## 1. Study title

Genome Wide Association Study of Modulators of Pregnancy in Long Term Survivors of Pediatric Cancer

## 2. Investigators and Working Group

### 2.1. Investigators:


| Name | Email | Affiliation |
|---|---|---|
| Seth Rotz, MD | rotzs@ccf.org | Department of Pediatric Hematology, Oncology, and Bone Marrow Transplantation, Cleveland Clinic Children's Hospital, Cleveland, OH |
| Bo Hu, PhD | hub@ccf.org | Department of Quantitative Health Sciences, Cleveland Clinic Foundation, Cleveland, OH |
| Peter Bazeley, MD, MS | bazelyp@ccf.org | Center for Clinical Genomics, Cleveland Clinic Foundation, Cleveland, OH |
| Sarah Worley, MS | worleys@ccf.org | Department of Quantitative Health Sciences, Cleveland Clinic Foundation, Cleveland, OH |
| Wendy Leisenring, ScD | wleisenr@fredhutch.org | Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA |
| Melissa Hudson, MD | Melissa.Hudson@STJUDE.ORG | Cancer Survivorship Division, St. Jude Children's Research Hospital, Memphis, TN |
| Les Robison, PhD | Les.Robison@STJUDE.ORG | Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, Memphis, TN |
| Yutaka Yasui, PhD | Yutaka.Yasui@STJUDE.ORG | Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, Memphis, TN |
| Kevin C Oeffinger, MD | kevin.oeffinger@duke.edu | Department of Community and Family Medicine, Duke Cancer Institute, Duke University, Durham, NC |
| Smita Bhatia, MD, MPH | sbhatia@peds.uab.edu | Institute for Cancer Outcomes and Survivorship, UAB Comprehensive Cancer Center, University of Alabama Birmingham, Birmingham, AL |
| Navneet Majhail, MD | majhain@ccf.org | Department of Medical Oncology, Taussig Cancer Institute, Cleveland Clinic Foundation, Cleveland, OH |
| Dennis John Kuo, MD, MS | dekuo@ucsd.edu | Division of Pediatric Hematology-Oncology, University of California, San Diego, San Diego, CA |
| Hui-Chun Irene Su, MD, MSCE | hisu@ucsd.edu | Department of Reproductive Medicine, University of California, San Diego, San Diego, CA |

| | | |
|---|---|---|
| Debashis Sahoo, PhD | dsahoo@ucsd.edu | Department of Pediatrics, University of California, San Diego, San Diego, CA |
| Jennifer Levine, MD, MS | jel9022@med.cornell.edu | Division of Pediatric Hematology-Oncology, Weill Cornell Medical College, New York, NY |
| Saro Armenian, DO, MPH | SArmenian@coh.org | Departments of Pediatrics and Population Sciences, City of Hope Cancer Center, Duarte, CA |
| Catherine Su | c2su@ucsd.edu | School of Medicine, University of California, San Diego, San Diego, CA |

### 2.2. Working Groups

CCSS Genetics and Chronic Conditions Working Groups

### 2.3. PI Contact Information

Seth J. Rotz MD
rotzs@ccf.org
9500 Euclid Avenue
Cleveland, Ohio 44195
216.217.8106

## 3. Background and Rationale

Infertility concerns have a significant impact on quality of life in survivors of cancer.[1,2] Emerging technologies are expanding the pool of cancer patients who may be able to preserve fertility despite intensive therapy.[3-5] Among females ages 15-44 followed by the CCSS, the relative risk of a survivor ever being pregnant was 0.81 (95%CI, 0.73-0.90; P< .001) compared with female siblings.[6] A follow up study published in 2016 demonstrated a hazard ratio (HR) of 0.87 (95%CI 0.81-0.94; P=0.00015) of having been pregnant, compared to siblings.[7] In a self-reported survey of female CCSS participants 455/3,531 (13%) reported that they were clinically infertile (i.e., >1 year of attempts at conception without success).[8] Among males aged 15-44 followed by the CCSS the HR of siring a pregnancy was 0.63 (95%CI 0.58-0.68; P<0.0001), compared to siblings.[7] In another survey of male CCSS patients self-reported rates of infertility were assessed. The prevalence of infertility was 46.0% (412/938) in survivors versus 17.5% in siblings (RR=2.64, 95%CI 1.88-3.70; P< 0.001).[9]

Multiple chemotherapy agents have been associated with decreased fertility. Differences exist in germ cell tolerance to these agents between males and females.[7] In male survivors, reduced likelihood of siring a pregnancy was associated with cyclophosphamide, ifosfamide, procarbazine, and cisplatin. However, in female survivors, only busulfan, lomustine, and cyclophosphamide (doses in the highest quartile) are significantly associated with reduced pregnancy. Notably, pelvic and cranial radiotherapy is strongly associated with reduced rate of pregnancy in childhood cancer survivors,[6] and previous studies have excluded this population for analysis.[7]

Cyclophosphamide is used in a wide variety of pediatric cancers and substantial inter-patient variability exists in germ cell tolerance to this agent.[10-13] Males exposed to high doses of cyclophosphamide have high rates of azoospermia.[11] Male participants in the CCSS in the highest tertile of cumulative cyclophosphamide dose exposure (9,360 to 143,802 mg/m2) had a HR of 0.42 (95%CI, 0.31-0.57; P<0.001) of siring a pregnancy compared to those not receiving cyclophosphamide.[14] Likewise, female participants in the CCSS who were treated with cyclophosphamide had a lower likelihood of pregnancy (RR, 0.81; 95% CI, 0.68 to 0.93; P = .005)

in a multivariate model.[6] Multiple enzymes in the cytochrome P450 family play a role in cyclophosphamide metabolism. In previous smaller studies of women with lupus nephritis, heterozygosity or homozygosity for *CYP2C19*2* polymorphism was associated with lower rates of ovarian toxicity, whereas in women with breast cancer, *CYP3A4*1B* heterozygosity or homozygosity was associated with ovarian failure.[12,13] Although the role of polymorphisms in enzymes responsible for cyclophosphamide on ovarian failure has been explored, its effect on becoming pregnant in females or siring a pregnancy in males is not known, but are logical loci to investigate.

