**Attachment 3 – Justification for Sample Size Calculations in Multi-Site Study (Supporting information)**

- **Attachment 3a – Child Sample Size Calculations**
- **Attachment 3b – Adult Sample Size Calculations**

Sample size calculations were conducted using OpenEpi Version 3.03 (Dean AG, Sullivan KM, Soe MM. OpenEpi: Open Source Epidemiologic Statistics for Public Health, www.OpenEpi.com, updated 2014/09/22). For some health-related endpoints, calculations could not be conducted because of a lack of information in the studies on the parameters needed to make the calculations.

Sample size calculation for mean difference:

$$n_1 = \frac{(\sigma_1^2 + \sigma_2^2 / \kappa)(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}$$

$$n_2 = \frac{(\kappa * \sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}$$

The notation for the formulae are:

$n_1$ = sample size of Group 1

$n_2$ = sample size of Group 2

$\sigma_1$ = standard deviation of Group 1

$\sigma_2$ = standard deviation of Group 2

$\Delta$ = difference in group means

$\kappa$ = ratio = $n_2/n_1$

$Z_{1-\alpha/2}$ = two-sided Z value (eg. Z=1.96 for 95% confidence interval).

$Z_{1-\beta}$ = power

Where $N_2/N_1$ is the ratio of the two sample sizes. Then $N_2$ is simply this ratio multiplied by $N_1$. For a type 1 error (or α error) of .05, the $Z_{1-\alpha/2}$ value is 1.96. This calculation is for a two-tailed hypothesis test and equivalent to using a 95% confidence interval to determine statistical significance. For a one-tail test with α =.05, the $Z_{1-\alpha/2}$ in the above equation is replaced by $Z_{1-\alpha}$ and its value is 1.65, equivalent to using a 90% confidence interval to determine statistical significance. The $Z_{1-\beta}$ in the above equation is the Z value for the selected power. For 80% power, $Z_{1-\beta}$ = 0.84, for 90% power, $Z_{1-\beta}$ = 1.28, and for 95% power, $Z_{1-\beta}$ = 1.65. (See Rosner B. Fundamentals of Biostatistics, 7th Edition, equation 8.27, p. 302).

The sample size calculations for odds ratios, risk ratios, etc. are as follows:

The sample size formula *with<u>out</u>* the correction factor by Fleiss is:

$$n_1 = \frac{\left[Z_{\alpha/2}\sqrt{(r+1)\bar{p}\bar{q}} + Z_{1-\beta}\sqrt{rp_1q_1 + p_2q_2}\right]^2}{r(p_1 - p_2)^2}$$

$$n_2 = r\,n_1$$

For the Fleiss method ***with*** the correction factor, take the sample size from the uncorrected sample size formula and place into the following formula:

$$n_{1cc} = \frac{n_1}{4}\left[1 + \sqrt{1 + \frac{2(r+1)}{n_1 r|p_2 - p_1|}}\right]$$

$$n_{2cc} = r\,n_{1cc}$$

When the input is provided as an odds ratio (OR) rather than the proportion of exposed with disease, the proportion of exposed with disease is calculated as:

$$p_1 = \frac{p_2 OR}{1 + p_2(OR - 1)}$$

When the input is provided as a risk (or prevalence) ratio (RR) rather than the proportion of exposed with disease, the proportion of exposed with disease is calculated as:

$$p_1 = p_2 RR$$

Fleiss JL. Statistical Methods for Rates and Proportions. John Wiley & Sons, 1981.

**<u>Note</u>**: In some studies, the standard deviation is not presented but instead, the interquartile range (IQR) is given. Assuming a normal distribution for the outcome under evaluation (e.g., thyroid function measures), the standard deviation can be calculated by dividing the IQR range by 1.35. However if the outcome is not normally distributed, this formula could underestimate the standard deviation. In particular, if the outcome under evaluation has been log-transformed presumably to achieve a normal distribution, the untransformed outcome is unlikely to have a normal distribution. Therefore, using this formula when the outcome does not have a normal distribution may underestimate the SD by as much as 20% according to simulations conducted in Wan X. 2014. A higher SD would increase the sample size requirement.

**Attachment 3a. Child Sample Size Calculations**

The following notes provide comments and information on the parameters (e.g., standard deviation, disease prevalence) used in the sample size calculations provided in Table 1 for the children study.

Sample size calculations were conducted with type 1 ("α error") set at .05 and type 2 error ("β error) set at .20. Sample sizes per stratum group were calculated. It was considered important that a study have a total sample size so that exposures could be categorized into tertiles (i.e., reference level, medium level, and high level) or quartiles (i.e., reference level, low, medium and high).

Studies were selected that were considered the most representative of U.S. populations exposed via drinking water to PFOA, PFOS and/or PFHxS as a result of the migration of these PFAS chemicals into ground water or surface water sources from the use of aqueous film forming foam (AFFF). The PFOS, PFOA and PFHxS results in the NHANES studies were used in many of the sample size calculations. For those outcomes not included in NHANES studies, the C8 studies were used.  Where applicable studies from Taiwan or other major industrialized countries were also used.

