Endogenous Estrogen Metabolites and Gastric Cancer Risk Among Postmenopausal Women

Statistical Analysis and Results

Isaac Zhao

01/08/19 Version 2

Materials and Methods

Study Design

Table 1: Demographics Table

	Cohort						Case Control					
	Iran		Korea (KMCC)		Germany		Korea (SNU)		Japan		Overall	
	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, 16α-hydroxyestrone, 16-ketoestradiol, estriol, 17-epiestriol, and 16-epiestriol) in an aliquot of 500 μL 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.

Table 2: Table of estrogen batch information

Batch	n	Study
1	40	Germany and Korea (KMCC)
2	40	Korea (KMCC)
3	40	Korea (KMCC)
4	36	Japan
5	40	Iran
6	40	Iran
7	40	Iran
8	20	Iran
9	44	Korea (SNU)
10	45	Korea (SNU)
11	44	Korea (SNU)
12	46	Korea (SNU)
13	45	Korea (SNU)
14	44	Korea (SNU)
15	47	Korea (SNU)
16	40	Korea (SNU)

Results

Covarying Relationships Between Metabolites

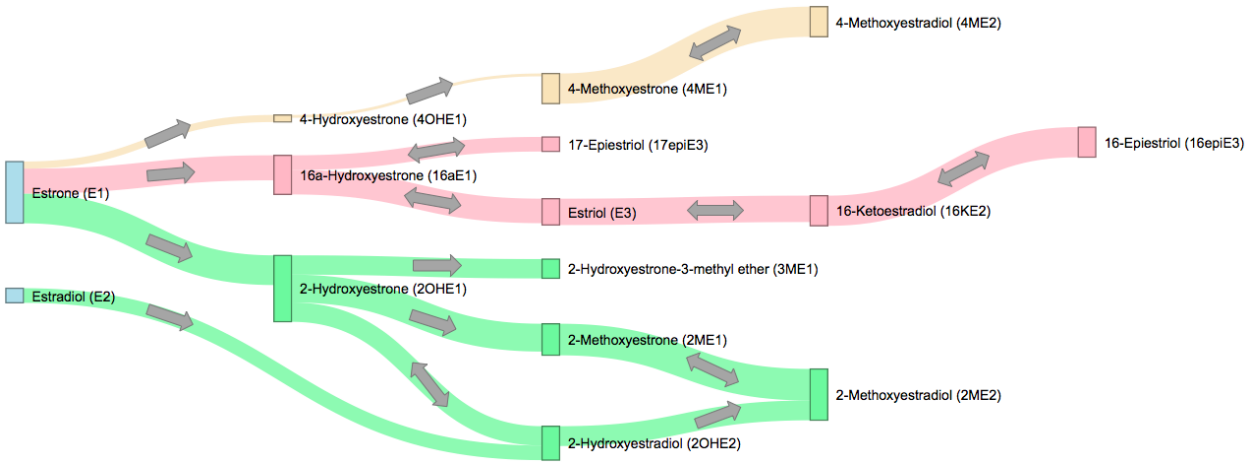


Figure 1: Estrogen metabolite hydroxylation pathway sankey diagram. Line thickness proportional to spearman correlations between substrates and products on hydroxylation pathways (2-OH, green; 4-OH, tan; 16-OH, pink), ranging from 0.06 to 0.67. 4-Hydroxyestradiol and 16a-Hydroxyestrone not shown due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

Correlations between metabolites were relatively equal throughout the pathways except for the 4-Hydroxylation pathway connection with the 4OHE1 metabolite.