The CCSS conducted genotyping on 4.1 million loci for 5,739 childhood cancer survivors and has collected data on patient drug exposure, radiation exposure, age at treatment, and history of pregnancy and siring a pregnancy. Similarly, the SJLIFE study has collected clinical and GWAS data on survivors treated over several decades at St. Jude Children's Research Hospital.[15] Genetic polymorphisms and deleterious mutations play a role in infertility in the general population, although our understanding of these contributions remains incomplete.[16-19] For example, GWAS approaches have been explored to analyze causes of polycystic ovarian syndrome (PCOS), a condition leading to infertility in many women.[16]

Methods such as the cyclophosphamide equivalent dose (CED) are available to compare the risk of infertility across different treatment regimens, independent of study populations.[20] A CED ≥4,000 mg/m$^2$ has been associated with a decreased HR of pregnancy among partners of male childhood cancer survivors and increased rate ratio for non-surgical premature menopause in female childhood cancer survivors.[20] However, current methods do not take into account individual patient variability in drug tolerance, metabolism, and yet undiscovered genetic factors.

## 4. Specific aims/objectives/research hypotheses

Long-term survivors of pediatric cancer have multiple concerns regarding reproductive health.[21] Female survivors have a decreased likelihood of having been pregnant and male survivors have a decreased likelihood of having sired a pregnancy compared to sibling controls.[7] Genetic polymorphisms and deleterious mutations play a role in infertility in the general population, although our understanding of these contributions remains incomplete.[16-19] Greater doses of chemotherapeutic agents have been associated with reduced likelihood of having been pregnant or siring a pregnancy in pediatric cancer survivors. Presumably, genetic polymorphisms play a role in germ cell tolerance to chemotherapeutics and may have a modifying effect on rates of pregnancy in pediatric cancer survivors. *We hypothesize* that single-nucleotide polymorphisms (SNPs) are associated with decreased rates of pregnancy or siring a pregnancy in long-term survivors of pediatric malignancies. We will test this hypothesis with the following specific aims:

**Primary Aim 1:** Using a GWAS approach from the Childhood Cancer Survivor Study (CCSS) cohort, identify genetic risk factors associated with reduced likelihood of pregnancy in long-term female survivors of pediatric cancer. Individual SNPs association with the likelihood of pregnancy will be tested for both main effects and gene-environment interaction (GxE) with cyclophosphamide equivalent dose (≥4,000 mg/m$^2$).

**Primary Aim 2:** Using a GWAS approach from the CCSS cohort, identify genetic risk factors associated with reduced likelihood of having sired a pregnancy in long-term male survivors of pediatric cancer. Individual SNPs association with the likelihood of having sired a pregnancy will be tested for both main effects and GxE with cyclophosphamide equivalent dose (≥4,000 mg/m$^2$).

**Primary Aim 3:** Using a Boolean analysis of the CCSS GWAS dataset, identify genetic risk factors associated with reduced likelihood of pregnancy in long-term female survivors of pediatric cancer focusing on the evaluation of candidate genes identified in Secondary Aims 5 and 6.

**Primary Aim 4:** Using a Boolean analysis of the CCSS GWAS dataset, identify genetic risk factors associated with reduced likelihood of having sired a pregnancy in long-term male survivors of pediatric cancer focusing on the evaluation of candidate genes identified in Secondary Aims 5 and 6.

**Secondary Aim 1:** Determine if SNPs reaching genome-wide significance in primary aim 1 may be replicated in an independent cohort from the St. Jude life (SJLIFE) genomics project.

**Secondary Aim 2:** Determine if SNPs reaching genome-wide significance in primary aim 2 may be replicated in an independent cohort from SJLIFE.

**Secondary Aim 3:** Using a candidate gene approach from the CCSS cohort, determine if CYP3A4*1B (rs2740574) and/or CYP2C19*2 (rs12769205, rs4244285) haplotypes are associated with reduced likelihood of having been pregnant in long-term female survivors of pediatric cancer.

**Secondary Aim 4:** Using a candidate gene approach from the CCSS cohort, determine if CYP3A4*1B (rs2740574) and/or CYP2C19*2 (rs12769205, rs4244285) haplotypes are associated with reduced likelihood of having sired a pregnancy in long-term male survivors of pediatric cancer.

**Secondary Aim 5:** Using publicly available pre-existing large gene expression datasets, identify sets of genes associated with the normal development and function of ovaries, testes, ova and sperm. This type of analysis has been performed before to identify developmentally regulated genes in normal B-Cells[22]. In this analysis we used publicly available 4,787 human and 2,167 mouse microarrays on diverse tissue types and diseases. The algorithm started with two known genes KIT and CD19 that were used as endpoints of the developmental pathway and predicted other genes that are supposed to turn on during normal development of B-Cells by using Boolean implication relationships from KIT and CD19. Later we used similar algorithms to identify genes associated with development and differentiation in bladder cancer[23], colon cancer[24] and prostate cancer[25]. Currently, we have a database of publicly available 25,955 human (GSE119087) and 11,758 mouse (GSE119085) microarrays. We will perform Boolean analysis on these databases to identify the genes associated with normal development and function of ovaries, testes, ova and sperm. These sets of genes would be used to identify the candidate SNP's from the CCSS GWAS dataset to be used in statistical analyses to discover associations with pregnancy in female survivors (Primary Aim 3) and siring a pregnancy in male survivors (Primary Aim 4).

**Secondary Aim 6:** Using publicly available pre-existing large gene expression datasets, identify sets of genes associated with the normal development and endocrine function of the pituitary gland as mentioned above. We will start with genes such as OTX2, PITX1, PITX2, HESX1, and TBX19 that are known to be involved with the development of the pituitary gland[26] and use strong Boolean implication relationships to identify other genes. These sets of genes would be used to identify the candidate SNP's from the CCSS GWAS dataset to be used in statistical analyses to discover associations with pregnancy in female survivors (Primary Aim 3) and siring a pregnancy in male survivors (Primary Aim 4).

The expected outcome of this study is the determination that patients with specific SNPs will have decreased rates of pregnancy and siring pregnancy after correcting for other risk factors (types and total doses of alkylating agents and similar drugs,[7] age at follow-up, and age at treatment). This information will be immediately clinically applicable as patients receiving chemotherapy may have genotyping performed soon after diagnosis to assist in risk counseling, and inform treatment decisions and fertility preservation options.

