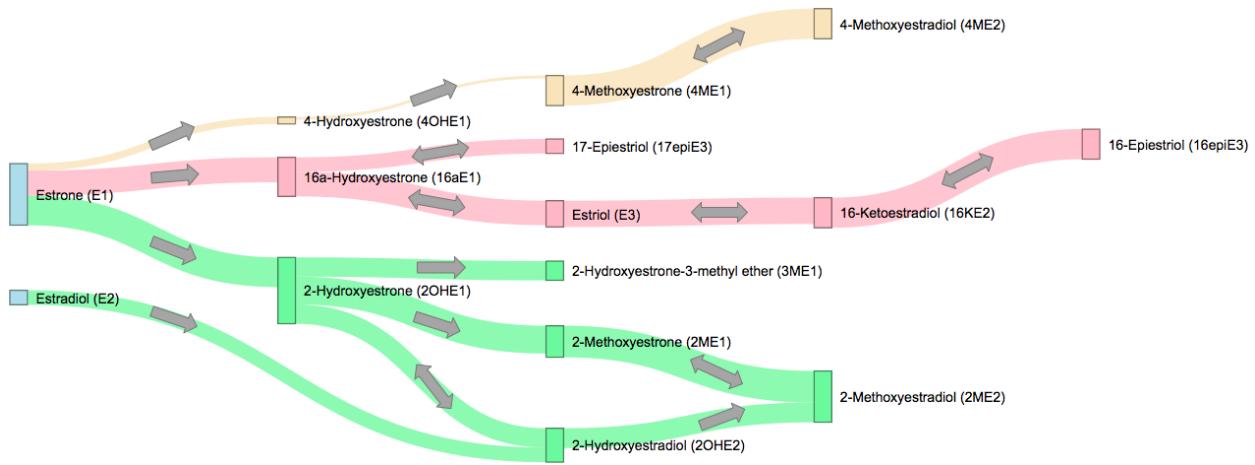


Figure 1: Estrogen metabolite hydroxylation pathway sankey diagram. Line thickness proportional to spearman correlations between substrates and products on hydroxylation pathways (2-OH, green; 4-OH, tan; 16-OH, pink), ranging from 0.06 to 0.67. 4-Hydroxyestradiol and 16a-Hydroxyestrone not shown due to extremely low concentrations.

Correlations between metabolites were relatively equal throughout the pathways except for the 4-Hydroxylation pathway connection with the 4OHE1 metabolite. 4OHE1 may have weak correlations due to an artifact of lab mis-measurement (many NF values).



Figure 2: Metabolite correlation matrix. Redder regions indicate stronger positive correlations. Scatterplot and histogram on log scale.

The relationship between estrogen metabolites were computed using spearman correlations on log transformed concentrations. Spearman correlation was chosen to account for nonlinear relationships among the metabolites that are more resistant to outliers and influential points. Log transformations were used to better visualize the paired scatterplots due to highly skewed distributions.

The histograms show that the estrogen metabolite concentration distributions were still highly right tail skewed even after log transformations. Extremely high outliers may be an artificial result of laboratory mis-measurement and should be cautioned to be used as true values. The correlation matrix shows that all estrogens were positively associated with each other with highly significant p-values, even after adjusting for bonferroni corrections of significance levels. In other words, an individual with higher levels of one metabolite will in general have higher levels of all other metabolites as well. The strongest correlation was among parent estrogens E1 and E2, which was expected. The lowess smooth curve regression line with span 2/3 (in dark blue on the scatterplot) shows that on the log scale, there did not appear to be any major nonlinear higher order interactions (i.e. relationships were mostly linear on the log scale). Thus analysis was conducted without inclusion of polynomial terms. The span parameter was chosen as a general rule of thumb to apply to all metabolites and may not necessarily produce the best residual sum of squares, although it summarizes the true shape of relationships decently well.

## Batch Effect and Not Found Data

Table 2: Table of estrogen batch information

Batch	n	Total estrogen NF (n)	Cases (n)	Controls (n)	Study
1	36	94	25	11	Germany and Korea (KMCC)
2	36	28	14	22	Korea (KMCC)
3	32	14	0	32	Korea (KMCC)
4	30	52	15	15	Japan
5	36	10	13	23	Iran
6	36	35	10	26	Iran
7	36	38	13	23	Iran
8	16	19	7	9	Iran
9	38	50	37	1	Korea (SNU)
10	39	64	39	0	Korea (SNU)
11	35	78	31	4	Korea (SNU)
12	35	44	32	3	Korea (SNU)
13	36	50	4	32	Korea (SNU)
14	37	86	2	35	Korea (SNU)
15	32	12	0	32	Korea (SNU)
16	30	9	1	29	Korea (SNU)

Note: Sample sizes are after QC sample removal

Batch 3, 13, 14, 15, and 16 consisted of almost entirely controls; while batch 9, 10, 11, and 12 consisted of almost entirely cases. Since some batches tended to be significantly higher/lower than other batches (see QC graph), there may be evidence of a batch effect. In other words, differences for cases vs. controls in metabolite concentrations or lack thereof for Korea (KMCC) and Korea (SNU) may be artificial due to batch differences rather than a true association.

In general, different batches had different number of NF values. The batches with the lowest number of NF values also were the batches with mostly controls (batch 3, 15, and 16). Thus, controls were more represented than planned. Secondly, batch 1 containing Germany had substantially higher NF values for metabolite concentrations. Thus, Germany was slightly more under-represented than planned.

Table 3: Table of values not found (NF) for metabolites and creatinine

Variable	NF (n)	NF (%)
E1	25	4.6
E2	23	4.3
4OHE1	183	33.9
4ME1	34	6.3
4ME2	23	4.3
2OHE1	28	5.2
3ME1	24	4.4
2OHE2	61	11.3
2ME1	22	4.1
2ME2	23	4.3
16aE1	77	14.3
17epiE3	29	5.4
E3	54	10
16KE2	42	7.8
16epiE3	35	6.5
creatinine	20	3.7

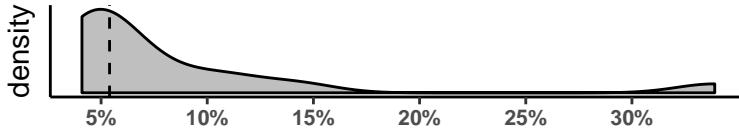


Figure 3: Distribution of NF percent among all metabolites. Vertical line indicates median.

The median percent of NF values for estrogen metabolites was about 5%, with 4OHE1 having the highest percentage of NF values. NF values for metabolites were caused by measurements not being identified as targeted peaks due to mis-alignment of co-eluting peaks and strong unsuppression from the LC-TMS device. The peaks could not be fully re-integrated nor enriched even after re-processing. NF values for creatinine were due to low sample volumes. The NF values were not imputed due to the unknown relationship that mis-measurement of the LC-TMS device had with the probability of a measurement to be un-recorded. In other words, it could not be determined if the distribution of NFs were independent of laboratory procedures which would bias predicted results if true. Furthermore, even if assumptions hold true for imputation, prediction models were not accurate enough to distinguish concentrations between tertile level cutoffs for the NF values (best prediction models for each metabolite were at most 60-70% accuracy when predicting tertiles and mean absolute error higher than the difference between tertile cutpoints per estrogen when predicting absolute concentrations).

