Risk Prediction for Nonsurgical Premature Menopause in Childhood Cancer Survivors

by

Rebecca Anne Clark

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Epidemiology

School of Public Health
University of Alberta

Abstract

Childhood cancer survivorship has increased drastically over the previous several decades, consequently increasing the frequency of chronic conditions in survivors. Female childhood cancer survivors are at an increased risk of developing nonsurgical premature menopause (NSPM) due to toxicities from their treatment. NSPM occurs when ovarian function is retained for at least 5 years following cancer diagnosis, but menopause develops naturally before age 40. Such a condition can negatively impact quality of life and reduce potential reproductive years. The literature details risk factors including an older age at cancer diagnosis, and treatment with high doses of alkylating agents and radiation. In order to aid physicians, patients and their families have informed discussions regarding fertility preservation, I aimed to develop prediction algorithms of the absolute risk an individual has of developing NSPM.

The Childhood Cancer Survivor Study cohort was the primary data source for this project. Due to the presence of both stratified random sampling and participant loss to follow-up within the cohort, I initially investigated methods for combining sampling and censoring weights in the estimation of model accuracy measures to aid in model evaluation. I designed and implemented four simulation studies, varying the relationship between sampling design, censoring distribution and risk score distribution, and assessed weighting scenarios with distinct combinations of censoring and sampling weights. Depending on the study setting, different weighting scenarios gave reasonable estimates, and ignoring or inadequately accounting for weights resulted in biased accuracy estimates.

Candidate risk prediction models were developed on a training set of 4,054 observations from the Childhood Cancer Survivor Study cohort using a time-specific logistic regression model with competing risks (TLR-CR), a Fine-Gray regression (FGR) model and a random survival forest model with competing risks (RSF-CR). Model performance and accuracy were measured using the time-specific area under the ROC curve (AUC_t), the time-specific average positive predictive value (AP_t), and calibration curves on both the training set and an internal validation set of 1,454 observations.

Model accuracy values and curves were presented for 15 years post cancer diagnosis as an illustration of overall model performance. All three models performed similarly on the training set. The estimated AUC_t values decreased when internal validation was conducted; however AP_t values were still larger than the event rate. The AP_t / Event Rate ratio for the TLR-CR model increased from the training set performance. AUC_t and AP_t values on the test set calculated over 10-20 years post cancer diagnosis displayed similar findings. The models were well calibrated for low risk patients, however only the TLR-CR model was consistently well calibrated for high risk patients on both datasets. Moving forward, model performance on individuals with clinically verified ovarian status will be assessed through validation on an external cohort. The future practical application of the risk estimates as a risk scoring system aims to have a positive impact on the quality of life of survivors well into their adulthood.

Preface

This thesis is an original work by Rebecca Anne Clark. The research project, of which this thesis is apart, received research ethics approval from the University of Alberta Research Ethics Board, Project Name “Risk prediction for ovarian failure in childhood cancer survivors”, No. Pro00067066, August 22, 2016.

For my family

Acknowledgements

I would like to give a big thank you to everyone who has supported me through the process of writing my thesis. Particularly I would like to thank Dr. Yan Yuan, my supervisor and mentor who has been there for me from the moment I entered my graduate program; encouraging me to become more independent and to take initiative. Without her guidance, I would not be as confident and assured as I am today.

I would like to thank my committee members, Dr. Paul Nathan and Dr. Yutaka Yasui, for their advice and dedication to my work. I would also like to thank the remaining members of the project team, including Dr. Tarek Motan and Dr. Melanie Barwick who provided valuable insight and feedback on my research directions, as well as Dr. Charles Sklar and Dr. Sogol Mostoufi-Moab.

I would like to thank my parents and family for always believing in me, and for being motivating when I needed it the most. Their support for my decision to move across the country was unwavering, and I would not have been able to make the journey without such a reassuring family. I want to thank my friends, both in Alberta as well as back home in Ontario, for being encouraging and keeping me sane for these past years!

I would like to thank all the members of Dr. Yuan's research group, past and present, for listening to my presentations and providing valuable feedback. In particular, I want to acknowledge Dr. Khanh Vu for his guidance, advice, and for always taking the time to help me.

I would like to say a huge thank you to the Women & Children's Health Research Institute (WCHRI) and the Stollery Children's Hospital Foundation for supporting me through the WCHRI Graduate Studentship and the WCHRI trainee travel grant, without which I would not have had the many opportunities to share my research at various conferences. I would also like to thank the Canadian Institutes for Health Research for supporting this project.

I would like to acknowledge and thank the Childhood Cancer Survivor Study for providing the dataset for this research, as well as thank the study participants and their families for taking the time and effort to complete the surveys. I would like to acknowledge and thank the individuals from the Fred Hutchinson Cancer Research Center in Seattle for their excellent data management ability, and for answering my many long lists of questions!

Finally, I would like to thank the School of Public Health at the University of Alberta for providing me with this opportunity, and the professors and staff who always went the extra mile to research answers to my questions. Throughout these past years I have been given many occasions to grow, to prove myself, and to develop as a researcher. I feel privileged to have established such a solid foundation to move forward with in the world of academia.