## 5. Analysis framework

### 5.1. Outcome(s) of interest

#### 5.1.1 Primary

Ever having been pregnant or ever having sired a pregnancy (all outcomes combined including live births, miscarriages, abortions)

- Ability to define infertility is limited in the CCSS questionnaires as it needs to take into account marital/cohabitation status, interval of time (at least 1 year) with potential for pregnancy (such data are only available on the baseline but not follow-up surveys), and a desire to become pregnant. Thus, the primary outcome will be focused on pregnancy.[7]
- Sensitivity analyses: will also examine results from participant baseline questionnaire: "Was there ever a period in your life when you and a partner tried for one year or more to become pregnant, without success?"

### 5.2. Subject Population

- Female and Male CCSS GWAS subjects will be included.
- Patients will be stratified for analysis by sex for analysis
- Patients will be excluded if they received cranial or pelvic/gonadal radiotherapy of any dose **(Table 1)**.
- Only subjects exposed to Cyclophosphamide will be included for secondary aims 3 and 4 **(Table 2)**.
- Note: After completing primary aims 1 and 2, we will determine if SNPs which meet genome-wide significance in the CCSS cohort (discovery cohort) remain significant in the replication cohort (SJLIFE). Data will be made available (generously agreed to by Drs. Hudson and Robison) and replication analyses will be conducted by the St. Jude analytic team with oversight by Yutaka Yasui, PhD. Note: CCSS participants will be excluded from the SJLIFE cohort analysis.[15] Primary outcomes and eligibility otherwise per primary aims.

### 5.3. Exploratory Variables

- Patient age at survey
- Patient age at treatment
- Race/ethnicity
- Patient Malignancy
- Patient marital status
- Patient exposure to cyclophosphamide (total dose)
- Cyclophosphamide Equivalent Dose (absolute score, and increments of 4,000mg/m2)
- Patient exposure to radiation (location and dose)
- Patients history of surgical sterility (orchiectomy, oophorectomy, or hysterectomy) and age of surgical sterility

### 5.4. Analytic Approach

We will work with the CCSS statistical team to finalize the appropriate analysis for the proposed study. Descriptive statistics will be generated and compared between survivors with and without pregnancy/ having sired a pregnancy. Cohort risk factors will be described using medians and quartiles or means and standard deviations for continuous variables, and counts and percentages for categorical variables. SAS version 9.4 and R version 3.2 will be used for analysis.

### 5.4.1. Primary Aims 1 and 2

#### 5.4.1.1. Overall Approach

1. Determine the frequency of exposure to other fertility risk factors in the CCSS GWAS study: CED, age at treatment, age at survey completion, marital status, malignancy, and sex (**Table 3)**.
2. Subsequent steps will be stratified by patient sex.
3. Determine the frequency of having ever been pregnant in females and siring a pregnancy among males.
4. Determine the relative frequency of SNP polymorphisms in the cohort, after performing quality control of SNPs. *(Quality Control section below).*
5. Using a multivariate model, determine if SNP polymorphisms are associated with rates of pregnancy or siring a pregnancy are when accounting for other risk factors *(see statistical overview).***(Figure 1, Table 4)**
6. Suggestive SNPs ($P<5x10^{-4}$) from *Primary Aims,* will be tested for the interaction of GxE (CED < OR ≥ 4,000 mg/m2). **(Table 5, Figure 2)**

#### 5.4.1.2. Statistical Overview

In order to assess the association between each genotype and occurrence of pregnancy or siring a pregnancy, we will use multivariable left truncated Cox proportional hazards models,[27] for the age at first pregnancy/siring of pregnancy, with the subjects' risk time starting at the later of 1) age of entrance into the CCSS cohort, or 2) age 15, censored at the age of last CCSS follow-up. Death prior to pregnancy/siring of pregnancy will be treated as a competing risk. For multivariable analysis, we will assume an additive genetic effect, and adjust for covariates including year of diagnosis, age at diagnosis, and treatments known to affect fertility, and for the interaction between the allele (for these meeting genome wide significance) and cyclophosphamide equivalent dose (as a nominal variable). $P<5x10^{-8}$ will be considered statistically significant at the genome-wide level, and hazard ratios and confidence intervals for significant SNPs will be reported.[28] Suggestive SNPs ($P<5x10^{-4}$ as the main effect) will be tested for the interaction of GxE (CED < OR ≥ 4,000 mg/m2). Note: Significant SNPs will be then validated using the replication (SJLIFE) cohort. The same statistical models will be used.

Any SNPs that meet genome-wide significance in both discovery and validation cohorts, as well as any SNPs with at least suggestive significance (P<1e-6) in both cohorts and in high LD ($r^2>0.7$) with these lead SNPs, will undergo functional assessment, including: 1) evaluation as possible expression QTLs in testicular and ovarian tissues found in GxE **(Figure 2)**; 2) assessment of Human Omni 5 annotations including exonic, splice site, and promoter markers; 3) evaluation of overlap with DNA regulatory regions found in the ENCODE database as well as microRNA and lncRNA coding regions; 4?) gene-set enrichment analysis if multiple exonic SNPs are significant

#### 5.4.1.3. Sample size and power considerations

We will compare groups defined by presence/absence of a SNP on pregnancy-free rate using a left-truncated survival model. We assume time 0 is age 15 with no accrual time and maximum follow-up time of 30 years (analyzed ages 15-45), with pregnancy rates by age 45 obtained from subject counts in the study proposal. Pregnancy-free rates are assumed to have piecewise linear

survival curve with proportional hazards: 90% of first pregnancies/ siring of first pregnancy occur evenly between ages 15 and 35 (year 20), 10% of first pregnancies/ sired occur evenly between ages 35 and 45 (year 30). Minimum detectable hazard ratios depend on prevalence of SNP of interest **(Table 6).** Sample size calculations were performed using SAS 9.4 PROC POWER. Note, once subject data is obtained, if a large number of pregnancies occur in females or males >45 years of age we will include participants up to 55 years of age.