All estrogen batches were re-analyzed by the laboratory which allowed us to compare non-overlapping missing values. (i.e. NF for a sample in the first analysis but not the second analysis), which determined that NFs did not appear to be correlated with how high or low the concentrations were. NF values for a sample in both batch analysis could not be determined if the nature was due to being below the LLOD range or not. Thus, the standard approach of replacing all NF values with half lower limit of quantification (LLOQ) or lower limit of detection (LLOD) would be inappropriate. Since NF was determined to be independent of measured variables, (i.e. NF status had nothing to do with concentrations themselves) complete cases (i.e. removing all missing values) was used accordingly. Furthermore, since the nature of NFs could not be fully determined, using complete cases was the safest and simplest approach. However, it should be noted that results that are borderline significant at the  $\alpha=0.05$  level may change if NF values were treated differently or corrected instead. Sensitivity analysis or “worst-case-scenario” imputation (i.e. replacing all NF values with all high values or all low values) is another option to be considered if the laboratory cannot amend accordingly.

## Quality Control (QC) Analysis

Laboratory personnel was blinded to the case-control status of sample donors. A quality control (QC) set of 20 masked duplicate samples plus 4 additional laboratory control replicates from subjects with high available volumes representative of all studies was performed. Coefficient of variation (CV) and intraclass correlation coefficient (ICC) was calculated for standardized estrogen metabolite concentrations for the QC samples to assess within- and between-batch variations for assay reliability. The formula used for ICC on the original and log scale was  $\frac{\sigma_{bs}^2}{\sigma_{bs}^2 + \sigma_{bb}^2 + \sigma_{ws}^2}$ , the formula used for CV on the original scale was  $\frac{\sqrt{\sigma_{ws}^2 + \sigma_{bb}^2}}{\mu}$ , and the formula used for CV on the log scale was  $e^{\sigma_{ws}} - 1$ ; where  $\sigma_{bs}^2$  = variance between subject,  $\sigma_{bb}^2$  = variance between batch, and  $\sigma_{ws}^2$  = variance within subject. The variance components were computed using a two stage multilevel model with varying intercept for ID and batch. Log transformations were necessary to adjust for high skew and large outliers to improve normality of distributions to meet assumptions necessary for the computation of variance which would otherwise lead to exaggerated variance components.

NF values were treated as missing and were not included in calculations since any estimated numeric representations may give biased results. In other words, replacing missing values with expected values from a prediction model would favor a better reproducibility measurement). The lower limit of quantitation (LLOQ) for each analyte was 0.04 ng/mL. The median for each sample was well above the LLOQ and thus was not a major concern. However, samples with lower concentrations (i.e. postmenopausal sample) had higher measurement error since the measurement device is less sensitive at these levels. Thus, CV was also calculated for the post-menopausal sample by itself ; otherwise the ICC and CV measurements from mixing all four samples together would be a mis-representative measure of the study population of just postmenopausal women.

## Laboratory Control QC Analysis

Table 4: Laboratory control QC samples reproducibility calculations

Estrogen	Original Scale						Log Transformed				
	$\sigma_{bs}^2$	$\sigma_{bb}^2$	$\sigma_{ws}^2$	Mean	ICC	CV	$\sigma_{bs}^2$	$\sigma_{bb}^2$	$\sigma_{ws}^2$	ICC	CV
E1	25.8	0.4	2.6	8.0	89.4	21.7	0.4	0.0	0.0	89.1	20.1
E2	6.4	0.1	0.4	2.9	92.0	25.8	0.8	0.0	0.0	93.5	21.2
4OHE1	843330.8	0.0	2524.4	339.3	99.7	14.8	7.1	7.1	0.1	49.6	41.4
4ME1	0.0	1.2	1.8	0.8	0.0	231.0	0.0	0.9	0.5	0.0	103.0
4ME2	0.0	0.0	0.0	0.1	17.5	139.9	0.1	0.9	0.7	7.7	129.3
2OHE1	8631.2	30252.9	46181.0	57.1	10.1	484.6	1.0	1.0	0.9	34.2	156.3
3ME1	0.0	0.0	0.0	0.5	19.1	49.1	0.1	0.1	0.1	16.2	42.3
2OHE2	317.5	0.0	146.1	4.0	68.5	302.5	3.3	0.2	1.2	69.3	204.4
2ME1	3.2	10.2	120.9	5.0	2.4	230.6	0.3	0.1	0.5	29.0	107.0
2ME2	0.0	0.0	0.0	0.5	36.1	54.1	0.2	0.1	0.1	42.9	40.1
16aE1	6.2	0.1	0.8	2.7	87.4	35.5	1.4	0.1	0.3	78.2	68.6
17epiE3	2.5	0.1	0.3	1.2	84.6	57.7	2.0	0.5	0.7	62.7	124.1
E3	87.8	83.2	95.9	14.9	32.9	89.8	1.7	0.5	0.2	71.5	48.7
16KE2	7.4	0.1	0.7	2.5	90.1	36.0	1.4	0.0	0.1	94.1	30.3
16epiE3	4.9	2.5	3.4	2.8	45.7	85.6	1.0	0.3	0.3	65.4	66.4

*Note:*

bs = between subject, bb = between batch, ws = within subject

Units in pmol/mg creatinine

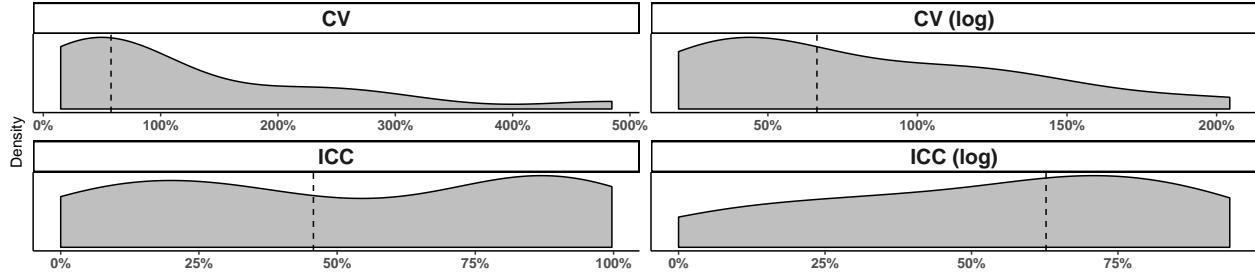


Figure 4: Distribution of CV and ICC for all metabolites for laboratory control QC samples. Vertical line indicates median.