Table of Contents

List of Figures

List of Tables

List of Appendices

List of Appendix Figures

List of Appendix Tables

List of Abbreviations

ALL	Acute lymphoblastic leukemia
AMH	Anti-Mullerian hormone
AP	Average positive predictive value
AOF	Acute ovarian failure
AUC	Area under the receiver operating characteristic curve
AUCPR	Area under the precision recall curve
BMT	Bone marrow transplantation
CCS	Childhood cancer survivor
CCSS	Childhood Cancer Survivor Study
CED	Cyclophosphamide equivalent dose
CI	Confidence interval
CIF	Cumulative incidence function
CNS	Central nervous system
CSH	Cause specific hazard
CV	Cross validation
DNA	Deoxyribonucleic acid
FGR	Fine-Gray regression
FN	False negative
FP	False positive
FSH	Follicle stimulating hormone
Gy	Gray

HPO	Hypothalamic-pituitary-ovarian
HR	Hazard ratio
ICD-O	International Classification of Diseases for Oncology
IPCW	Inverse probability-of-censoring weight
IR	Incidence rate
KM	Kaplan-Meier
LTFU	Loss to follow-up
MH	Menstrual health
NSPM	Nonsurgical premature menopause
OR	Odds ratio
PH	Proportional hazards
PNET	Primitive neuroectodermal tumor
PPV	Positive predictive value
PY	Person-years
PR	Precision-Recall
RF	Random forest
ROC	Receiver operating characteristic curve
RR	Relative risk
RSF	Random survival forest
RSF-CR	Random survival forest with competing risks
RT	Radiation therapy
SCT	Stem cell transplantation
SEER	Surveillance, Epidemiology, and End Results

SD	Standard deviation
SHR	Subdistribution hazard ratio
SJLIFE	St Jude Lifetime Cohort Study
SMN	Second malignant neoplasm
SPM	Surgical premature menopause
TBI	Total body irradiation
TLR	Time-specific logistic regression
TLR-CR	Time-specific logistic regression with competing risks
TN	True negative
TP	True positive
US	United States
USA	United States of America
VIMP	Variable importance

1 Introduction

Advancements in cancer treatment over the previous few decades have dramatically increased the survival rate of childhood cancer[1]. Today, over 80% of children diagnosed with cancer will survive for more than 5 years, and it is estimated that by 2020 there will be close to 500,000 childhood cancer survivors (CCSs) living in the United States (US)[2]. However, CCSs are at an increased risk of developing long-term morbidities resulting from their primary cancer as well as its therapies[3,4]. These long-term conditions, known as *late effects*, have the ability to impact any organ system and can present through conditions including myocardial infarction, congestive heart failure, as well as premature gonadal failure[4]. Late effects can appear during treatment, shortly after its completion, or even years in the future. Recent studies have indicated that all long term CCSs will develop at least one late effect in their lifetime as a direct consequence of their previous treatment, emphasizing how frequently these conditions are observed[5].

A predominant late effect arising in female childhood cancer survivors is premature ovarian dysfunction. Premature ovarian dysfunction can be categorized into acute ovarian failure (AOF) and nonsurgical premature menopause (NSPM), dependent upon when ovarian function is compromised. AOF is diagnosed when an individual either never experiences menarche following cancer treatment, or permanently ceases to menstruate within 5 years of being diagnosed and treated with cancer[6]. Approximately 6.3% of female CCSs (215 out of 3,390 total) developed AOF once their treatment was finished as measured in a retrospective cohort study conducted by Chemaitilly et al. in 2006[6]. Conversely, NSPM occurs when females maintain regular ovarian function for a minimum of 5 years after cancer diagnosis and treatment, but menstruation stops naturally for at least 6 months before reaching age 40 (not resulting from

pregnancy, surgery or medication)[7]. In the general population, the prevalence of premature menopause is approximately 1%[8], whereas a 2018 study by Levine et al. reported that 9% of CCSs develop NSPM by age 40[9]. Additionally, cancer survivors have 10.5 times (95% Confidence Interval (CI) = (4.2, 26.3)) the risk of developing NSPM compared to their otherwise healthy siblings[9].

There is an urgent need to identify individuals at a high risk of developing NSPM, due to the impact of this condition in CCSs and the potential time-sensitivity of intervening. Although risk factors have been identified for NSPM, physicians lack the ability to obtain a risk estimate for specific patients. Therefore, using data from the Childhood Cancer Survivor Study (CCSS), a retrospective cohort study of 5-year cancer survivors from across North America, the main aim of this research is to develop prediction models which will predict the absolute risk of an individual childhood cancer patient developing NSPM.

Future extensions of the model will ideally provide oncologists, family physicians and obstetricians with personalized risk estimates of their patients developing NSPM. Collaboration with knowledge translation experts will help to facilitate the development of a risk scoring system, which can be applied in a clinical setting by practitioners to aid in discussions of fertility preservation. Individuals with a risk estimate above a threshold risk can require further discussion of fertility preservation either immediately or in the years following treatment completion. If the risk of developing NSPM is low until the patient reaches their mid to late 20s, then discussions and decision making can be deferred until after treatment is completed. If the risk is high immediately post treatment, discussions can occur beforehand, potentially allowing

for the preservation of reproductive opportunities in the future. Individuals at a low risk of developing NSPM at any time can be consoled and spared from undergoing unnecessary procedures. Ultimately, the research goal is to improve the lives and wellbeing of female cancer survivors well into their adulthood, and to ensure that any lasting impacts from treatment exposures are considered and appropriate action is taken.

This thesis is structured as follows. The remainder of Chapter 1 reviews the literature on premature menopause and risk factors in CCSs, as well as describing the statistical methods to be used during model development and evaluation. Chapter 2 explores the characteristics of the CCSS dataset obtained. Chapter 3 presents simulation studies to assess various weighting methods for model evaluation. Chapter 4 highlights model development and evaluation. Finally, Chapter 5 summarizes the findings, discusses study limitations, and provides recommendations for future research.

1.1 Literature Review

1.1.1 Background

Menopause is an inevitable life stage in surviving females[10]. *Natural menopause* is defined as the permanent cessation of menstruation (amenorrhea for at least 12 months) resulting from the loss of ovarian follicular activity[11]. It follows the period known as the *menopausal transition* (or "perimenopause"), which is characterized by menstrual variability and the increased frequency of anovulatory cycles (menstrual cycles where no ovulation occurs)[11]. In the general population, natural menopause occurs at an average age of 50.4 years, with the majority of women entering menopause between ages 45 and 55[4,12]. Early menopause is defined as menopause between ages 40 and 45, and premature menopause is defined as menopause before age 40[12].