### 5.4.2. Secondary Aims 3 and 4

Using a candidate gene approach will significantly reduce the risk of false positive results, and allow the testing of a specific hypothesis.

#### 5.4.2.1. Overall Approach

1. Genotype data will undergo quality control as in the Primary Aims. However, to increase genotype density for the candidate genes, an additional analysis with data imputed with the Human Reference Consortium mixed haplotype panel will be included. Phasing of genotypes will be performed with Eagle, to allow for analysis of the CYP2C19*2 (rs12769205, rs4244285) haplotype.
2. Subsequent steps will be stratified by patient sex.
3. Determine the frequencies of the haplotypes of interest in the CCSS GWAS cohort for: CYP3A4*1B (rs2740574); CYP2C19*2 (rs12769205, rs4244285)
4. Perform analysis using multivariable Cox proportional hazards models with diagnosis, age at diagnosis, year of diagnosis and CED to determine if the above SNPs are significant when accounting for other risk factors (see statistical considerations).***(note for secondary aims 3 and 4 tables will be created similar to tables 3-5 and figure 2 for the group who was exposed to cyclophosphamide and for the polymorphisms listed above)***

#### 5.4.2.2. Variant/ Haplotype Frequency in Candidate Genes

In previous smaller studies of women exposed to cyclophosphamide, heterozygosity or homozygosity for *CYP2C19*2 and CYP3A4*1B* effect the risk of ovarian failure.[12,13] The Variant/ Haplotype Frequency of these polymorphisms in the 1000 Genomes Phase 3 version 5 release are noted in **(Table 7).**

#### 5.4.2.3. Statistical Overview

For the candidate gene analysis univariable and multivariable Cox proportional hazards models for age at pregnancy or siring of pregnancy (as defined for the primary aims) will be constructed. First, we will assess each treatment and risk factor (including year of diagnosis, age at diagnosis, treatments known to affect fertility, and cyclophosphamide dose) in a univariable model to assess which factors are associated with fertility. Then, each allele will be assessed in an individual multivariable model adjusting for the risk factors which were assessed to be important in the univariable models. Model fit and assumptions will be assessed by analyzing Martingale and Schoenfeld residuals. As each of two candidate SNP alleles will be analyzed separately, significance criteria for each analysis will be 0.025 (Bonferroni adjustment).

### 5.4.3. Primary Aims 3 and 4, Secondary Aims 5 and 6

#### 5.4.3.1. Boolean analysis

Boolean analysis is a simple mathematics of two values, i.e., high/low, 1/0, or positive/negative. The gene expression levels are converted to Boolean values (high and low) using StepMiner algorithm.[29] First the expression values are sorted from low to high and a rising step function is fitted to the series to identify the threshold. Boolean analysis is performed to determine relationship between the expression levels of pairs of genes. For example, there are six possible

types of relationships between the expression levels of two genes: two symmetric ones (equivalent and opposite) and four asymmetric ones (low => low, high => low, low => high, high => high).[30] These relationships are called Boolean implication relationships because they are represented by logical implication (=>) formula. BooleanNet statistics is used to assess the significance of the Boolean implication relationships.[30]

Boolean analysis utilizes an "if…then" statement based on gene expression profiles. A Boolean relationship is essentially a logical statement that is either always true or false. An example of a Boolean relationship is: if gene X is high, then gene Y is high. Boolean analysis differs from correlations and associations between genes by offering a platform where asymmetric relationships can be discovered in addition to symmetric ones. https://www.frontiersin.org/files/Articles/25283/fphys-03-00276-HTML/image_m/fphys-03-00276-g001.jpg

A specific Boolean relationship is called an invariant if all samples in a particular universe follow them. We hope to establish Boolean invariants in a specific universe in the context of fertility. The invariant is a stated relationship or variable that is constant. The invariants in a particular category of samples provides a set of logical relationships in that universe.

One way we can use this relationship in biology is by looking at a path of gene differentiation in an organism.[22] In an example pathway, only gene X is present during the more immature stages and only gene Y is present after maturation. In this case, if gene X is high, then gene Y is low and vice versa along the pathway. Hypothetically, we could identify a gene that is high when gene X is low but also high when gene Y is high. Based on this gene's Boolean relationship, we can characterize it as a gene that appears during the middle of the differentiation pathway. Using this principle, we hope to filter gene candidates to test SNPs. Predictive SNPs in different pathways would allow for more accurate estimations of the risk of gonadal damage and infertility in survivors of pediatric cancer.

We have developed a method termed Mining Developmentally Regulated Genes (MiDReG) to predict genes whose expression is either activated or repressed as precursor cells differentiate[22]. MiDReG bases its predictions on Boolean implications mined from large-scale microarray databases and requires two or more "end point" markers for a given developmental pathway. For example, in studies of B cell development, we used two known genes KIT and CD19 that are expressed early and late respectively during B cell development. MiDReG searched for genes X that are expressed during development and satisfy the implications "KIT high => X low" and "CD19 high => X high". There is a robust Boolean implication KIT high => CD19 low is observed in the diverse collection of microarray dataset both in humans and mice. Genes that are expressed at an intermediate step and remain high till the end are discovered by identifying genes with KIT high => X low and CD19 high => X high Boolean implications. After comprehensive review of the literature it was observed that 41 of these 62 predicted genes have been knocked out in mice and 26 of these (63%) exhibit defects in B cell function and differentiation. In addition, MiDReG enabled the discovery of a new a branchpoint in B cell and T cell development[31]. MiDReG is a general method that can be applied to any genes of interest. This proposal is based on application of this tool to increase statistical power of GAWS study by focusing on relevant genes.

By applying the above computational technique of Boolean analysis to large, pre-existing, publically available gene expression datasets (25,955 human; GSE119087 and 11,758 mouse; GSE119085), we will identify genes associated with important molecular pathways within healthy and cancerous tissues that could potentially affect fertility. We would apply these techniques to identify three different sets of genes that would be important in the analysis:

1. Sets of genes associated with the normal development and function of ovaries, testes, ova and sperm. Example seed genes for MiDReG analysis are SOX17 and WT1 for ovary, DDX4, TPTE for testis and sperm.
2. Sets of genes associated with the normal development and endocrine function of the pituitary gland. Example seed genes for MiDReG analysis are OTX2, PITX1, PITX2, HESX1, and TBX19.
3. Sets of genes associated with the pharmacogenomics and pharmacodynamics of chemotherapy.