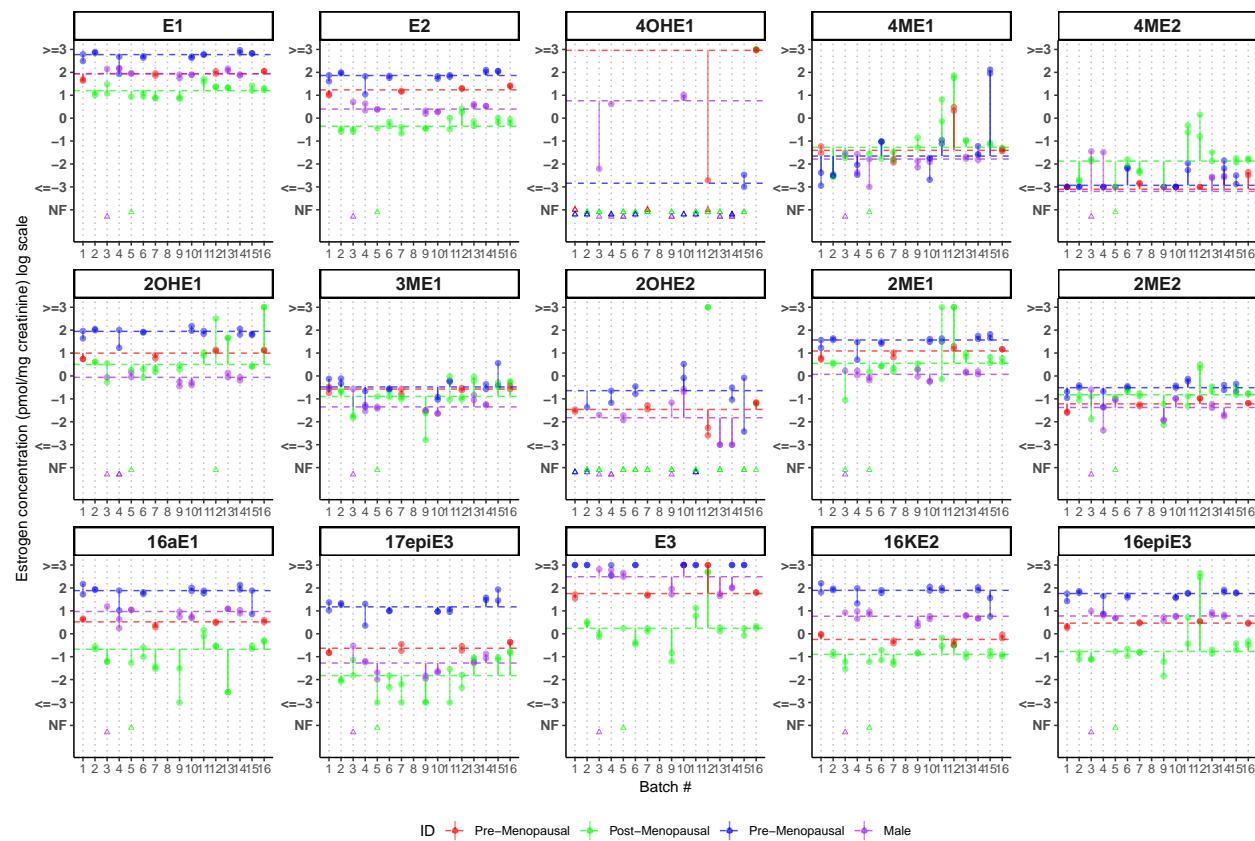


Figure 5: Batch (x-axis) plotted against measured standardized concentrations.

The laboratory control QC samples consisted of 4 distinct urine sample types: 2 pre-menopausal ( $n=8$  and  $16$ ), 1 post-menopausal ( $n=22$ ), and 1 male urine ( $n=14$ ), totaling 60 observations. In order to assess within and between variation, four samples were allocated within in each batch, some of which had duplicate sample types in the same batch. All laboratory control QC samples were removed when assessing the association between gastric cancer and estrogens.

Most metabolites had ICCs  $\geq 60\%$  with most CVs  $\leq 60\%$ , indicating moderate reproducibility generalizable across all batches. It should be noted that although 4OHE1 had a high ICC, it may be due to a low sample size and shouldn't be trusted (i.e. the only two samples representative of sample B - postmenopausal were the samples with extreme concentrations that are most likely inaccurate). The QC samples plotted on the log scale illustrated that a few measurements for the same sample were many fold different than their counterparts. Two batches were specifically problematic, often having measurements many fold higher than

other batches across different estrogens. Concentrations of these batches were not adjusted downwards since the sample size of QC samples in these batches were not sufficient enough for justification. Within subject variation was so significant that when concentrations were classified into tertiles, many replicates were in different tertile categories.

## QC Sample Removal

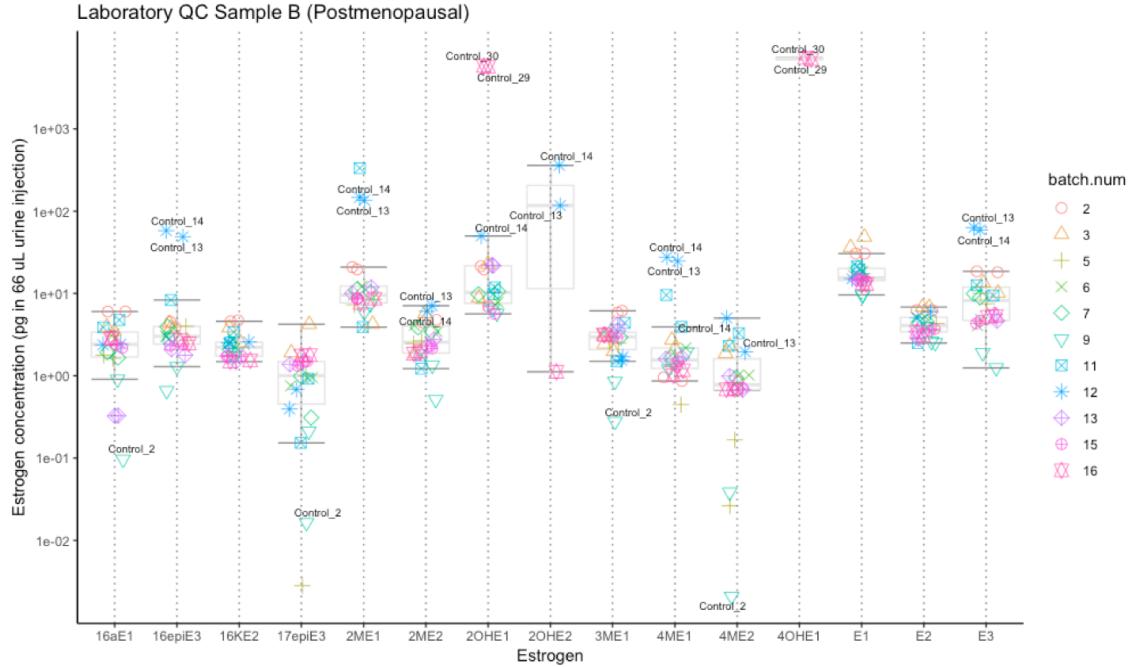


Figure 6: Problematic QC samples highlighted for postmenopausal

Table 5: Laboratory control QC samples reproducibility calculations

Estrogen	ICC	CV	ICC (log)	CV (log)
E1	88.6	21.5	89.0	21.2
E2	91.7	25.0	95.5	19.5
4OHE1	0	148.9	0	48.8
4ME1	3.4	250.9	0	103.0
4ME2	14.9	114.1	6.9	74.8
2OHE1	86.1	31.8	86.9	31.6
3ME1	22.4	47.7	27.9	36.8
2OHE2	40.6	79.3	43.4	92.2
2ME1	0	262.2	46.1	90.4
2ME2	51.3	32.3	41.7	36.6
16aE1	86.4	34.5	85.3	51.2
17epiE3	83.5	56.0	68.6	113.3
E3	30.3	90.8	76.7	53.3
16KE2	89.3	35.1	94.1	29.4
16epiE3	91.4	25.1	90.5	32.5