Physiology of Menopause

Ovarian follicles are the functional units of the ovaries, with each ovarian follicle containing theca and granulosa cells as well as an immature oocyte[13]. The immature oocyte has the potential to mature into an egg in preparation for fertilization[14]. The quantity of ovarian follicles in the ovaries reaches a peak level of 7 million[15] at approximately 6 months gestation, and begins to decline exponentially thereafter[4,13,16]. By the age of puberty, roughly 300,000 ovarian follicles remain, and only around 400 mature oocytes will be released during ovulation over an entire lifetime[4,13,16]. The number and quality of ovarian follicles in the ovaries defines an individual's *ovarian reserve*, which in itself determines the probability of successful reproduction[16]. Following ages 35-37, the exponential decrease in ovarian follicle reserve is accelerated[4], and it is suggested that among other factors menopause is triggered when the number of follicles falls

below a set level[7]. In postmenopausal women, the number of follicles remaining is minimal, and it is possible for none to be present at all[10]. The decrease in ovarian follicle reserve (therefore, the decrease in eggs available for fertilization) along with the increase in anovulatory cycles leads to the decline in reproductive capability of females with advancing age[11].

The cycle of ovarian follicle development and stimulation is regulated by hormones released from the hypothalamus and pituitary gland. The gonadotropin hormones, (follicle stimulating hormone (FSH) and luteinizing hormone), are released from the anterior pituitary and stimulate the growth of a subset of approximately 30 dormant ovarian follicles, of which one follicle will attain more rapid growth and become the dominant follicle[10,13]. These hormones are under negative feedback from ovarian steroids and inhibins[13]. FSH in particular is inhibited by inhibin B, which is produced by the group of ovarian follicles along with estradiol[10]. As the number of follicles decreases with advancing age, the production of inhibin B and estradiol are decreased[10,11]. These low inhibin B levels no longer inhibit the secretion of FSH thereby allowing it to increase in concentration[10,11]. Elevated FSH levels are a characteristic present during the initial transition to the early menopausal phase. However, in women experiencing regular menstrual cycle length it is unlikely that this change would be detected without testing[11]. The anti-Mullerian hormone (AMH) is also produced by the group of ovarian follicles, and similar to inhibin B, decreases in concentration with the decline of follicle numbers[16]. Monitoring AMH levels can provide an estimate of the magnitude of the ovarian reserve. A high concentration of AMH is associated with sufficient follicle numbers, and lower concentrations with limited reserve[16].

Physiologically, women may begin to notice irregularity in menstrual cycle length and an increase in the frequency of anovulatory cycles resulting from low estradiol and high FSH levels indicative of the perimenopause stage[10,11]. Subsequently, the late menopausal phase is characterized by bouts of amenorrhea for at least 60 days, which eventually culminates in the final menstrual period signifying that menopause has been reached[10,11]. High FSH levels accompanying low inhibin B and estradiol levels are prominent characteristics around the time of the final menstrual period and continue into the stage of post-menopause[11]. Along with the distinctive loss of menstruation, menopause can include symptoms such as hot flashes, night sweats, and an increase in the rate of bone loss due to the absence of estradiol in the brain and ovaries[10].

Impacts of Premature Menopause

There are a variety of ways in which the process of ovulation and menstruation can be jeopardized prematurely, but regardless of the cause, any form of early ovarian dysfunction negatively impacts the quality of life of women, and increases the probability of developing chronic diseases[10,17-19]. Women entering menopause prematurely undergo the loss of important ovarian hormones (such as estrogen), and are at an increased risk of overall mortality, as well as for developing conditions such as osteoporosis and various cardiovascular diseases (including coronary heart disease)[10,17,18].

Women with premature menopause are significantly more likely to record lower values of physical and mental health compared to women with normal ovarian function, as well as reporting a decreased overall quality of life[19]. Informational, social and emotional support from

the diagnosing physician were mentioned as being absent when women received their diagnosis of NSPM, and many recalled the entire experience as traumatic[19]. Additionally, a higher proportion of women with premature ovarian dysfunction were diagnosed with anxiety and depression compared to the general population[8,17,20]. An increase in the quantity of psychological support around the time of diagnosis with ovarian dysfunction was mentioned as potentially being useful to alleviate a portion of the stress accompanying the diagnosis[19].

A primary concern of women diagnosed with ovarian dysfunction is reproductive inability, and the associated feelings of inadequacy and shame from infertility[19,21]. In a study by Singer et al. in 2011 specifically conducted to observe the experiences of women with premature menopause, 92% of participants indicated that the impact on fertility was a major consequence of their diagnosis, and 75% of participants identified infertility as a specific concern[19]. The lack of reproductive ability is reported to be linked to a lower sense of self-worth and feelings of abnormality[19]. Many women felt fearful of being rejected by potential partners due to the humiliation of their condition and therefore were hesitant to disclose their infertility to others[19,20]. The inability to conceive can cause negative emotions such as jealousy and resentment toward friends and family members with children, due to the emphasis that is placed on women to become mothers[20].

Fertility Preservation

With regards to childhood cancer patients, future fertility potential may not be at the forefront of a patient or her family's mind as she prepares to initiate cancer treatment, particularly if the patient is very young. However, it is important to be sufficiently informed of the potential

consequences that a treatment plan may pose later in life before it is undertaken. Compromised reproductive ability may be highly likely to occur based on the proposed doses and agents that are to be used during cancer treatment. Interventions to offer reproductive opportunities in the future should be evaluated and discussed if deemed necessary (this is termed *oncofertility counselling*[22]). Unfortunately, there is evidence that fertility procedures are not being discussed with patients and their families as often as they should be[23]. When pressed for reasons for not exploring these options, physicians cited "not at a significant risk" and "too young" in 29% and 27% of circumstances respectively[23].