We will search only the regions around the above identified genes to check if associations exist between candidate SNPs/genes in the CCSS GWAS database and the clinical outcome variables of interest:

1. Achieving pregnancy in long-term female survivors of pediatric cancer
2. Having sired a pregnancy in long-term male survivors of pediatric cancer

## 6. Examples of Specific Tables and Figures

Investigators are asked to provide examples of specific tables and figures that will illustrate important relationships and that may or may not be part of the final manuscript.

| Row | Table 1. Exposures for Primary Aims | Female | Male |
|---|---|---|---|
| **1** | Total CCSS GWAS study participants | 2,958 | 2,781 |
| **2** | Total Row 1 participants not receiving Pelvic/Gonadal or Cranial Radiation | 999 | 1,325 |
| **3** | Total Row 2 participants who have been pregnant ( have sired a pregnancy) | 659 | 717 |
| **4** | Total Row 2 participants that didn't have a pregnancy (or sired a pregnancy) | 340 | 608 |
| **5** | Total Row 4 participants that died, or lost to follow-up, or surgical sterility prior to age 45 | 14 | 70 |

| Row | Table 2. Exposures of Secondary Aims | Female | Male |
|---|---|---|---|
| **1** | Total GWAS study participants | 2,958 | 2,781 |
| **2** | Total Row 1 exposed to cyclophosphamide | 1,119 | 1,139 |
| **3** | Total Row 2 and not receiving Pelvic/Gonadal or Cranial Radiation | 398 | 601 |
| **4** | Total Row 3 participants who have been pregnant ( have sired a pregnancy) | 265 | 308 |
| **5** | Total Row 3 participants that didn't have a pregnancy (or sired a pregnancy) | 133 | 293 |
| **6** | Total Row 5 participants that died, or lost to follow-up, or surgical sterility prior to age 45 | 14 | 70 |

*Table 1 and 2 will presumably be adapted into a figure 1 (Flow Chart) demonstrate the inclusion process for this study. Similar table/figures will be created from the replication cohort*

| **Table 3.** | | | | |
|---|---|---|---|---|
| Characteristic | Pregnant (n=X) | No Pregnant (n=X) | Sired Pregnancy (n=x) | No Pregnancy Sired (n=X) |
| Median age at diagnosis of primary cancer, years (range) | | | | |
| Age at original diagnosis, years, n (%) | | | | |
| <5 | | | | |
| 5-9 | | | | |
| 10-14 | | | | |
| 15-20 | | | | |
| Median age at last follow-up, years (range) | | | | |
| Current age, years, n (%) | | | | |
| <20 | | | | |
| 20-29 | | | | |
| 30-39 | | | | |
| >40 | | | | |
| Race/ethnicity, n (%) | | | | |
| White, non-Hispanic | | | | |
| Black, non-Hispanic | | | | |
| Hispanic | | | | |
| Other | | | | |
| Primary Cancer Diagnosis, n (%) | | | | |
| Leukemia | | | | |
| CNS Tumor | | | | |
| Hodgkin Disease | | | | |
| Non-Hodgkin Lymphoma | | | | |
| Renal Tumors | | | | |
| Neuroblastoma | | | | |
| Soft Tissue Sarcoma | | | | |
| Bone Tumors | | | | |
| Cyclophosphamide exposure, n (%) | | | | |
| Median Cyclophosphamide Dose (range) | | | | |
| Cyclophosphamide Equivalent dose, n (%) | | | | |
| <1 mg/m2 | | | | |
| 1-3,999 mg/m2 | | | | |
| 4,000-7,999 mg/m2 | | | | |
| 8,000-11,999 mg/m2 | | | | |
| ≥12,000 mg/m2 | | | | |
| Radiotherapy | | | | |
| None | | | | |
| Neck | | | | |
| Chest | | | | |
| Arms | | | | |
| Sterilizing Procedure | | | | |

| **Table 4. Multivariate Model** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Women | | | | Men | | | |
| | Pregnant | No Pregnant | Hazard Ratio (95% CI) | *P* | Sired Pregnancy | No Pregnancy Sired | Hazard Ratio (95% CI) | *P* |
| Cyclophosphamide Equivalent dose, n (%) | | | | | | | | |
| <1 mg/m2 | | | | | | | | |
| 1-3,999 mg/m2 | | | | | | | | |
| 4,000-7,999 mg/m2 | | | | | | | | |
| 8,000-11,999 mg/m2 | | | | | | | | |
| ≥12,000 mg/m2 | | | | | | | | |
| | | | | | | | | |
| Age at original diagnosis, years, n (%) | | | | | | | | |
| <5 | | | | | | | | |
| 5-9 | | | | | | | | |
| 10-14 | | | | | | | | |
| 15-20 | | | | | | | | |
| | | | | | | | | |
| *GENE1 rs1234567* | | | | | | | | |
| AA | | | | | | | | |
| GA | | | | | | | | |
| GG | | | | | | | | |
| | | | | | | | | |
| *GENE2 rs7654321* | | | | | | | | |
| AA | | | | | | | | |
| GA | | | | | | | | |
| GG | | | | | | | | |

*Table 4 illustrates the presentation of a Multivariate Hazard Model taking into account variables found to be significant in the Univariate analysis (in this case presuming Cyclophosphamide Equivalent Dose and age at diagnosis)*

| **Table 5** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Women | | | | Men | | | |
| | Pregnant | No Pregnant | Hazard Ratio (95% CI) | *P* | Sired Pregnancy | No Pregnancy Sired | Hazard Ratio (95% CI) | *P* |
| **Gene Environment Interaction (GXE)** | | | | | | | | |
| *GENE1* GXE | | | | | | | | |
| CED <4,000 mg/m2 and AA/GA | | | | | | | | |
| CED <4,000 mg/m2 and GG | | | | | | | | |
| CED <4,000 mg/m2 and AA/GA | | | | | | | | |
| CED <4,000 mg/m2 and GG | | | | | | | | |
| *GENE2* GXE | | | | | | | | |
| CED <4,000 mg/m2 and AA/GA | | | | | | | | |
| CED <4,000 mg/m2 and GG | | | | | | | | |
| CED <4,000 mg/m2 and AA/GA | | | | | | | | |
| CED <4,000 mg/m2 and GG | | | | | | | | |