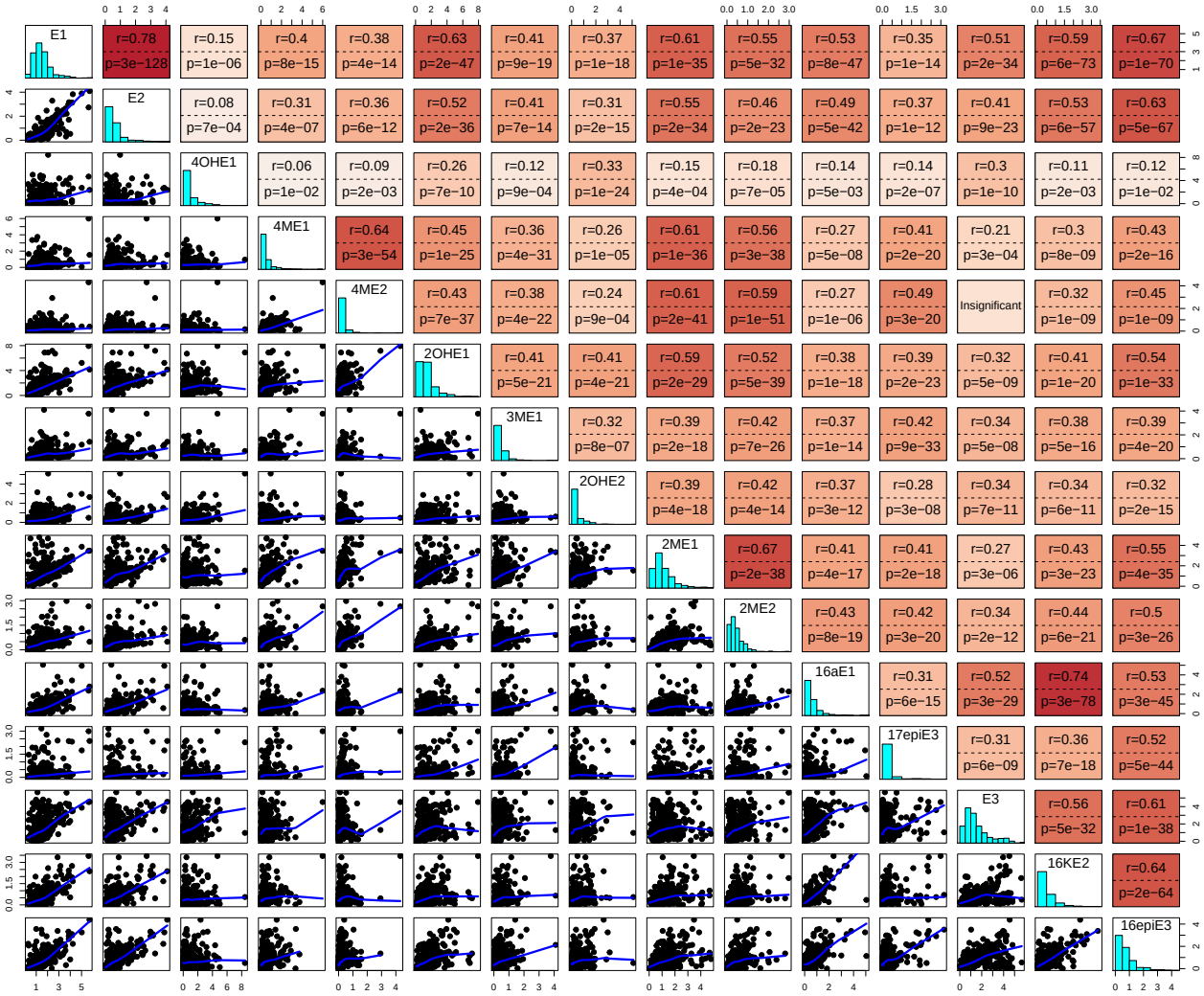


Figure 2: Metabolite correlation matrix. Redder regions indicate stronger positive correlations. Scatterplot and histogram on log scale.

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 use 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 bonferonni 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.

Not Found Data

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

Variable	NF (n)	NF (%)
E1	25	4.63
E2	23	4.26
4OHE1	183	33.89
4ME1	34	6.3
4ME2	23	4.26
2OHE1	28	5.19
3ME1	24	4.44
2OHE2	61	11.3
2ME1	22	4.07
2ME2	23	4.26
16aE1	77	14.26
17epiE3	29	5.37
E3	54	10
16KE2	42	7.78
16epiE3	35	6.48
creatinine	20	3.7

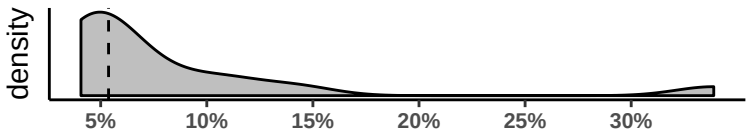


Figure 3: Distribution of NF percent among all metabolites. Vertical line indicates median.

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-eluding 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 of a measurement to be un-recorded. In other words, it could not be determined if the distribution of missingness was independent of laboratory procedures.

All estrogen batches were re-run to compare non-overlapping missing values (i.e. NF for a sample in the first run but not the second run), which determined that NFs did not appear to be correlated with how high or low the concentrations were. 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 justified and 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 α=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) may also be considered if the laboratory cannot amend accordingly.

NF values that were missing in both batch runs could not be determined if the nature was due to falling in the LLOD or not. However, since metabolites were strongly correlated with one another, one possibility to

diagnose is to plot one metabolite vs. another metabolite most correlated and see where the NF values lie.

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 to calculate ICC was $\frac{\sigma^2_{bs}}{\sigma^2_{bs}+\sigma_{bb}+\sigma^2_{ws}}$ and the formula for the computation of CV was $\frac{\sqrt{\sigma^2_{ws}+\sigma^2_{bb}}}{\mu}$; where σ^2_{bs} = variance between subject, σ^2_{bb} = variance between batch, and σ^2_{ws} = variance within subject. The variance components were computed using a two stage multilevel model with varying intercept for ID and batch.