Fertility preservation services may be feasible for childhood cancer patients[24]. Oocyte and ovarian tissue cryopreservation are two procedures that are worth mentioning in particular. Oocyte cryopreservation involves extracting and freezing oocytes for future use, and is currently the preferred method for female cancer patients[22,25,26]. Cryopreservation is the process of freezing cells such that all biological functions are arrested, with the intention of later allowing them to thaw and resume normal function[26]. Although the patient is required to be post-pubescent, the oocytes do not need to be fertilized, which makes it a viable option for those without a partner. The procedure requires ovarian stimulation which may take up to a few weeks[25,27]. This may not be possible beforehand if treatment needs to be initiated immediately following diagnosis, though may be completed post treatment before entering premature menopause[4,6].

Ovarian stimulation is the process of suppressing the pituitary gland and stimulating the ovaries to induce follicular growth and mature the oocytes[22,25]. The mature oocytes can then be harvested

through oocyte pick up, and frozen through cryopreservation for future use. Ovarian stimulation itself can pose challenges and potential drawbacks, as it requires the use of a transvaginal ultrasound and oocyte pickup using a needle, which can be traumatic and painful without general anesthetic[28]. As oocyte cryopreservation in cancer patients is still a relatively new procedure, there are limited measures of its success in attaining pregnancy and live birth[22]. However, live births have been reported in the small sample of cancer survivors who thawed and used cryopreserved oocytes[22]. Data has indicated that the rate of pregnancy from the use of cryopreserved and thawed oocytes is comparable between cancer survivors and the general population[22].

Alternatively, ovarian tissue cryopreservation, although still considered experimental, has demonstrated significant advantages over alternate methods in many cases. It is performed without ovarian stimulation and therefore is the only option for prepubertal girls as well as those who are time-sensitive, allowing the procedure to be completed as quickly as necessary[4,28,29]. Methods for performing ovarian tissue cryopreservation were described in a study published in 2016 conducted by Abir et al[30]. Patients underwent general anesthetic for the laparoscopic removal of ovarian tissue[30]. A partial oophorectomy was performed for post-pubertal adolescents due to the larger volume of ovarian tissue available, whereas a complete oophorectomy was required for prepubertal females with small ovaries[30]. The retrieved ovarian tissue was subsequently sliced and cryopreserved for future use[30].

While promising, oocyte and ovarian tissue cryopreservation can unfortunately be invasive, expensive, and particularly traumatic to young patients[4,25]. Complications, such as bleeding and

infection, can occur in both procedures[27,31]. Ovarian tissue cryopreservation has increased risks due to the invasive nature of the surgery required to obtain the tissue, as well as the potential to reintroduce the cancer cells back into the individual upon use of the tissue[4,30]. Therefore, it is crucial to identify whether fertility preservation interventions should be discussed right away or if it is safe to postpone the discussion until the individual is older without compromising future reproductive potential.

1.1.2 Risk Factors for Premature Menopause in the General Population

Certain lifestyle characteristics have been identified to be significantly independently associated with a decrease in the age at natural menopause in the general population of women. Analysis has shown that smoking increases the risk of premature menopause by 43% in otherwise healthy women[32]. Women recruited in the study by Hyland et al. in 2015 who were ever smokers had a significantly higher odds ratio (OR) for developing premature menopause compared to those women who were never-smokers (OR: 1.27, 95% CI = (1.18, 1.37))[33]. Women who began smoking before the age of 15 were menopausal approximately 21.6 months earlier than those women who were never-smokers[33]. After adjusting for other factors, women who were never-smokers but were exposed to high doses of second-hand smoke attained menopause 13 months earlier than the standard average menopausal age (OR = 1.17 (95% CI = (1.05,1.30)) indicating that any tobacco exposure may impact the age a woman enters menopause[33].

Furthermore, exposure to tobacco is hypothesized to influence the development of many adverse late effects in cancer survivors, as well as impacting the development of second malignant

neoplasms (SMN)[34]. Nevertheless, a substantial number of CCSs are current smokers, albeit at a lower rate than in the general population (14% of CCSs based on the CCSS follow-up 2003 survey vs. 29% in the general population)[35]. Specifically, data on smoking patterns in survivors of childhood cancer show that 15% of female cancer survivors were current cigarette smokers as of 2012[36]. The youngest age category (age 18−44) comprised the highest percentage (35.2%) of female smokers[36].

While studies have shown that timing of natural menopause is associated with ethnicity, this finding is still controversial[37,38]. In a study performed by Delellis Henderson et al. in 2008, ethnicity was found to be significantly associated with age of natural menopause in women from the Multiethnic Cohort Study after adjusting for other factors[37]. Compared to the reference population of non-Latina Whites, Japanese-American women were significantly less likely to develop early menopause (hazard ratio (HR) = 0.93, (95% CI = (0.90, 0.95)))[37]. Both US and non-US born Latina women were at a significant increased risk of developing early menopause compared to non-Latina Whites (HR = 1.10 (95% CI= (1.07, 1.14)), HR = 1.25 (95% CI = (1.21, 1.30)) respectively) and no significant difference was found between non-Latina Whites and African American women (HR = 0.99 (95% CI = (0.96, 1.02)))[37].

1.1.3 Risk Factors for Premature Menopause in Childhood Cancer Survivors

Extensive research has been undertaken on the various risk factors for the development of NSPM in CCSs following cancer treatment completion[39]. Treatment exposures, such as chemotherapy and radiation therapy (RT), have been identified as main risk factors for the development of late

effects in CCSs. Their toxic effects on the reproductive organs are known as *gonadotoxic effects*, and can contribute to ovarian dysfunction through multiple pathways and methods[40].