*Table 5 is an example of 2 genes with GXE that are statistically significant.*

**Table 6.** Minimum detectable hazard ratios for non-pregnancy at age 45,* subjects with vs subjects without polymorphism, with power = 80% at specified Type I error rate and sample size, and cumulative pregnancy outcome rates of 54% for males and 67% for females in the GWAS cohort, and 51% for males and 67% for females in the candidate gene cohort.

| Allele or SNP prevalence (%) | GWAS analyses: significance criteria $\alpha = 5x10^{-8}$ | | Candidate gene analyses: significance criteria $\alpha = 0.025$ | |
|---|---|---|---|---|
| | Males, N=1,325 | Females, N=999 | Males, N=601 | Females, N=398 |
| 5% | 3.5 | 4.3 | 2.5 | 3.4 |
| 10% | 2.5 | 3.0 | 2.0 | 2.5 |
| 15% | 2.2 | 2.5 | 1.8 | 2.1 |
| 22%# | 2.0 | 2.2 | 1.6 | 1.9 |
| 25% | 1.9 | 2.1 | 1.6 | 1.9 |
| 50% | 1.7 | 1.9 | 1.5 | 1.7 |
| 77%& | 1.8 | 2.0 | 1.6 | 1.8 |

Minimum detectable hazard ratio for two-sided log-rank test comparing subjects with and without SNP or allele of interest at 80% power. Significance criteria for GWAS analyses will be $5x10^{-8}$. Significance criteria for candidate gene study will be controlled at a false discovery rate of 0.025, here estimated with a Bonferroni correction for 2 tests. *No reported pregnancy or siring of pregnancy, regardless of pregnancy outcome. Censoring rates were calculated assuming 52% of the cohort is currently aged 31-45 years and has the same overall pregnancy/siring rate as the analysis cohort; censored patients consist of subjects aged 31-45 with non-pregnancy plus the specified number of subjects who died, were lost to follow-up, or underwent surgical sterility; #frequency of CYP2C9*2; &frequency of CYP3A4*1B

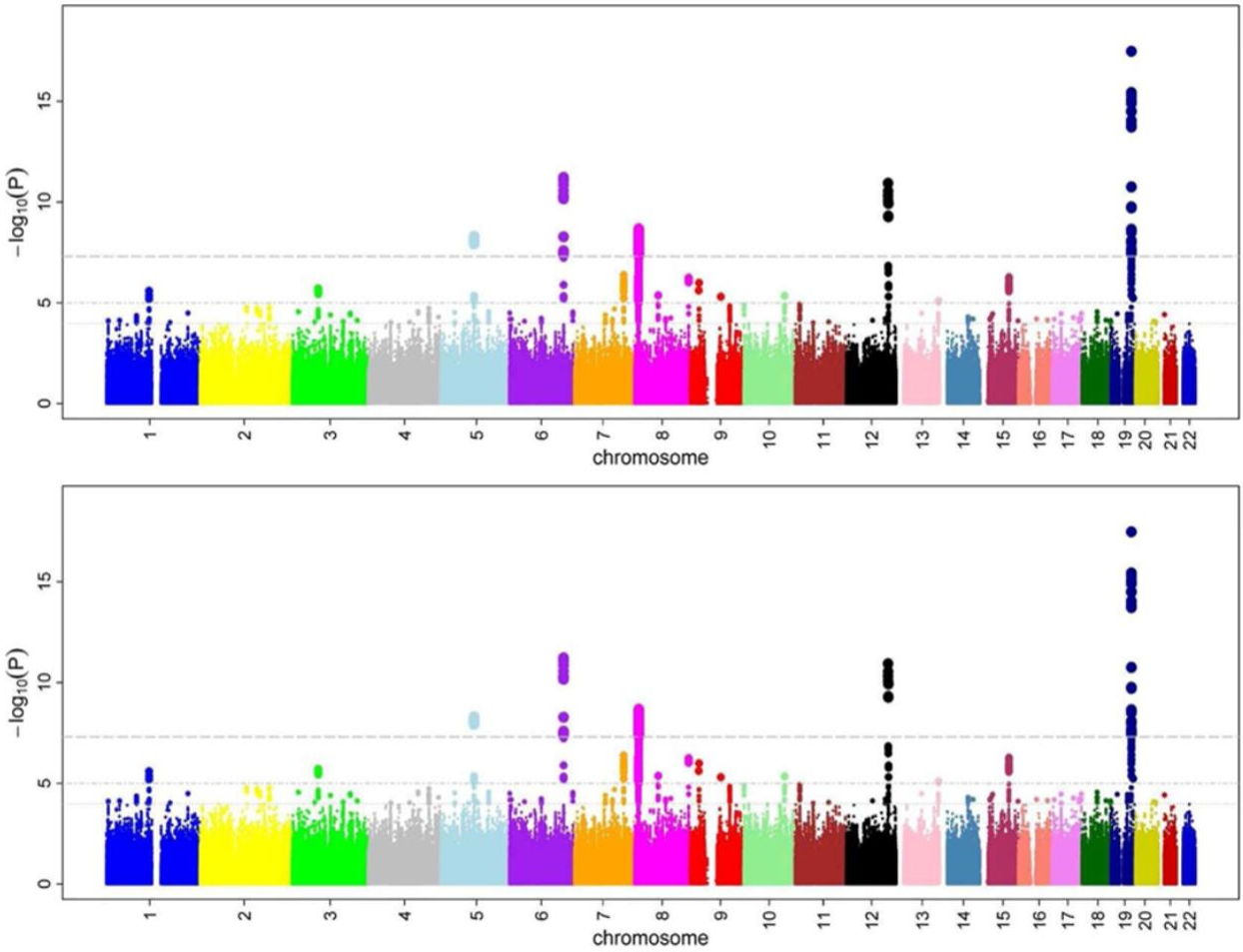
*Table 6 for illustration of power calculation, unlikely to be included in a final manuscript*