Note: QC sample 2, 13, 14, 29, 30 removed

According to figure 5, the postmenopausal sample was the most problematic, specifically with batch 12 and batch 16. Figure 6 highlights the specific samples of batch 12 and 16 from postmenopausal QC sample to identify which samples were giving poor ICC and CV readings. These samples were removed from QC calculations to see how they affects ICC and CV. These two changes fixed the extreme skewness of the QC dataset and thus the ICC and CV on the original scale were much closer to that of the ICC and CV on the log scale. Using this adjustment, the ICC for 2OHE1 and 16epiE3 became over >85% compared to an ICC of 10% and 46% respectively before adjustment. The ICC and CV for other metabolites did not change much by this adjustment. It should be noted that removing extreme outliers and batches for the entire dataset is unjustifiable since there is not enough strong evidence to support this claim.

### Korea SNU QC Analysis

Table 6: Korea SNU QC samples reproducibility calculations

Estrogen	Original Scale					Log Transformed					
	$\sigma_{bs}^2$	$\sigma_{bb}^2$	$\sigma_{ws}^2$	Mean	ICC	CV	$\sigma_{bs}^2$	$\sigma_{bb}^2$	$\sigma_{ws}^2$	ICC	CV
E1	5.0	12.3	11.0	3.9	17.6	123.5	0.4	0.2	0.3	47.6	71.8
E2	0.5	0.9	0.6	0.7	25.2	167.1	0.7	0.1	0.3	64.3	69.0
4OHE1	0.1	3.3	21.2	2.3	0.5	217.7	0.8	0.9	1.1	27.4	189.5
4ME1	1.8	0.0	1.3	1.3	59.3	86.9	1.2	0.5	0.7	48.0	135.5
4ME2	0.0	0.0	0.1	0.3	41.1	97.8	1.0	0.0	0.5	66.6	102.0
2OHE1	4149.5	0.0	4589.3	21.4	47.5	316.7	1.6	0.3	0.4	71.1	82.0
3ME1	0.2	0.2	1.2	0.5	13.7	222.2	0.6	0.5	0.9	29.9	154.1
2OHE2	0.0	0.7	0.1	0.4	0.0	233.6	1.1	1.3	0.7	35.1	136.1
2ME1	3.8	0.0	11.4	2.7	25.2	122.6	0.7	0.0	0.4	62.0	94.1
2ME2	0.0	0.2	0.2	0.6	0.0	103.6	0.2	0.5	0.4	17.1	82.5
16aE1	0.4	0.2	0.4	0.7	38.0	104.5	0.6	0.1	0.2	66.8	63.1
17epiE3	1.0	1.2	1.7	0.9	25.2	185.7	2.1	0.9	0.7	56.4	131.4
E3	215.4	496.2	834.7	17.6	13.9	206.8	0.5	2.8	0.9	12.4	159.7
16KE2	0.2	0.3	0.4	0.6	20.3	145.9	0.5	0.0	0.5	51.9	96.2
16epiE3	4.5	0.4	2.0	1.7	65.3	89.7	1.3	0.0	0.3	80.0	77.8

Note:

bs = between subject, bb = between batch, ws = within subject

Units in pmol/mg creatinine

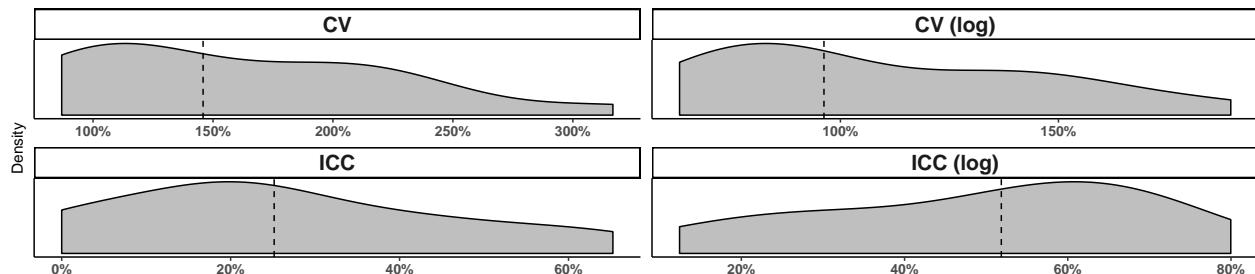


Figure 7: Distribution of CV and ICC for all metabolites for Korea SNU QC samples. Vertical line indicates median.