Chemotherapy

The risk of developing premature menopause is associated with the specific chemotherapy agent used and the cumulative dosage of exposure to agents[4,41]. Chemotherapy agents fall into numerous classes, including alkylating agents, anthracyclines and antimetabolites, all of which are categorized based on their structure and action[42]. They are appropriate for cancer treatment as the rapidly dividing cancer cells are more sensitive to DNA damage and do not have time for repairs[42].

Alkylating agents work by interacting with DNA and preventing cell division and growth, and are the class of chemotherapy agents that have the highest potential to cause gonadotoxic damage[6,27,40,42]. In contrast to other treatment exposures, the gonadotoxic effect of alkylating agents for women is specific to the ovaries, leaving the uterus unharmed[40]. Although the primary goal of alkylating agents is to cause damage to the cancer cells, these drugs can also do damage to the surrounding healthy tissues (to the ovaries in particular, as they are attracted to the maturing cells)[40,42]. Alkylating agents instigate follicle depletion in the ovaries, reducing the number of ovarian follicles available for maturation and reproduction, and increasing the potential for menopause[40]. Compared to the general population, patients exposed to alkylating agents as their only treatment had reduced ovarian reserve, with procarbazine (an alkylating agent) exposure associated with a significant decrease in reserve size and levels of important ovarian hormones[43].

Normalization of the cumulative doses of alkylating agents can be attained through the cyclophosphamide equivalent dose (CED). The CED standardizes the exposures of 10 common alkylating agents (cyclophosphamide included) to the units of cyclophosphamide to allow for quantification and comparison of exposures independent of the study cohort[44]. The risk of NSPM is significantly larger with an increased exposure to alkylating agents in total, corresponding to an increased value of CED[9,44]. In a 2006 study by Sklar et al. on premature menopause in the CCSS, it was determined that although the risk associated with NSPM was increased with any exposure to these agents, the risk was further increased with increasing dosage levels, indicative of a dose-response relationship[45]. For patients with a CED value greater than or equal to 4000 mg/m^2 and less than 8000 mg/m^2, the relative risk (RR) of developing NSPM was 2.74 (95% CI = (1.13, 6.61), $p = 0.025$) compared to individuals without any alkylating agent exposure[44]. This risk was further increased for individuals with a CED value greater than 8000 mg/m^2, with a RR of developing NSPM equal to 4.19 (95% CI = (2.18, 8.08), $p < 0.001$) times that of an individual without any exposure[44].

A specific alkylating agent included in the calculation of the CED is procarbazine, which is a drug predominantly used to treat Hodgkin lymphoma[46]. Of all the agents included in the CED calculation, procarbazine is found to have the most significant impact on NSPM development[9]. In the original cohort of the CCSS, 39.7% of survivors who were treated with a procarbazine dose ≥ 4000 mg/m^2 had developed NSPM by age 40[9]. Furthermore, univariate analysis showed that the CED variable is no longer significant when the contribution of procarbazine is removed[9]. In the 2018 study published by Levine et al., exposure to a procarbazine dose ≥ 4000 mg/m^2 led to an OR of 8.96 (95% CI = (5.02, 16.00), $p < 0.0001$) for developing NSPM compared to no

exposure[9]. However, treatment protocols have recently been modified in order to reduce the use of high quantities of harmful exposures[46]. This has led to the limited use of procarbazine as a treatment exposure except in the specific cases of Hodgkin lymphoma, and therefore, is not a strong risk factor in the majority of patients who remain unexposed[46].

Radiation Therapy

RT uses energy from electrically charged particles to invoke cell death in the exposed tissue, with the aim of destroying cancer cells and causing the least amount of harm to normal cells[47]. Oocytes are particularly vulnerable to the genomic damage caused by radiation exposure, with cell death causing significant decreases in the reserve of ovarian follicles[40]. Childhood exposure to radiation has been implicated in numerous outcomes related to ovarian dysfunction, as the gonadotoxic effects of radiation impact not only the ovaries but the uterus as well[40]. The likelihood of becoming pregnant in adulthood is significantly reduced after exposure to abdominal or pelvic radiation[48]. Pregnancy complications, such as the risk of spontaneous abortion, preterm birth, low birth weight, and stillbirth, are increased after exposure to radiation[40,49].

Radiation impacts the risk of premature menopause depending on the specific site it is administered to and dosage level[4,40,49]. Direct radiation to the ovaries (through abdominal, pelvic or total body radiation) at doses of radiation greater than 10 Grays (Gy) has been linked to a high risk of developing NSPM and ovarian dysfunction, although exposure to scatter radiation from other body areas can confer significant damage[40,45]. Nevertheless, individuals exposed to any dose of radiation experience an increased risk of developing NSPM compared to those without

any exposure[45]. The OR was 2.73 (95% CI = (1.33, 5.61), p = 0.0062) for the development of NSPM for individuals with ovarian radiation doses < 5 Gy compared to no radiation exposure, and increased to 8.02 (95% CI = (2.81, 22.85), p < 0.0001) for individuals with ovarian radiation doses ≥ 5 Gy[9].

Radiation to the Hypothalamic-Pituitary-Ovarian (HPO) axis has additionally been linked to a high risk of ovarian complications and dysfunction[4,48]. As stated previously, the hypothalamus and pituitary gland release hormones which are necessary for ovarian development and stimulation. Damage to the hypothalamus and pituitary through cranial radiation modifies the timing of the release of these hormones, consequently contributing to atypical ovarian development, and may result in lower pregnancy rates[4]. In particular, doses of radiation greater than 35 Gy have been identified as high risk for impacting fertility, compared to lower doses[48]. In contrast to AOF and NSPM, the effects of the damage caused to the pituitary and hypothalamus can be moderated through the routine administration of gonadotropic hormones allowing the patient to achieve normal ovarian function and providing the opportunity for future reproductive possibilities[50,51].