Since concentrations were highly skewed, log-transformed concentrations were also computed for ICC but not for CV. Estrogen concentrations were log transformed to improve normality of distributions to meet assumptions necessary for the computation of variance - otherwise the interpretation could be misleading due to exaggerated variance components. A constant of 1 was added to all metabolite concentrations before log transformations to ensure the resulting mean concentrations were all positive values since many un-transformed concentrations were close to 0. The same constant was applied to each metabolite regardless of specific minimum concentrations per metabolite for consistency and comparability. CV could not be calculated on the log scale due to the constant added term affecting the mean component, which would otherwise produce negative log mean concentrations on the original scale.

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) showed higher variance since the measurement device is less sensitive at these levels. Thus, CV was 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 all postmenopausal women.

Laboratory Control QC Analysis

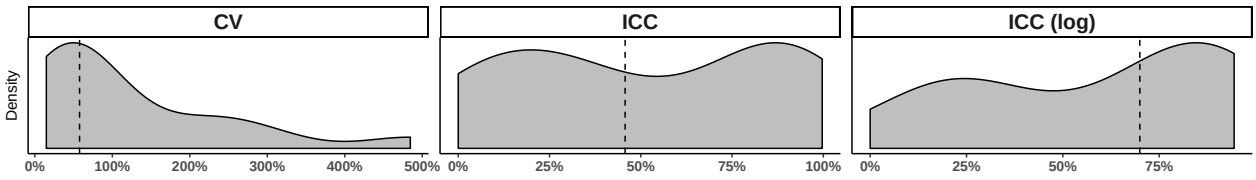


Figure 4: Distribution of CV and ICC for all metabolites for laboratory control QC samples. Vertical line indicates median.

Table 4: Laboratory control QC samples reproducibility calculations

	Original Scale						Log Transformed			
Estrogen	σ^2_{bs}	σ^2_{bb}	σ^2_{ws}	Mean	ICC	CV	σ^2_{bs}	σ^2_{bb}	σ^2_{ws}	ICC
E1	25.75	0.39	2.65	8.05	89.44	21.67	0.30	0.01	0.02	89.68
E2	6.39	0.15	0.41	2.89	92.02	25.76	0.39	0.01	0.02	94.40
4OHE1	843330.75	0.00	2524.43	339.34	99.70	14.81	6.62	1.69	0.00	79.66
4ME1	0.00	1.25	1.83	0.76	0.00	231.01	0.00	0.14	0.12	0.00
4ME2	0.01	0.01	0.02	0.13	17.48	139.93	0.00	0.01	0.01	20.54
2OHE1	8631.21	30252.90	46181.02	57.05	10.15	484.57	0.59	0.77	0.75	28.06
3ME1	0.01	0.03	0.04	0.51	19.13	49.09	0.01	0.01	0.01	20.35
2OHE2	317.52	0.00	146.12	4.00	68.48	302.49	1.29	0.04	0.34	77.19
2ME1	3.17	10.24	120.89	4.97	2.36	230.63	0.12	0.10	0.35	21.24
2ME2	0.03	0.03	0.03	0.45	36.11	54.08	0.01	0.01	0.01	41.61
16aE1	6.22	0.08	0.82	2.67	87.38	35.47	0.43	0.01	0.04	90.52
17epiE3	2.49	0.11	0.35	1.17	84.56	57.70	0.37	0.01	0.02	92.81
E3	87.82	83.18	95.87	14.91	32.91	89.77	1.01	0.32	0.12	69.96
16KE2	7.42	0.09	0.73	2.50	90.13	36.04	0.51	0.00	0.03	94.43
16epiE3	4.91	2.46	3.37	2.82	45.71	85.55	0.35	0.09	0.11	63.44

Note:
Log transformations calculated by log(concentration+1)
bs = between subject, bb = between batch, ws = within subject
Units in pmol/mg creatinine