**Table 7. Gene Frequency**

| Gene | Polymorphism | Variant/ Haplotype Frequency[32] |
|---|---|---|
| CYP2C19 | *2 (rs12769205, rs4244285) | 0.2212 |
| CYP3A4 | *1B (rs2740574) | 0.7692 |

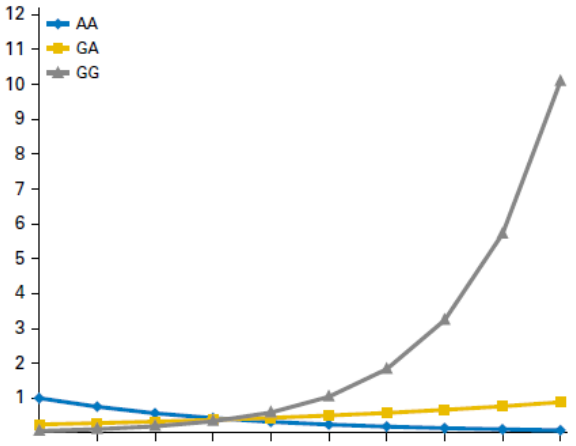
*Table 7 for illustration of analytic approach, unlikely to be included in a final manuscript*

**Figure 1.**




*Example of Manhattan plot to be used for potential Univariate analysis of GWAS and Pregnancy/ Suring a pregnancy. P<5x10-8 will be considered statistically significant at the genome-wide level. Plots will likely have an A and B panel to segregate by sex. Credit: https://en.wikipedia.org/wiki/Manhattan_plot*

**Figure 2**




*Example Figure of OR of Pregnancy or Siring a pregnancy based on genotype and CED in a GXE. CED would be on the X-axis, Hazard Ratio on the T-axis. Separate plots would be generated for male and female. As OR will be calculated for pregnancy (as opposed to not pregnancy) the shape of the curve would be different than the one pictured above. Credit (Wang X…Bhatia S., JCO. 2016)*

## 7. Special consideration

### 7.1. Additional Information

#### 7.1.1. Quality Control for GWAS Data

We will perform seven quality control (QC) steps for the CCSS GWAS data. (1) Missingness. We will exclude SNPs missing in at least 20% of the population. We will also exclude individuals with >80% genotype missingness; (2) Sex discordance. We will check for sex discordance based on the recorded data and the X chromosome homozygosity; (3) Minor allele frequency. We will only include SNPs with MAF>0.05; (4) Hardy Weinberg Equilibrium (HWE). We will exclude markers that deviate from HWE (P<1e-6); (5) Heterozygosity. We will remove individuals who deviate ±3 SD from the mean heterozygosity rate; (6) Relatedness. We will calculate identity by descent (IBD) of all sample pairs and exclude highly possible duplicates or relatives (if PLINK pi-hat >0.20 between a pair, the sample with the highest sample missingness rate will be removed); (7) Population structure. To maximize sample size, all individuals will be included in the analysis. If there is evidence of population structure, assessed by QQ plots of SNP main effect p-values or a genomic inflation factor>1.2, a second analysis will be tailored to European-ancestry individuals. Namely, after excluding high-LD regions (e.g. HLA region on chromosome 6) and LD pruning, we will perform principal components analysis to identify European and non-European ancestry groups (compared to the 1000 genomes ancestry groups). To account for any residual population structure, principal component analysis will be performed again without these individuals and the top 10 principal components will be tested for association with time to pregnancy/ siring a pregnancy and included as covariates if significant (p<0.05).

After conducting QC, we will perform imputation with the University of Michigan Imputation server using the 1000 Genomes Phase 3 version 5 mixed haplotype panel.[15] Post-imputation quality control will exclude monomorphic, extremely rare (MAF < 5%) and low confidence (minimac AvgCall <95%, Rsq <0.5) SNPs.

### 7.1.2. Limitations

This proposal is not without limitations. First, having been pregnant or siring a pregnancy does not necessarily capture the true burden of infertility. CCSS data relies on patient self-reports of pregnancy and pregnancies may go unnoticed in females, and siring a pregnancy as well as non-paternity may occur in males. Additionally, CCSS data does not capture patient's intent or desire to become pregnant, use of contraceptives, or use of fertility techniques such as *in vitro* fertilization. Additionally, we have chosen as an endpoint in this study pregnancy, and not live-birth and certain factors in survivors could conceivable play a role in miscarriage. However, we have chosen to exclude patients with pelvic radiation, so such effects as uterine restriction from radiation should be controlled for. Despite these limitations, we feel this proposal will add valuable knowledge to our understanding of the ability of pediatric cancer survivors to become pregnant and sire a pregnancy. We believe the study is sufficiently powered to detect clinically significant genetic polymorphisms affecting pregnancy in long-term cancer survivors **(Table 4)**. However, in the event that polymorphisms do not reach the level of statistical significance for our primary aims ($\alpha=5x10^{-8}$) we would discuss amending the protocol (with the CCSS Genetics Working Group) considering a candidate gene approach for cyclophosphamide exposed patients at additional enzymes known to be involved in cyclophosphamide metabolism (CYP2B6, CYP2C9, CYP2C19, CYP3A4, CYP3A5, CYP2A6, CYP2C8, GSTA1, GSTAM1, GSTP1, GSH1, GSTT1).