Additional Treatment Related Risk Factors

The combined use of both radiation and chemotherapy in an individual is a common treatment protocol, with 27.4% of patients in the original cohort of the CCSS having received exposure to alkylating agents and ovarian radiation[9,42]. Although exposure to alkylating agents or ovarian radiation during treatment individually are both classified as major risk factors for the development of NSPM, their combined use in a patient poses the greatest risk[39,45]. The

cumulative incidence of NSPM in CCSs who were exposed to both alkylating agents and abdominal/pelvic radiation approached 30% by age 40[45].

The risk of treatment in preparation for stem cell transplantation (SCT) and bone marrow transplantation (BMT) on the development of NSPM has recently been assessed in the original cohort of the CCSS[9]. 17 individuals received SCT, and exposure was found to have a significant increase in the odds of the development premature menopause (OR = 6.35, 95% CI = (1.19, 33.93), p = 0.0307)[9]. During the 1970s and early 1980s, preparation for SCT treatment was generally preceded by total body irradiation (TBI), where the body is flushed with high doses of radiation[46]. In more recent periods however, preparation with TBI was replaced by treatment with high doses of chemotherapy agents, particularly cyclophosphamide and busulfan[40].

An additional risk factor demonstrated to increase the risk of premature menopause in CCSs is an older age at initial cancer diagnosis and therefore an older age at treatment. In a study of cancer patient survivors treated with chemotherapy only, the risk of AOF increased significantly with an older age of cancer diagnosis[52]. As the ovarian follicle reserve decreases with increasing age, an older patient will have a lower number of ovarian follicles compared to a younger patient. Radiation and chemotherapy both accelerate the depletion of the number of ovarian follicles and cause damage to the ovaries, contributing to an earlier timing of menopause[4,7]. Therefore, in general, the older a female is when she is treated for cancer, the less follicle reserve she possesses and the more vulnerable she is to potential damage caused by treatment exposures[7,40]. Younger ovaries have been shown to be more resistant to the toxins administered through radiation[6,7]. For example, for a female at age 12, the mean sterilizing dose of radiation to

the ovaries is 18 Gy[4]. Conversely, a female at age 45 requires an ovarian radiation dose of only 9.5 Gy to produce the same effect[4].

1.2 Methodology Review

1.2.1 Measures of Association

It is crucial to ensure sufficient knowledge of the terminology used during risk prediction, particularly with respect to the measures estimated by the models and their subsequent implications. Of particular importance to epidemiological studies are the measures of prevalence and incidence, and their relationship to the absolute risk, relative risk and the odds ratio.

Prevalence and Incidence Measures

The prevalence and incidence of an event are common epidemiological measures estimated in a population. The population for which these measures are estimated is termed the *population at risk*[53]. Individuals are defined as being within the population at risk under the condition that if they were to develop the event during the pre-specified time period, they would be counted as cases in the calculation[53]. The *prevalence* is a static measure of event frequency in the population, as it represents the proportion of the population that observed an event at a specific time t_0[54]. Prevalence is calculated from the ratio of the number of cases present at t_0 (known as *prevalent cases*) and the size of the risk set of the population[54].

$$\text{Prevalence} = \frac{\text{Number of Cases at } t_0}{\text{Total Population at Risk}}$$

The prevalence can also be calculated over a specific time period using the *period prevalence*. The numerator includes those cases that developed within the specified time period, and the denominator is the average population within the specified time period.

$$\text{Period Prevalence} = \frac{\text{Number of Cases in a Specified Time Period}}{\text{Average Population at Risk}}$$

The *incidence* of an event is a population measure used to examine the change in event occurrence over time[54]. Incidence can be represented by the cumulative incidence, as well as the incidence rate. The *cumulative incidence* is the ratio of the number of new cases which developed during a specified time period and the number of individuals at risk at the beginning of the time period[54]. A condition posed by the cumulative incidence is that it must only be used in closed populations, where there is no entry or exit of participants during the study period and all individuals included in the analysis are at risk throughout the entire period[53,54]. This assumption is rarely realistic as often individuals are not under observation for the entire duration of the time period, and this contradiction may lead to bias in the estimated cumulative incidence measure[53].

$$\text{Cumulative Incidence} = \frac{\text{Number of New Cases Over the Time Period}}{\text{Total Population at Risk Initially}}$$

The *incidence rate* provides an estimate of the development of cases for each unit of person-time measured[53]. It may be used in situations which do not require participants to be observed for the entire study period as it incorporates the time at risk for each individual in the study. This can be measured in terms of the number of person-years (PY) at risk contributed to the study. The denominator in the incidence rate is composed of the sum of the total PY contributed by the all individuals within the study[53].

$$\text{Incidence Rate} = \frac{\text{Number of New Cases Over the Time Period}}{\text{Total PY at Risk}}$$

Absolute Risk, Relative Risk and the Odds Ratio

Absolute risk is the term used to define the incidence of an event and indicates the magnitude of the risk of an event within a population[53]. Essentially, the absolute risk represents the probability of an individual developing the event of interest[53]. The *relative risk* (RR) is used to compare the incidence (or risk) of an event in a group with an exposure present (the exposed group) to the risk of the event in a group without an exposure present (the unexposed group)[53]. It is computed as the ratio of the incidence in the exposed group to the incidence in the unexposed group (or the probability of the event occurring in the exposed group compared to the probability of the event occurring in the unexposed group)[53].

$$\text{RR} = \frac{\text{Probability of Event in the Exposed Group}}{\text{Probability of Event in the Unexposed Group}}$$

The RR is useful for comparing event risk between exposure statuses. A value for the RR greater than 1 indicates that the risk of event in the exposed group is greater than that of the unexposed group; similarly, a RR less than 1 indicates that the risk of the event in the exposed group is less than that of the unexposed group[53]. As incidence is a dynamic measure, the RR can be estimated from studies where participants are observed over a length of time[53]. Therefore, there are limitations on when the RR is an appropriate measure.