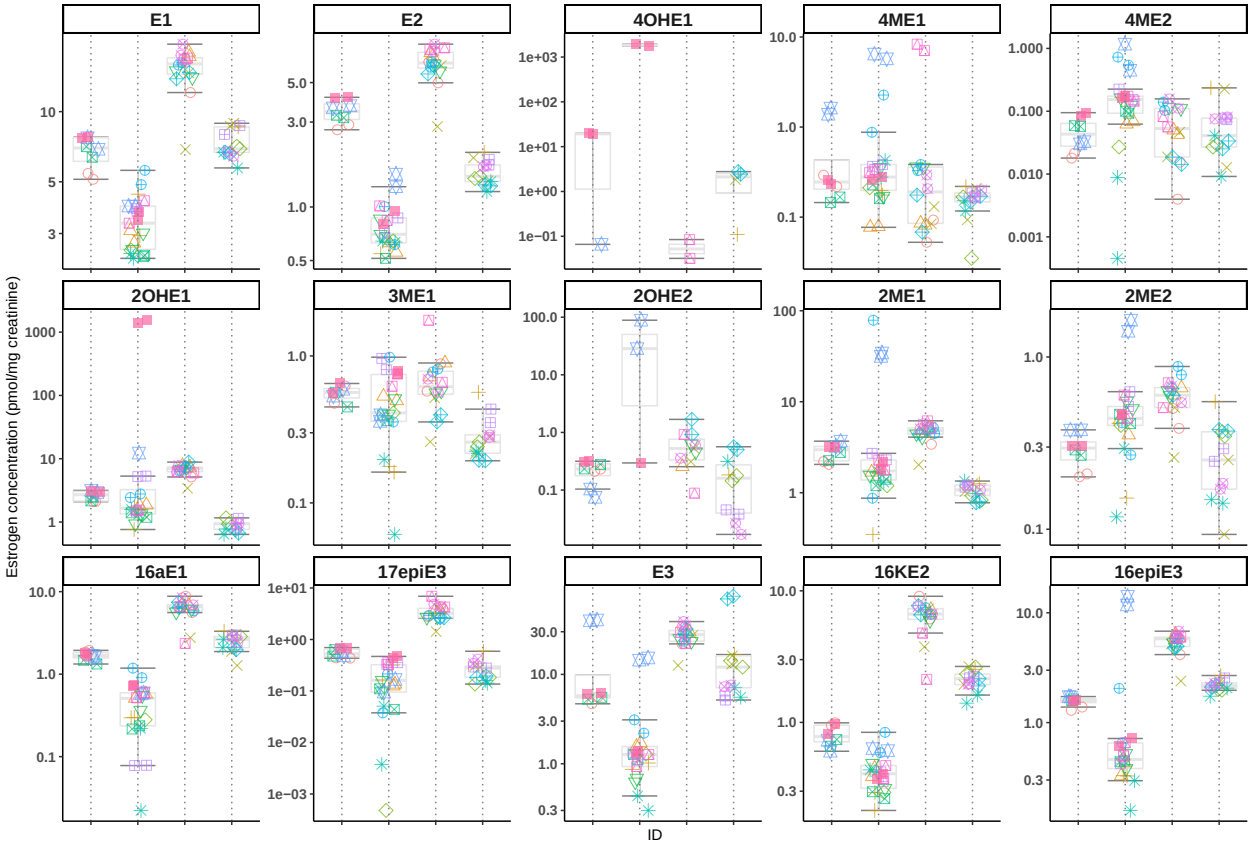


Figure 5: Four laboratory control QC sample IDs (x-axis) plotted against measured standardized concentrations. Batch identification indicated by color and shape.

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.

Most metabolites had ICCs $\geq 75\%$ with most CVs $\leq 50\%$, indicating moderate reproducibility generalizable across all batches. 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 measurments 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.

Korea SNU QC Analysis

Table 5: Korea SNU QC samples reproducibility calculations

	Original Scale						Log Transformed			
Estrogen	σ^2_{bs}	σ^2_{bb}	σ^2_{ws}	Mean	ICC	CV	σ^2_{bs}	σ^2_{bb}	σ^2_{ws}	ICC
E1	4.99	12.34	10.98	3.91	17.63	123.53	0.21	0.06	0.13	52.64
E2	0.50	0.89	0.59	0.73	25.20	167.11	0.08	0.03	0.04	52.37
4OHE1	0.13	3.33	21.20	2.28	0.54	217.70	0.17	0.27	0.25	25.14
4ME1	1.83	0.00	1.25	1.29	59.27	86.92	0.25	0.06	0.10	61.57
4ME2	0.04	0.00	0.06	0.25	41.14	97.82	0.02	0.00	0.02	52.76
2OHE1	4149.46	0.00	4589.28	21.39	47.48	316.69	1.16	0.12	0.21	77.82
3ME1	0.22	0.23	1.16	0.53	13.72	222.25	0.04	0.03	0.10	24.84
2OHE2	0.00	0.74	0.09	0.39	0.00	233.56	0.00	0.15	0.03	0.00
2ME1	3.82	0.00	11.37	2.75	25.15	122.61	0.27	0.00	0.21	55.63
2ME2	0.00	0.21	0.19	0.61	0.00	103.63	0.00	0.06	0.04	2.99
16aE1	0.36	0.16	0.43	0.74	38.02	104.48	0.09	0.02	0.05	56.14
17epiE3	0.96	1.19	1.66	0.91	25.21	185.71	0.12	0.11	0.11	36.68
E3	215.41	496.20	834.66	17.64	13.93	206.77	0.36	1.51	0.51	15.15
16KE2	0.18	0.31	0.38	0.57	20.30	145.86	0.05	0.01	0.05	42.79
16epiE3	4.47	0.42	1.96	1.72	65.26	89.69	0.33	0.01	0.10	76.18