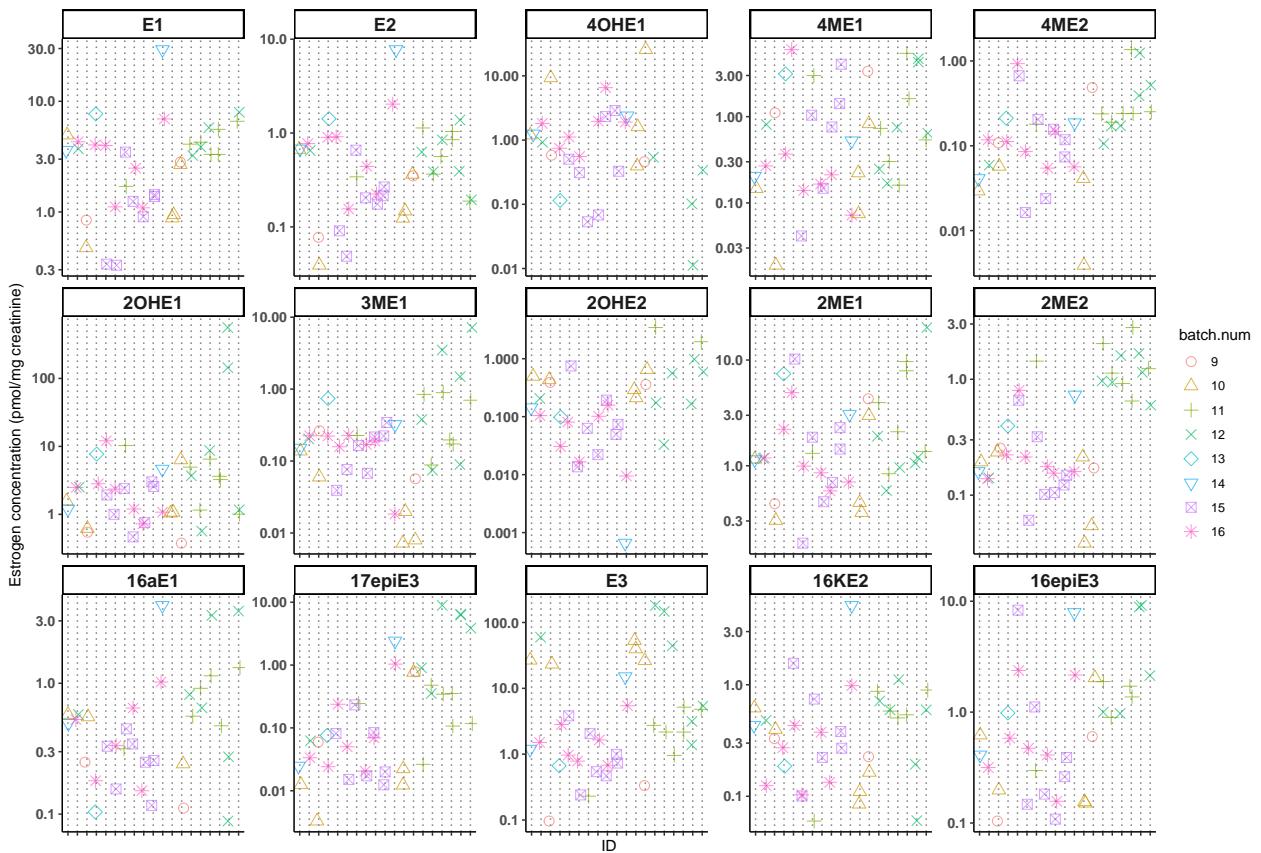


Figure 8: 19 Korea SNU QC sample IDs (x-axis) plotted against measured standardized concentrations. Batch identification indicated by color and shape.

The Korea SNU replicates consisted of 19 unique samples each with two measurements that were either both in the same batch or across different batches, totaling 38 observations. One sample from each duplicate was randomly kept while the other pair removed when assessing the association between gastric cancer and estrogens. Most metabolites had ICCs  $\geq 50\%$  with most CVs  $\leq 60\%$ , indicating moderately inconsistent measurements. Replication results for Korea SNU duplicates was worse than over all batches assessed with laboratory control samples. Many measurements for the same sample were many fold different than their counterparts. Thus, it should be noted that metabolite measurements for Korea SNU may not be reproducible and caution is advised.

### Combined QC Analysis

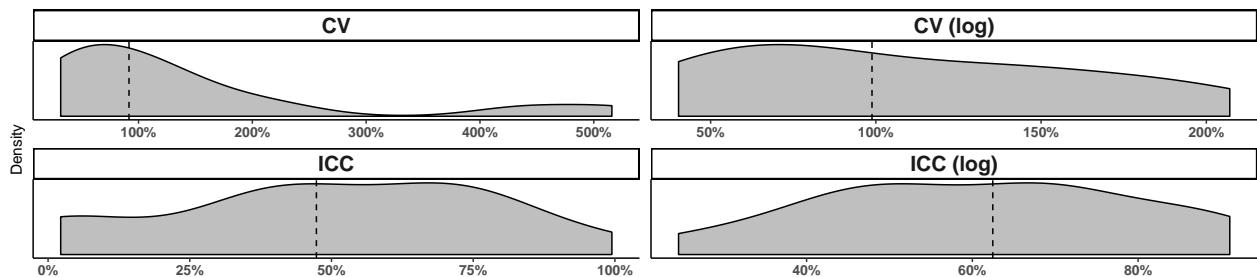


Figure 9: Distribution of CV and ICC for all metabolites for combined QC samples. Vertical line indicates median.

Table 7: Combined Laboratory QC and Korea SNU QC samples reproducibility calculations

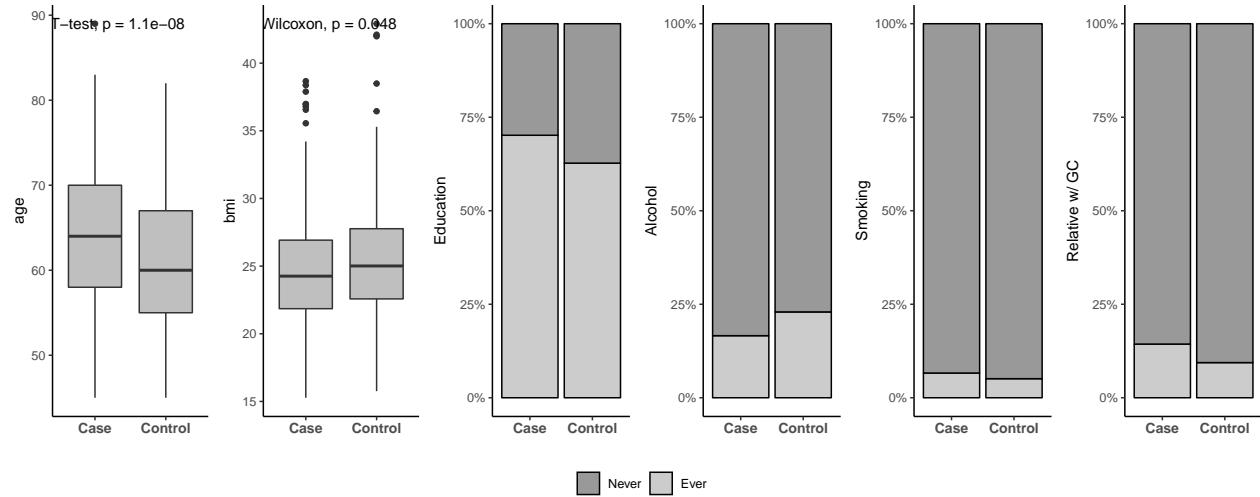
Estrogen	Original Scale						Log Transformed				
	$\sigma_{bs}^2$	$\sigma_{bb}^2$	$\sigma_{ws}^2$	Mean	ICC	CV	$\sigma_{bs}^2$	$\sigma_{bb}^2$	$\sigma_{ws}^2$	ICC	CV
E1	17.2	0.4	5.7	6.5	73.7	38.3	0.8	0.0	0.1	86.2	40.3
E2	2.5	0.1	0.5	2.0	79.5	39.7	1.3	0.0	0.1	91.1	41.0
4OHE1	168984.3	0.0	901.8	95.0	99.5	31.6	3.0	1.7	1.3	50.4	207.1
4ME1	0.1	0.4	2.5	1.0	3.8	176.2	0.8	0.6	0.8	36.7	138.1
4ME2	0.0	0.0	0.0	0.2	47.3	115.1	0.5	0.7	0.8	24.6	140.6
2OHE1	1280.8	3106.2	44931.1	42.5	2.6	515.9	1.0	0.3	1.1	41.6	190.3
3ME1	0.5	0.0	0.4	0.5	55.4	122.8	1.0	0.1	0.4	69.1	87.9
2OHE2	46.8	1.0	88.9	2.2	34.2	437.8	1.7	0.4	1.0	54.5	176.3
2ME1	1.9	2.5	82.1	4.1	2.2	225.5	0.6	0.1	0.5	48.0	109.3
2ME2	0.1	0.0	0.1	0.5	40.4	69.3	0.5	0.1	0.2	62.5	60.6
16aE1	2.0	0.1	0.7	2.0	70.6	46.3	1.0	0.1	0.3	72.7	68.4
17epiE3	2.2	0.2	0.8	1.1	70.4	91.3	2.9	0.4	0.8	69.5	147.1
E3	317.6	177.5	330.4	16.0	38.5	140.9	1.2	0.8	0.5	48.2	98.8
16KE2	1.9	0.1	0.7	1.8	72.0	49.0	1.0	0.0	0.2	85.8	51.6
16epiE3	3.1	1.4	3.5	2.4	38.6	91.6	1.1	0.2	0.3	69.4	74.7