For situations where the participants are not followed over time, the RR cannot be calculated directly[53]. In case-control studies, the *odds ratio* (OR) compares the odds of exposure in cases (individuals with the event) to the odds of exposure in controls (individuals without the event), where the odds of an event are calculated as the probability of the event occurring (P) divided by the probability of the event not occurring (1-P)[53].

$$\mathrm{OR}_{\text{Case Control}} = \frac{\text{Odds of Exposure in Cases}}{\text{Odds of Exposure in Controls}}$$

For cohort studies and other studies where the subjects are measured over time, the incidence is an appropriate measure to calculate. In these studies, the odds ratio can be calculated as the odds of becoming a case based on exposure.

$$\mathrm{OR}_{\text{Cohort}} = \frac{\text{Odds that an Individual with the Event is Exposed}}{\text{Odds that an Individual with no Event is Exposed}}$$

Similar to the interpretation of the RR, an OR greater than 1 indicates a positive relationship between the exposure and the event, and less than 1 indicates a negative relationship with the event[53]. The RR can be estimated from the OR assuming that the cases and controls selected are an adequate representation of the underlying population, and the prevalence of the event is small (the rare disease assumption)[53]. This assumption states that the OR is a good estimate of the RR if the prevalence of the event in the population is less than 10%[53,55].

1.2.2 Survival Analysis

Individuals within a prospective study are followed for a length of time to potentially observe the development of an event of interest. Depending on the study objective, this event of interest could be an injury, the diagnosis or relapse of an illness, or death from a certain cause. Investigators recruit participants to a study if they meet predetermined inclusion criteria. For example, the original cohort in the Childhood Cancer Survivor Study was composed of patients diagnosed at participating institutions with eligible cancers before age 21 between January 1, 1970 and December 31, 1986 who had survived at least 5 years[56].

Participants who meet the inclusion criteria and become enrolled are regarded as the *risk set* of the study. Being in the risk set implies that the individual is at risk for the development of the event of interest. When the individual no longer meets the criteria for being at risk for the event, they are removed from the risk set. The study design may allow individuals to enter the study at various calendar times, and therefore allow them to be under observation for different lengths of time. The majority of studies have set periods within which patients are initially recruited, are subsequently observed for a length of time, and monitored for the development of the event.

Survival Data

Survival analysis is primarily interested in the *time-to-event* for study participants, or the length of time from study entry until the event occurs. The random variable T_i represents the complete follow-up time of patient i within a study. Time is characterized as a continuous variable with T_i defined on $[0, \infty)$[57]. For those who observed the event of interest within the study period, the time from the start of observation to the event of interest for the ith individual is recorded and

denoted by the random variable T_i^*. However, not all individuals observe the event of interest. The study may terminate before the individual has experienced the event, or the individual may be lost to follow-up (LTFU) within the study period.

An individual is LTFU if they were at one time participating in the study, but are no longer able to be contacted or followed up with[57]. Individuals who do not have the event of interest within the study period or are LTFU are said to be right censored[57]. C_i^* denotes the time from the start of observation of patient i until the observation of the individual ceases (the censoring time). T_i, the complete follow-up time of patient i, can therefore be expressed as the minimum of T_i^* and C_i^* as only one time value is typically observed ($T_i = \min(T_i^*, C_i^*)$). δ_i is an indicator variable which takes the value 1 if the individual observes the event ($T_i^* < C_i^*$) and 0 if the individual is censored ($T_i^* > C_i^*$)[57]. Therefore, survival data includes a pair of outcome data for each individual, consisting of the follow-up time and the status indicator (T_i, δ_i).

Survival Functions

Three functions which can be estimated from survival data include the probability density function, the survivor function, and the hazard function. These functions are based on the random variable T_i (the complete follow-up time). The probability density function is defined as:

$$f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t)}{\Delta t}$$

and represents the probability of the event of interest occurring at time t[57]. The survivor function is given by:

$$S(t) = P(T \geq t) = \int_{t}^{\infty} f(x)\, dx$$

where the function $f(t)$ represents the probability density function[57]. In the context of survival analysis, the word "surviving" can be interpreted as time without the event. Therefore, the survivor function measures the probability of being event free up to time t[57]. The relationship between the survivor function and the probability density function can also be expressed as $f(t) = -S'(t)$[57]. The survivor functions is a decreasing continuous function with $S(0) = 1$ and $S(\infty) = \lim_{t \to \infty} S(t) = 0$, implying that the probability of survival at baseline is one hundred percent, and as the time of study approaches infinity, the probability of survival tends to zero[57].

The hazard function (also known as the hazard rate), measures the instantaneous incidence rate of the event of interest[57]. It represents the probability that the event will occur at time t given that the individual has been event free up to time t[57]. The equation for the hazard function is:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

where $\lambda(t)$ is the hazard rate at time t. The hazard rate can also be expressed as the ratio between the previous two functions:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

Cox Proportional Hazards Regression

The Cox proportional hazards (PH) model, proposed by David R. Cox in his 1972 paper, is the most common method used to model survival data[58]. It assesses the effect of predictors and covariates on the hazard rate, but leaves the baseline hazard unspecified[59]. The general form of this model is:

$$\lambda_i(t) = \lambda_0(t)\exp(\Sigma\beta_p X_{pi})$$

$\lambda_0(t)$ represents the baseline hazard function (a nonnegative unspecified function of time) and β_p is a column vector of coefficients. The X_{pi}'s represent the p explanatory variables. The Cox PH model falls into the category of semiparametric models, which implies that some of the parameters are estimated (β_p) while others are left unknown ($\lambda_0(t)$) [59].