Note:
Log transformations calculated by log(concentration+1)
bs = between subject, bb = between batch, ws = within subject
Units in pmol/mg creatinine

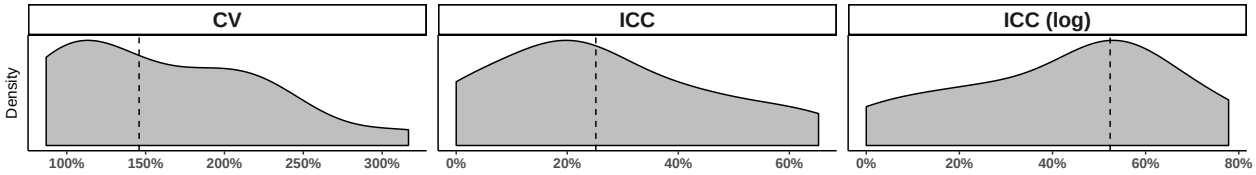


Figure 6: Distribution of CV and ICC for all metabolites for Korea SNU QC samples. Vertical line indicates median.

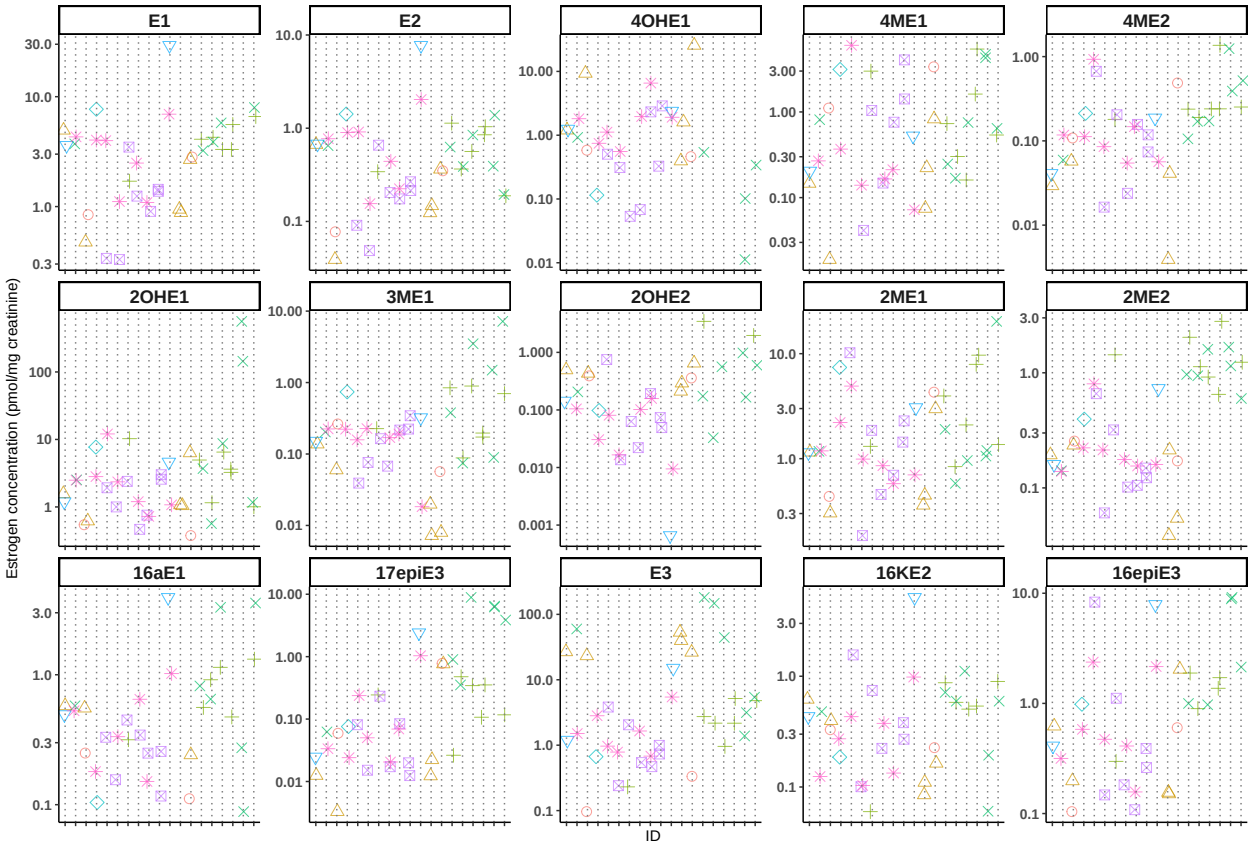
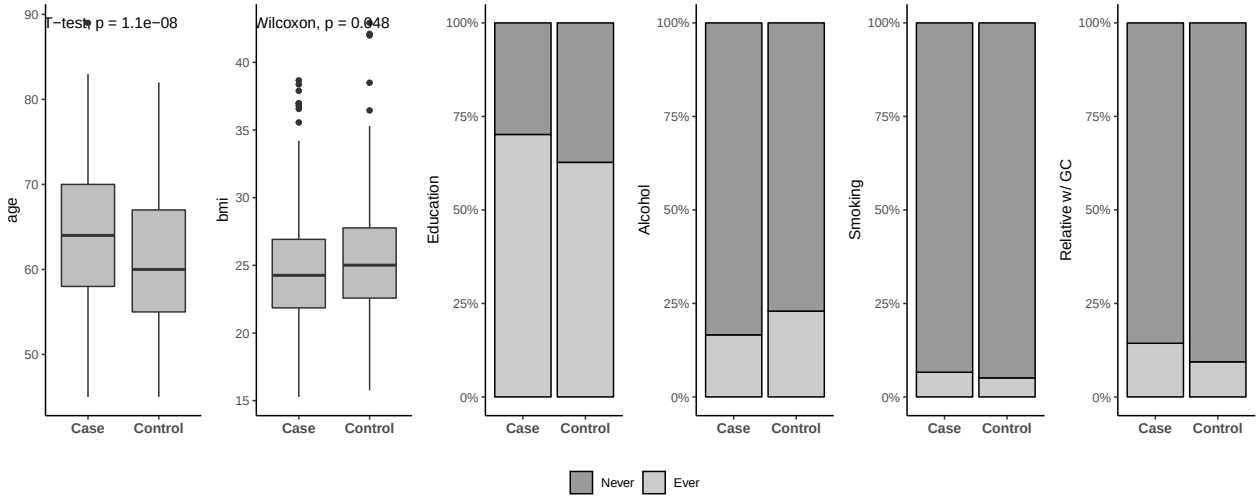