Note:

bs = between subject, bb = between batch, ws = within subject

Units in pmol/mg creatinine

## Assessing Association between Covariates and Gastric Cancer



Cases were on average an older age while controls had a higher BMI on average. Other covariates of education, alcohol, smoking, and relative with gastric cancer did not have major differences in distribution between cases and controls. Thus, it is expected that adjusting for age and BMI in the logistic model will ultimately reduce the effect size of the metabolite than if the metabolite was used in the model itself as a main effect. However, adjusting for other covariates should not affect the metabolite estimated effect size significantly, but may inflate variance if samples in one group are sparse. Since adjusting for these other covariates should theoretically have effect sizes approximately centered to the same degree as not adjusting for these covariates, it is justifiable to remove certain covariates from each study model to optimize the bias-variance tradeoff.

## Assessing Association between Estrogens and Gastric Cancer

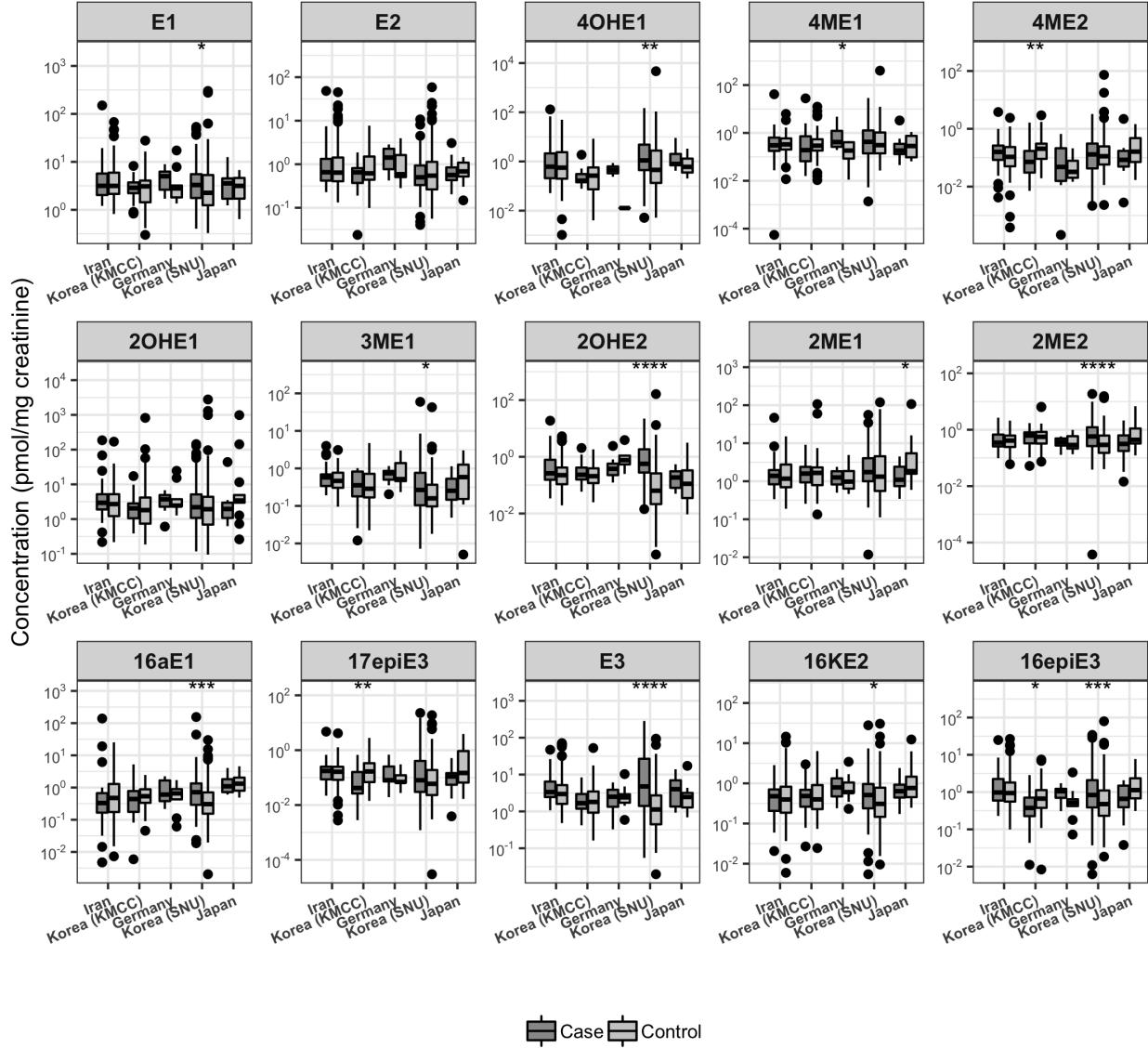


Figure 10: Study specific Wilcoxon-Mann-Whitney tests for cases vs. controls

Postmenopausal women with gastric cancer was compared with controls for both pre-diagnostic and early-stage case-control sample sets. Each study was analyzed separately to account for differences in the relationship between estrogens and gastric cancer by country of origin and study type. Each of the estrogen markers were separately analyzed to assess an association with gastric cancer. Wilcoxon-Mann-Whitney (also known as Mann-Whitney U) test was used to compare estrogen marker concentrations as a continuous value between cases and controls. Significance levels are indicated by stars above boxplots (\* :  $p \leq 0.05$ , \*\* :  $p \leq 0.01$ , \*\*\* :  $p \leq 0.001$ , \*\*\*\* :  $p \leq 0.0001$ ). According to the boxplot comparisons, over half of the metabolites had cases that had significantly higher concentrations for cases than controls at the  $\alpha=0.05$  level for Korea SNU. However, the three metabolites that were significant for Korea KMCC (4ME2, 17epiE3, 16epiE3) showed the opposite association (i.e. controls had significantly higher levels of these metabolites compared to cases). Since Korea SNU was significant for parent estrogen E1, adjusting for E1 when performing logistic regression for

this study may be appropriate since all other metabolites come from parent estrogens (i.e. other metabolites may be significantly associated with gastric cancer simply due to the fact that it was derived from the parent estrogen which is also significantly associated with gastric cancer).