**Examples:**

Lipids

In the C8 study (Frisbee 2010), the mean total cholesterol level in the study population was 160.7 mg/dL and the standard deviation (SD) was 29.3. The sample size calculations assumed the same SD in children exposed and the unexposed group. For hypercholesterolemia (total cholesterol ≥ 170 mg/dL), the prevalence in the C8 study was 34.2%.

Uric Acid

In the NHANES study (Geiger 2013), the mean uric acid level in the study population was 5.07 mg/dL with a SD of 1.19. The sample size calculations assumed the same SD in the exposed children and the unexposed group. The prevalence of hyperuricemia (uric acid ≥ 6 mg/dL) in the NHANES study was 16%.

Kidney Function

The mean estimated glomerular filtration rate (eGFR) in the C8 study of children and adolescents (Watkins 2013) was 133 mL/min/1.73 $m^2$ with a SD of 23.9. The sample size calculations assumed the same SD in the exposes children and the unexposed group.

Attention Deficit/Hyperactivity Disorder (ADHD)

In the C8 study (Stein 2011), the prevalence of participant-reported ADHD was 12.4% and the prevalence for participant-reported + used medications for ADHD was 5.1%. Sample size calculations used the 12.4% prevalence.

Hypersensitivity-related Outcomes

From an NHANES study (Stein 2016), the prevalence of current asthma and rhinitis among those aged 12-19 were 10.9% and 25.6%, respectively. For atopic dermatitis, the prevalence for children and adolescents (ages 5-17) is about 12% based on data from the National Health Interview Survey.

Sex hormones and Insulin-like growth factor – 1 (IGF-1)

C8 study of children (Lopez-Espinosa 2016)

a. Testosterone

For PFOS, there was a -6.6% difference in the natural log testosterone among girls (per interquartile range of the natural log of PFOS). Among girls, the median testosterone level was 15 ng/dL with an IQR of <LOD, 21 and the LOD of 10 ng/dL. For the sample size calculation of mean difference, the standard deviation was assumed to be equal for the exposed and unexposed groups and equal to 11.85. (Assuming LOD/2 was the lower limit of the IQR, the range = 21 - 5 = 16. Assuming a normal distribution, dividing 16 by 1.35 converts the IQR to a standard deviation, which equaled 11.85)[1]. To obtain the mean difference, the median testosterone level (15 ng/dL) was assumed to be the reference level (i.e., the level among the unexposed). The natural log of the median equals 2.71. A 6.6% decrease equals 2.53. Exponentiating 2.53 equals 12.55. The mean difference is then 15 – 12.55 = 2.45.

Assuming a 95% CI and 80% power, the sample size = 368/group; for a ratio of 2, the sample sizes = 552 and 276.

b. IGF-1

For PFHxS, there was a -2.5% difference in the natural log IGF-1 among boys (per interquartile range of the natural log of PFHxS). Among boys, the median IGF-1 level was 147 ng/mL with an IQR of 116, 187. For the sample size calculation of mean difference, the standard deviation was assumed to be equal for the exposed and unexposed groups and equal to 52.6. (The IQR range was 187 – 116 = 71. Assuming a normal distribution, dividing 71 by 1.35 converts the IQR to a standard deviation, which equaled 52.6)[1]. To obtain the mean difference, the median IGF-1 level (147 ng/mL) was assumed to be the reference level (i.e., the level among the unexposed). The natural log of the median equals 4.99. A 2.5% decrease equals 4.865. Exponentiating 4.865 equals 129.7. The mean difference is then 147 – 129.7 = 17.3.

Assuming a 95% CI and 80% power, the sample size = 146/group; for a ratio of 2, the sample sizes = 218 and 109.

For PFOS, there was a -5.9% difference in the natural log IGF-1 among boys (per interquartile range of the natural log of PFOS). This would require considerably smaller sample sizes for IGF-1 than those for PFHxS.

Thyroid function – Children/Adolescents

Taiwan study of children. (Lin 2013)

a. For males aged 12-19, there was a mean difference in the log TSH of -.50 mIU/L for PFOA levels in the 90$^{th}$ percentile (>9.71 ng/ml) compared to the reference level of PFOA exposure. The standard error for the reference group was 0.26 with N=32 in this group; and the standard error for the 90$^{th}$ percentile group was 0.33

with N=6. The standard deviations for the reference and 90th percentile groups were therefore 1.47 and 0.81, respectively.

Assuming a 95% CI and 80% power, the sample size = 89/group; for a ratio of 2, the sample sizes = 158 and 79.

b. For females aged 12-19, there was a mean difference in the log TSH of -.35 mIU/L for PFOA levels in the 90th percentile (>9.71 ng/ml) compared to the reference level of PFOA exposure. The standard error for the reference group was 0.18 with N=71 and the standard error for the 90th percentile group was 0.24 with N=14. The standard deviations for the reference and 90th percentile groups were therefore 1.52 and 0.90, respectively.

Assuming a 95% CI and 80% power, the sample size = 200/group; for a ratio of 2, the sample sizes = 348 and 174.