Figure 7: 19 Korea SNU QC sample IDs (x-axis) plotted against measured standardized concentrations. Batch identification indicated by color and shape.

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. 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.

Assessing Association between Covariates and Gastric Cancer



Cases were on average an older age while controls had a higher BMI on aveage. 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.

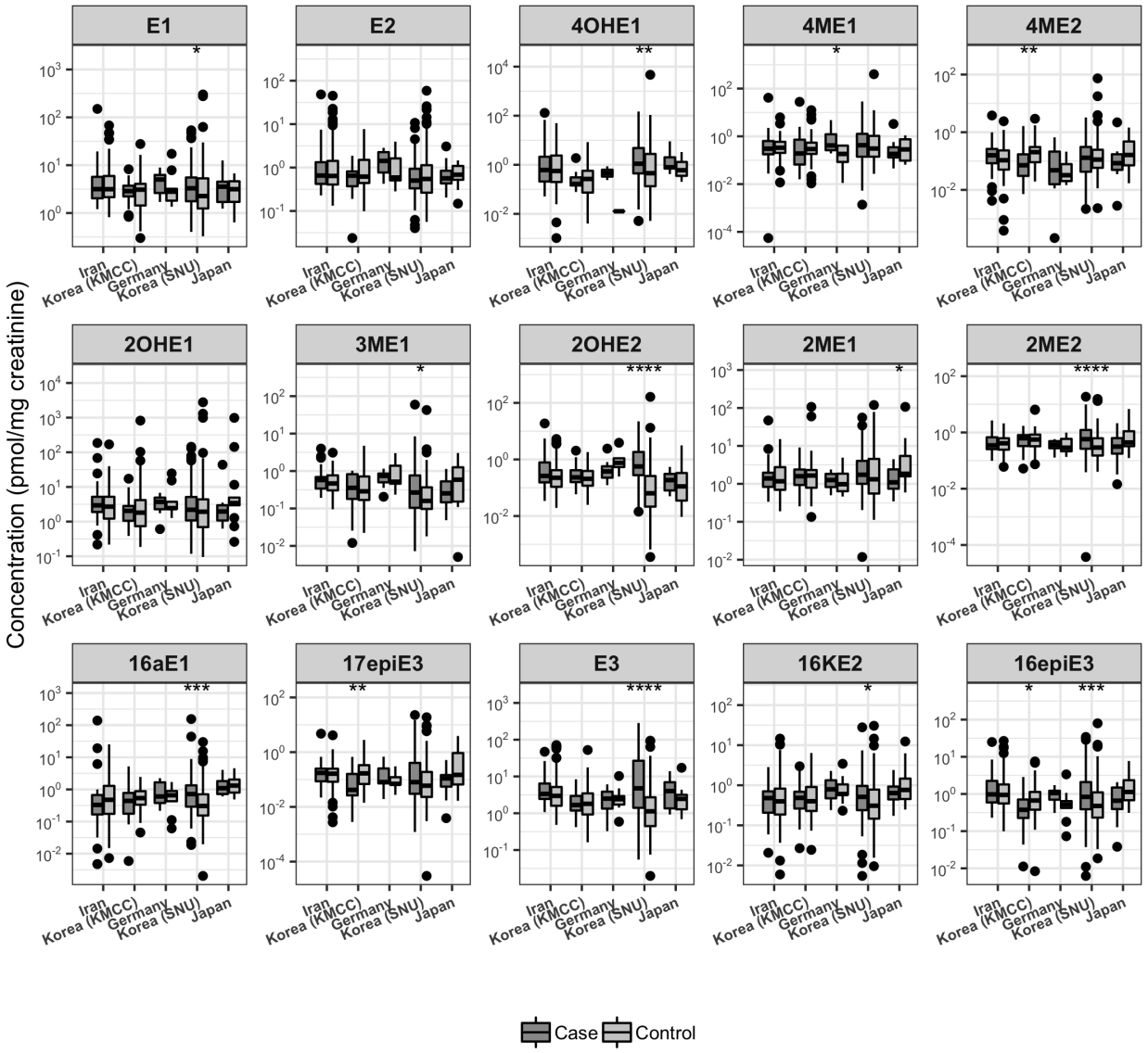


Figure 8: Study specific Wilcoxon-Mann-Whitney tests for cases vs. controls

Postmenopausal women with gastric cancer was compared with controls for both pre-diagnostic and early-stage case-control sample sets. Each study was analyzed seperately 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 α=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). These striking differences of associations of Korea SNU vs. Korea KMCC may indicate an inherent flaw that

needs to be studied more closely since it should not be expected that different study designs of cohort and early-stage case-control should affect observed associations drastically.