Table 8: Estrogen concentration median and tertile cutpoint

	Cohort			Case Control	
	Iran	Korea (KMCC)	Germany	Korea (SNU)	Japan
	Median (Tertile)				
Estrone (E1)	3.15 (2.37, 4.31)	3.08 (1.68, 3.62)	2.79 (1.97, 3.11)	2.26 (1.53, 3.95)	3.16 (2.31, 4.21)
Estradiol (E2)	0.64 (0.49, 1.02)	0.62 (0.46, 1.21)	0.59 (0.53, 1.03)	0.55 (0.33, 0.83)	0.69 (0.56, 0.82)
4-Hydroxyestrone (4OHE1)	0.56 (0.36, 1.38)	0.26 (0.12, 0.4)	0.01 (0.01, 0.01)	0.45 (0.23, 1.24)	0.68 (0.36, 1.13)
4-Methoxyestrone (4ME1)	0.33 (0.24, 0.47)	0.29 (0.22, 0.41)	0.18 (0.14, 0.29)	0.31 (0.19, 0.69)	0.28 (0.13, 0.73)
4-Methoxyestradiol (4ME2)	0.11 (0.07, 0.18)	0.2 (0.12, 0.24)	0.03 (0.02, 0.06)	0.11 (0.07, 0.18)	0.18 (0.08, 0.38)
2-Hydroxyestrone (2OHE1)	2.7 (1.51, 4.26)	1.81 (1.06, 3.29)	2.44 (2.38, 2.9)	1.93 (0.9, 2.92)	3.43 (3.04, 4.67)
2-Hydroxyestrone-3-methyl ether (3ME1)	0.46 (0.36, 0.63)	0.29 (0.22, 0.62)	0.52 (0.46, 0.89)	0.16 (0.11, 0.25)	0.6 (0.33, 1.11)
2-Hydroxyestradiol (2OHE2)	0.22 (0.13, 0.36)	0.21 (0.14, 0.32)	0.78 (0.58, 1.05)	0.06 (0.03, 0.14)	0.12 (0.05, 0.2)
2-Methoxyestrone (2ME1)	1.17 (0.86, 1.93)	1.66 (0.93, 2.13)	0.98 (0.78, 1.49)	1.32 (0.71, 2.68)	1.84 (1.7, 3.64)
2-Methoxyestradiol (2ME2)	0.41 (0.29, 0.55)	0.56 (0.45, 0.65)	0.29 (0.22, 0.44)	0.29 (0.19, 0.45)	0.44 (0.37, 0.94)
16a-Hydroxyestrone (16aE1)	0.48 (0.29, 0.87)	0.54 (0.43, 0.66)	0.66 (0.55, 0.77)	0.31 (0.2, 0.51)	1.32 (0.82, 1.57)
17-Epiestriol (17epiE3)	0.16 (0.11, 0.22)	0.17 (0.1, 0.24)	0.07 (0.07, 0.12)	0.06 (0.03, 0.13)	0.15 (0.07, 0.21)
Estriol (E3)	2.97 (2.25, 4.68)	1.81 (1.13, 2.63)	2.41 (1.99, 2.79)	1.13 (0.74, 2.1)	2.46 (1.57, 3)
16-Ketoestradiol (16KE2)	0.4 (0.27, 0.65)	0.4 (0.25, 0.52)	0.6 (0.59, 0.84)	0.31 (0.2, 0.59)	0.76 (0.52, 0.84)
16-Epiestriol (16epiE3)	0.94 (0.72, 1.37)	0.67 (0.48, 0.94)	0.52 (0.46, 0.58)	0.48 (0.3, 0.81)	1.14 (0.88, 2.1)

*Note:*

Tertile (33% and 66% quantiles) determined by control subjects

Units in pmol/mg creatinine

Study-specific estrogen marker tertiles were defined by the concentration distribution in controls to account for the extremely skewed distributions and improve interpretability. Tertile categories were treated as a numeric entry rather than a factor type to assess overall linear associations (i.e.  $\beta$  coefficient for metabolite markers interpreted as risk difference from one tertile to the next). Since reproducibility was poor, it may be worth increasing the number of quantiles to quartiles such that variance of measurement will not have such drastic changes on categorization. Finer tuning into smaller categories would not restrict the measurement to be placed in the extreme low or high end. However, due to small sample size of some studies, this may not be a plausible approach. Korea SNU appeared to have lower estrogen concentrations than other studies, especially for the metabolites with high concentrations overall (E1, 2OHE1, E3).