Additional notes:

Sample sizes for the categorical outcomes in Table 1 were based on the following prevalence in children (also listed in Appendix C Table 1):

Hypercholesterolemia: 34.2%
Hyperuricemia: 16%
Thyroid disease: 0.6%
ADHD 12.4% reported only; 5.1% reported with additional reporting on medications used for ADHD
Asthma: 11%
Rhinitis: 25.6%
Atopic dermatitis: 10.7%
Hypertension: 23.4%
Obesity: 17%

**Attachment 3b. Adult Sample Size Calculations**

The following provides information on the parameters and sample size calculation used in Table 2 for the adult study.

Sample size calculations were conducted with type 1 ("α error") set at .05 and type 2 error ("β error) set at .20. It was considered important that a study have a total sample size so that exposures could be categorized into tertiles (i.e., reference level, medium level, and high level) or preferably into quartiles (i.e., reference level, low, medium and high).

Studies were selected that were considered the most representative of U.S. populations exposed via drinking water to PFOA, PFOS and/or PFHxS as a result of the migration of these PFAS chemicals into ground water or surface water sources from the use of aqueous film forming foam (AFFF). The PFOS, PFOA and PFHxS results in the NHANES studies were used in many of the sample size calculations. For those outcomes not included in NHANES studies, the C8 studies were used. Where applicable studies from Taiwan or other major industrialized countries were also used.

**Example:**

**Liver Function – Adults**

In the C8 study (Darrow 2016), the mean alanine aminotransferase (ALT) level was 26 IU/L and the standard deviation was 19. The linear regression coefficient for the natural log ALT in the fifth quintile level of cumulative natural log PFOA was 0.058. Assuming that the reference group had an ALT level equal to the mean, the natural log of the mean ALT would be 3.26. Therefore the natural log of ALT for the fifth quintile cumulative log PFOA would be 3.32. Exponentiating 3.32 equals 27.6. The mean difference in the untransformed ALT is then 1.6.

Assuming a 95% CI and 80% power, the sample size = 2,214/group.

**Thyroid Function – Adults (not included in Table 2)**

In a study done by Shrestha 2015, the sample size was 87 adults aged 55-74. Mean and SD for TSH was 2.58 µIU/mL and 1.47, respectively. The linear regression of the natural log TSH resulted in a coefficient for the natural log PFOS of 0.129. Using a PFOS level of 15 ng/mL, the natural log of 15 is 2.71; multiplied by 0.129 equals 0.35. The reference level TSH was assumed to be the median TSH of 2.15 µIU/mL. The natural log of 2.15 is 0.77; adding 0.35 equals 1.12. Exponentiating 1.12 equals 3.06. The mean difference is then 3.06 – 2.15 = 0.91. The standard deviation of 1.47 was used for each group.

Assuming a 95% CI and 80% power, the sample size = 41/group.

Assuming a 95% CI and 95% power, the sample size = 68/group.

a. TSH

In Ji 2012, the sample size was 633, ≥12 years of age and the median TSH level was 1.37 µIU/mL with an IQR of 0.90, 2.01. The standard deviation was estimated as the IQR range divided by 1.35: (2.01 - .90)/1.35 = 0.82. This standard deviation was assumed for each group. For TSH, the linear regression coefficients for PFOS and PFHxS were 0.062 and 0.013, respectively. Using a PFOS level of 15 ng/mL and a PFHxS level of 9 ng/mL, the mean difference for PFOS and PFHxS are 0.93 and 0.12, respectively.

Assuming a 95% CI and 80% power, the sample size = 13/group for PFOS

Assuming a 95% CI and 80% power, the sample size = 733/group for PFHxS

b. $TT_4$ (total thyroxine)

In Ji 2012, the sample size was 633, ≥12 years of age and the median $TT_4$ level was 7.4 µg/dL and the IQR was 6.7, 8.1. The standard deviation was estimated: (8.1 – 6.7)/1.35 = 1.04. This standard deviation was assumed for each group. For $TT_4$, the linear regression coefficients for PFOS and PFHxS were -0.021 and -0.007, respectively. Using a PFOS level of 15 ng/mL and a PFHxS level of 9 ng/mL, the mean difference for PFOS and PFHxS are -0.32 and -0.06, respectively.

Assuming a 95% CI and 80% power, the sample size = 166/group for PFOS

Assuming a 95% CI and 80% power, the sample size = 4,716/group for PFHxS

Sample sizes for the categorical outcomes in Table 2 were based on the following prevalences in adults:

Hypercholesterolemia: 15%
Hyperuricemia: 24%
Thyroid disease: 6.5% (reported and confirmed by medical records); 11.5% (reported only)
Elevated ALT: 11.2%
Elevated GGT: 14%
Elevated bilirubin: 1.1%
Osteoporosis: 5%
Osteoarthritis: 7.6%
Cardiovascular disease: 13%
Ulcerative colitis: 0.5%
Rheumatoid arthritis: 1.2%

Health related Endpoints Not shown in Table 2:

Chronic kidney disease: 1.4%
Liver disease: 2%
Hypertension: 37%

Pregnancy-induced hypertension: 8.5%
Endometriosis: 7%
Lupus: 0.2%
Multiple sclerosis: 0.32%
Kidney cancer: 0.3%