Table 6: Estrogen concentration median and tertile cutpoint

	Cohort			Case Control	
	Iran	Korea (KMCC)	Germany	Korea (SNU)	Japan
	Median (Tertile)	Median (Tertile)	Median (Tertile)	Median (Tertile)	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. β 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 such that variance of measurement will not have such drastic changes on categorization. For example, a measurement categorized in tertile 2 could have also been in tertile 1 or 3 if measured again - changing the difference to the extreme low or high end. Finer tuning into deciles where the measurement would be categorized in decile 5 would not restrict the measurement to be placed in the extreme low or high end. Korea SNU appeared to have lower estrogen concentrations than other studies, especially for the metabolites with high concentrations overall (E1, 2OHE1, E3).

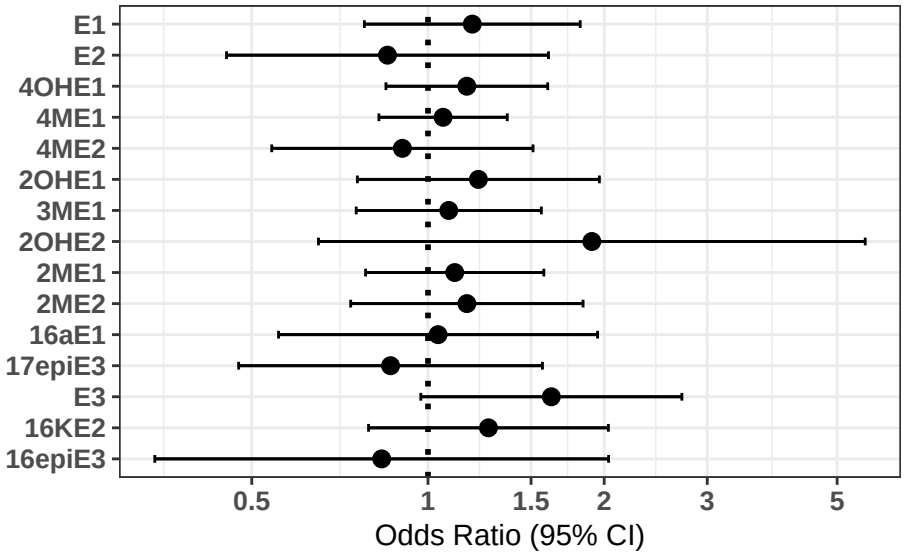


Figure 9: Forestplot of fully adjusted pooled random effect meta-analysis logistic regression model

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 sparisty 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.