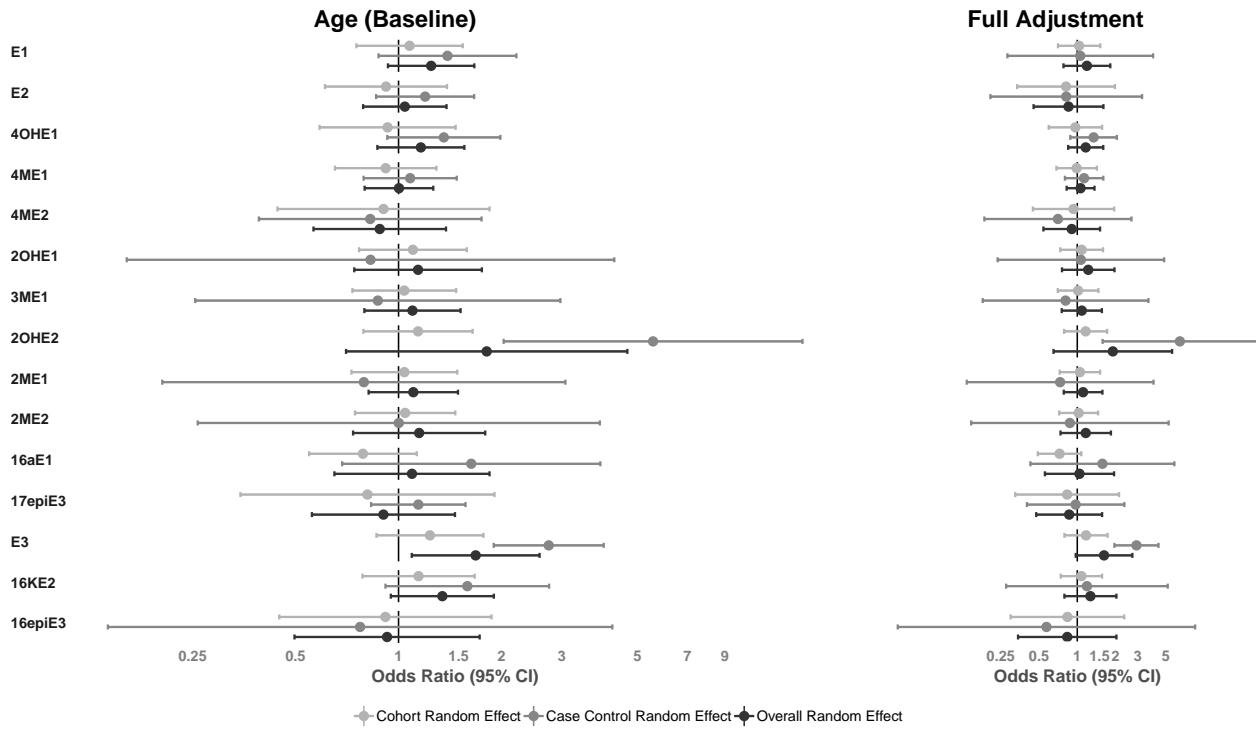


Figure 11: Study specific pooled random effect logistic model

Odds ratios (OR) and corresponding 95% confidence intervals (CI) were calculated using random effects maximum likelihood for each estrogen marker using multivariable unconditional logistic regression adjusted for age, body mass index, smoking, alcohol, education, and family history of gastric cancer as available. All studies could not be adjusted for all covariates due to either uncollected data or sparsity in categories. The model for Japan did not include education and Korea (KMCC) did not include family history of gastric cancer due to data not being collected for these covariates.

Multivariate logistic model used for each study:

- Iran:  $\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 E_i + \beta_2 age + \beta_3 BMI + \beta_4 education$
- Korea (KMCC):  $\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 E_i + \beta_2 age + \beta_3 BMI + \beta_4 education + \beta_5 smoke$
- Germany:  $\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 E_i + \beta_2 age + \beta_3 BMI$
- Korea (SNU):  $\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 E_i + \beta_2 age + \beta_3 BMI + \beta_4 education + \beta_5 smoke + \beta_6 GCrelative$
- Japan:  $\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 E_i + \beta_2 age + \beta_3 BMI + \beta_4 GCrelative$

Where  $\pi = Pr(Y = 1|E, X)$  and  $E_i$  = metabolites 1-15 as numeric tertile entries.

ORs were pooled by random effect meta-analysis separately for cohort and case-control studies to assess design-specific effect differences. Overall OR random effects were calculated by pooling all 5 studies' fixed effects together. Overall effects were not calculated from pooling the separate cohort and case-control random effects studies together since the assumption is that cohort and case-control effects and variances should inherently be relatively equal. However, pooling the two studies' random effects together rather than all 5 studies simultaneously only increased variance by a minute amount of  $\approx 0.1$ . All p-values were two-sided. Given the exploratory nature of our study p-values were not corrected for multiple comparisons. All statistical analyses and data visualizations were performed using R package "metafor" for meta-analysis and ggplot2.

The fully adjusted model showed that no estrogens were statistically significant at the  $\alpha=0.05$  level, although E3 was close. However, it should be noted that fixing the extreme outliers, NF values, missing creatinine

readings, and low reproducibility can drastically change results.

## Additional Analysis

**Table 1: Comparison of median estrogen metabolite concentrations reported from each study**

Estrogen	Nomenclature	Constanza et al. 2018	Moore et al. 2016 (% Difference) [Absolute Difference]	Sampson et al. 2016 (% Difference) [Absolute Difference]
E1	Estrone	3.02	2.04 (68%) [-0.98]	5.8 (192%) [2.78]
E2	Estradiol	0.57	0.42 (74%) [-0.15]	1.2 (211%) [0.63]
4OHE1	4-Hydroxyestrone	0.55	0.22 (40%) [-0.33]	0.7 (127%) [0.15]
4ME1	4-Methoxyestrone	0.33	0.05 (15%) [-0.28]	0.1 (30%) [-0.23]
4ME2	4-Methoxyestradiol	0.12	0.02 (17%) [-0.1]	0.1 (83%) [-0.02]
2OHE1	2-Hydroxyestrone	2.33	1.62 (70%) [-0.71]	4.9 (210%) [2.57]
3ME1	2-Hydroxyestrone-3-methyl ether	0.32	0.09 (28%) [-0.23]	0.3 (94%) [-0.02]
2OHE2	2-Hydroxyestradiol	0.24	0.38 (158%) [0.14]	1.2 (500%) [0.96]
2ME1	2-Methoxyestrone	1.43	0.36 (25%) [-1.07]	1.1 (77%) [-0.33]
2ME2	2-Methoxyestradiol	0.44	0.16 (36%) [-0.28]	0.5 (114%) [0.06]
16aE1	16a-Hydroxyestrone	0.53	0.49 (92%) [-0.04]	1.5 (283%) [0.97]
17epiE3	17-Epiestradiol	0.11	0.12 (109%) [0.01]	0.4 (364%) [0.29]
E3	Estriol	2.42	1.83 (76%) [-0.59]	5.5 (227%) [3.08]
16KE2	16-Ketoestradiol	0.46	0.53 (115%) [0.07]	1.6 (348%) [1.14]
16epiE3	16-Epiestradiol	0.72	0.2 (28%) [-0.52]	0.6 (83%) [-0.12]

Note: Units in pmol/mg creatinine.

Standardized estrogen concentration medians agreed with Moore et al., 2016 and Sampson et al., 2016 which used similar datasets of the same demographics (Comparison table found in Other Analysis section below).

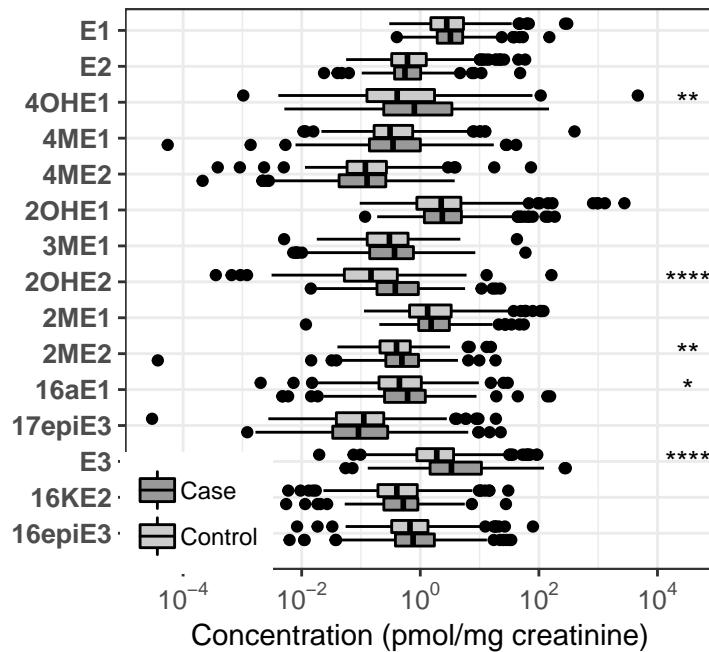


Figure 12: Pooled Wilcoxon-Mann-Whitney tests for cases vs. controls