## 7.2. References


1. Wenzel L, Dogan-Ates A, Habbal R, et al. Defining and measuring reproductive concerns of female cancer survivors. *J Natl Cancer Inst Monogr*. 2005(34):94-98.
2. Zebrack BJ, Casillas J, Nohr L, Adams H, Zeltzer LK. Fertility issues for young adult survivors of childhood cancer. *Psychooncology*. 2004;13(10):689-699.
3. Keros V, Hultenby K, Borgstrom B, Fridstrom M, Jahnukainen K, Hovatta O. Methods of cryopreservation of testicular tissue with viable spermatogonia in pre-pubertal boys undergoing gonadotoxic cancer treatment. *Hum Reprod*. 2007;22(5):1384-1395.
4. Kim SY, Kim SK, Lee JR, Woodruff TK. Toward precision medicine for preserving fertility in cancer patients: existing and emerging fertility preservation options for women. *J Gynecol Oncol*. 2016;27(2):e22.
5. Picton HM, Wyns C, Anderson RA, et al. A European perspective on testicular tissue cryopreservation for fertility preservation in prepubertal and adolescent boys. *Hum Reprod*. 2015;30(11):2463-2475.
6. Green DM, Kawashima T, Stovall M, et al. Fertility of female survivors of childhood cancer: a report from the childhood cancer survivor study. *J Clin Oncol*. 2009;27(16):2677-2685.
7. Chow EJ, Stratton KL, Leisenring WM, et al. Pregnancy after chemotherapy in male and female survivors of childhood cancer treated between 1970 and 1999: a report from the Childhood Cancer Survivor Study cohort. *Lancet Oncol*. 2016;17(5):567-576.
8. Barton SE, Najita JS, Ginsburg ES, et al. Infertility, infertility treatment, and achievement of pregnancy in female survivors of childhood cancer: a report from the Childhood Cancer Survivor Study cohort. *Lancet Oncol*. 2013;14(9):873-881.
9. Wasilewski-Masker K, Seidel KD, Leisenring W, et al. Male infertility in long-term survivors of pediatric cancer: a report from the childhood cancer survivor study. *J Cancer Surviv*. 2014;8(3):437-447.
10. Aubier F, Flamant F, Brauner R, Caillaud JM, Chaussain JM, Lemerle J. Male gonadal function after chemotherapy for solid tumors in childhood. *J Clin Oncol*. 1989;7(3):304-309.
11. Kenney LB, Laufer MR, Grant FD, Grier H, Diller L. High risk of infertility and long term gonadal damage in males treated with high dose cyclophosphamide for sarcoma during childhood. *Cancer*. 2001;91(3):613-621.

12. Singh G, Saxena N, Aggarwal A, Misra R. Cytochrome P450 polymorphism as a predictor of ovarian toxicity to pulse cyclophosphamide in systemic lupus erythematosus. *J Rheumatol*. 2007;34(4):731-733.
13. Su HI, Sammel MD, Velders L, et al. Association of cyclophosphamide drug-metabolizing enzyme polymorphisms and chemotherapy-related ovarian failure in breast cancer survivors. *Fertil Steril*. 2010;94(2):645-654.
14. Green DM, Kawashima T, Stovall M, et al. Fertility of male survivors of childhood cancer: a report from the Childhood Cancer Survivor Study. *J Clin Oncol*. 2010;28(2):332-339.
15. Brooke RJ, Im C, Wilson CL, et al. A High-risk Haplotype for Premature Menopause in Childhood Cancer Survivors Exposed to Gonadotoxic Therapy. *J Natl Cancer Inst*. 2018.
16. Joshi N, Chan JL. Female Genomics: Infertility and Overall Health. *Semin Reprod Med*. 2017;35(3):217-224.
17. Aston KI. Genetic susceptibility to male infertility: news from genome-wide association studies. *Andrology*. 2014;2(3):315-321.
18. Kosova G, Scott NM, Niederberger C, Prins GS, Ober C. Genome-wide association study identifies candidate genes for male fertility traits in humans. *Am J Hum Genet*. 2012;90(6):950-961.
19. Krausz C, Escamilla AR, Chianese C. Genetics of male infertility: from research to clinic. *Reproduction*. 2015;150(5):R159-174.
20. Green DM, Nolan VG, Goodman PJ, et al. The cyclophosphamide equivalent dose as an approach for quantifying alkylating agent exposure: a report from the Childhood Cancer Survivor Study. *Pediatr Blood Cancer*. 2014;61(1):53-67.
21. Kenney LB, Cohen LE, Shnorhavorian M, et al. Male reproductive health after childhood, adolescent, and young adult cancers: a report from the Children's Oncology Group. *J Clin Oncol*. 2012;30(27):3408-3416.
22. Sahoo D, Seita J, Bhattacharya D, et al. MiDReG: a method of mining developmentally regulated genes using Boolean implications. *Proc Natl Acad Sci U S A*. 2010;107(13):5732-5737.
23. Volkmer JP, Sahoo D, Chin RK, et al. Three differentiation states risk-stratify bladder cancer into distinct subtypes. *Proc Natl Acad Sci U S A*. 2012;109(6):2078-2083.
24. Dalerba P, Sahoo D, Paik S, et al. CDX2 as a Prognostic Biomarker in Stage II and Stage III Colon Cancer. *N Engl J Med*. 2016;374(3):211-222.
25. Sahoo D, Wei W, Auman H, et al. Boolean analysis identifies CD38 as a biomarker of aggressive localized prostate cancer. *Oncotarget*. 2018;9(5):6550-6561.
26. Kelberman D, Rizzoti K, Lovell-Badge R, Robinson IC, Dattani MT. Genetic regulation of pituitary gland development in human and mouse. *Endocr Rev*. 2009;30(7):790-829.
27. Morton LM, Sampson JN, Armstrong GT, et al. Genome-Wide Association Study to Identify Susceptibility Loci That Modify Radiation-Related Risk for Breast Cancer After Childhood Cancer. *J Natl Cancer Inst*. 2017;109(11).
28. Sobota RS, Shriner D, Kodaman N, et al. Addressing population-specific multiple testing burdens in genetic association studies. *Ann Hum Genet*. 2015;79(2):136-147.
29. Sahoo D, Dill DL, Tibshirani R, Plevritis SK. Extracting binary signals from microarray time-course data. *Nucleic Acids Res*. 2007;35(11):3705-3712.
30. Sahoo D, Dill DL, Gentles AJ, Tibshirani R, Plevritis SK. Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biol*. 2008;9(10):R157.
31. Inlay MA, Bhattacharya D, Sahoo D, et al. Ly6d marks the earliest stage of B-cell specification and identifies the branchpoint between B-cell and T-cell development. *Genes Dev*. 2009;23(20):2376-2381.
32. Zhang W, Ng HW, Shu M, et al. Comparing genetic variants detected in the 1000 genomes project with SNPs determined by the International HapMap Consortium. *J Genet*. 2015;94(4):731-740.