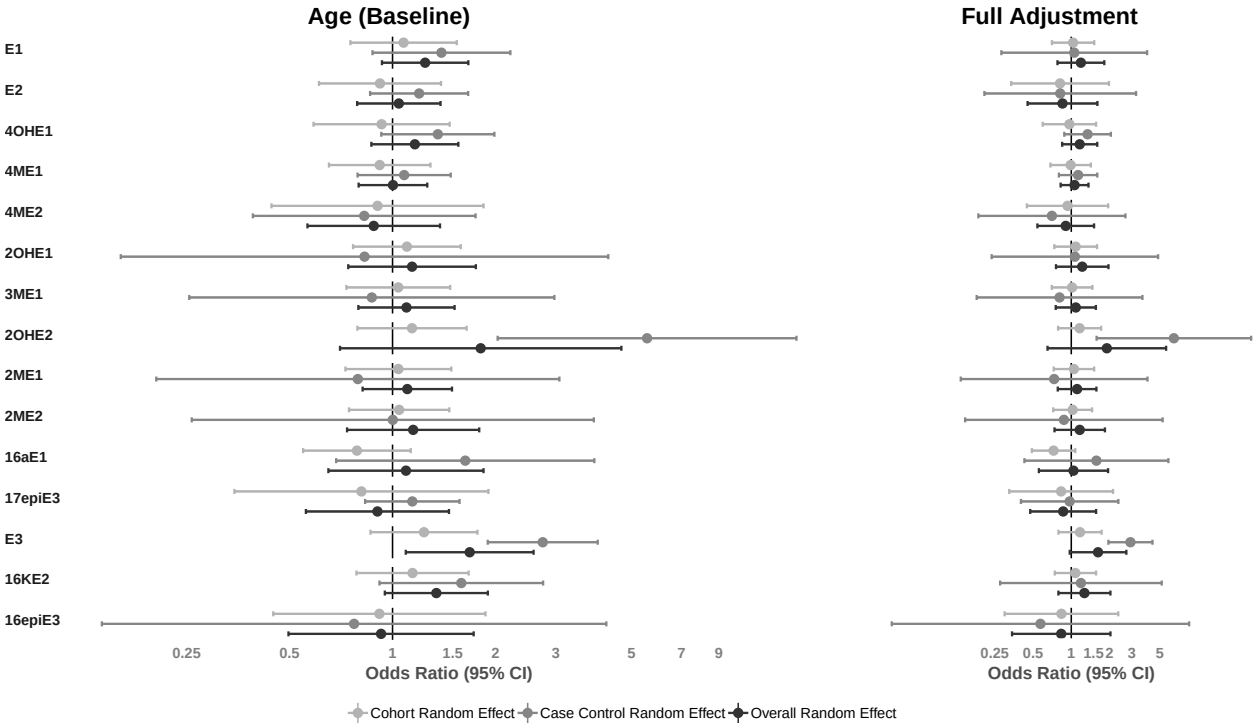
Multivariate logistic model used for each study:

- 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 ≈ 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.

Markers with statistically significant ordinal trends ($p<0.05$) were further examined for associations of each tertile with gastric cancer risk comparing to the lowest tertile.



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-Epiestriol	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-Epiestriol	0.72	0.2 (28%) [-0.52]	0.6 (83%) [-0.12]

Note: Units in pmol/mg creatinine.

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

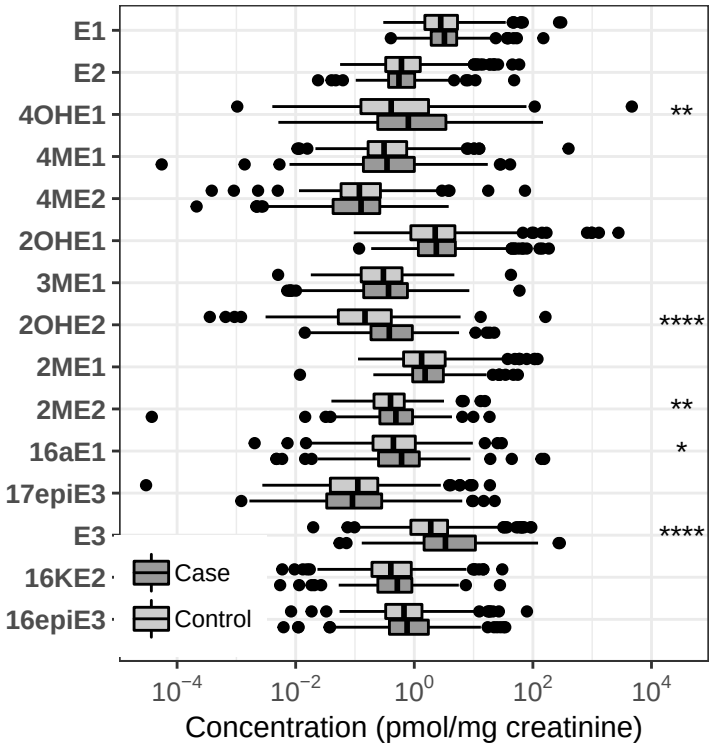


Figure 10: Pooled Wilcoxon-Mann-Whitney tests for cases vs. controls