This semiparametric nature can be achieved due to the assumption that while holding all else constant, the hazard ratio (HR) remains constant over time between two levels of a covariate (the proportional hazards assumption)[59]. The HR is the ratio of hazard rates for two distinct individuals ($\lambda_1(t)$ and $\lambda_2(t)$) who take on different values for a specific covariate X (X_1 and X_2 respectively)[59]:

$$\text{HR} = \frac{\lambda_1(t)}{\lambda_2(t)}$$

$$\text{HR} = \frac{\lambda_0(t)\exp(X_1\beta)}{\lambda_0(t)\exp(X_2\beta)}$$

$$\text{HR} = \exp(\beta(X_1 - X_2))$$

Based on the model, the baseline hazard function, $\lambda_0(t)$, is the same for both individuals and cancels out. The $\exp(\beta)$ produced by the Cox PH model represents the HR for the specific covariate X corresponding to β and is interpreted as the change in hazard from one level of covariate to another[59].

The underlying assumption of proportional hazards is crucial to use the Cox PH model with the desired interpretation of the hazard ratio. This assumption can be visually tested by plotting $-\log(\log(S(t)))$ (where $S(t)$ is the survivor function) against time for different covariate levels. If the proportional hazards assumption holds, the lines will be parallel and differ by β[57]. The assumption can also be tested by assessing the Schoenfeld residuals, whereby a large resulting p-value provides no evidence against the proportional hazards assumption[57]. A method for correcting a violation of the PH assumption is to fit time dependent variables into the model, which allows for the effect of a covariate to change over time[57]. The proportionality assumption can also fail when variables are omitted from the analysis in error[59]. However, if the proportional hazards assumption is not met, it is recommended to examine alternative models.

An additional assumption accompanying the proportional hazards assumption is that the continuous or ordinal independent variables (or a function of the continuous or ordinal independent variables) have a linear effect on the log hazard, as demonstrated by the log of the hazard equation[57]:

$$\log(\lambda_i(t)) = \log(\lambda_0(t)) + \Sigma\beta_p X_{pi}$$

Martingale residuals can be used to assess the functional form of the continuous independent variables in the model, and to confirm if a linear relationship between independent variables and the log hazard is appropriate, particularly in the presence of censoring events[59,60]. Essentially, martingale residuals compare the observed number of events for the ith individual to the expected number of events based on the individual specific hazard equation[59]:

$$\widehat{M_i} = \delta_i - \widehat{\Lambda_0}(t_i)\exp(\Sigma\hat{\beta}_p X_{pi})$$

δ_i is the censoring indicator, $\widehat{\Lambda_0}(t_i)$ is the cumulative baseline hazard function for the ith individual at their latest follow-up time, and $\Sigma\hat{\beta}_p X_{pi}$ are the estimated coefficients from the model, and the observed covariate values for the ith individual[59,60]. Each martingale residual has an expected value of 0 under the correct model, and the sum of all observed martingale residuals is also 0[59]. Once these residuals are calculated, they are plotted against the variable of interest and smoothed using a smoothing algorithm (such as a lowess smoother) to ease in the identification of a pattern in the data[59,61]. If the resulting curve is linear, then including the variable in the model as a linear variable is appropriate[59,61]. In contrast, for curves which do not appear linear when plotted, a transformation may have to be applied to the variable in order to correctly include it in the model[59].

Competing Risk Models

Although a study may be undertaken to explicitly observe one event of interest in its population, this objective may be precluded by the development of other events which prevent the individual from observing the event of interest[62]. These events are called *competing risk events*[62]. For

example, in a study measuring the time to recurrence of primary cancer in a group of cancer survivors, the study population is susceptible to many external events which would inhibit a relapse, including cancer in a new location or death. An individual who experiences a competing risk event is unlike an individual who is censored due to LTFU. Individuals who are censored are still assumed to be at risk for the event of interest, just unable to be observed. It is assumed that once a competing risk event has occurred, that individual is no longer at risk for the event of interest in the future[62].

All subjects recruited and observed in a study are under the assumption that they have the possibility to experience all the events. Usually, only the time to the first event is recorded, regardless of the events' importance to the study objective. That being said, there is still information provided by the participants who observe competing events which should be included in the model. With the introduction of competing events, δ_i, the status indicator of survival data pairs, is able to take more values. Previously, a value of 0 indicated that the individual was censored, and a value of 1 indicated that the individual observed the event of interest. With competing risk events introduced into the model, δ_i can take on K values in additi on to 0 and 1 ($K = 2, \dots, k$ where k represents the number of competing events in the situation).

The cause specific hazard for event k ($\mathrm{CSH_k}$), $\lambda_k(t)$, is defined as the instantaneous rate of failure at time t from cause k, given that no failure from any cause has occurred previously[63].

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t, \text{failure from cause } k \mid T \geq t)}{\Delta t}$$

Returning to the previous cancer recurrence example, this quantity would measure the instantaneous probability of primary cancer relapse (the event of interest) at time t given that up until time t the patient was completely event free. The CSH for event k can be modelled using the Cox PH model by treating the remaining events as censoring events provided all events are assumed to be independent[63]. In general, the assumption of independence between events is impractical and therefore often competing risk events cannot simply be treated as censoring.

The cumulative incidence function for event k ($\mathrm{CIF_k}$) at time t from cause k represents the probability of failure from cause k up to t[62].

$$\mathrm{CIF_k} = \Pr(T \leq t, \delta_i = k)$$

A semiparametric model was developed by Fine and Gray to model the CIF by involving the use of the subdistribution hazard function, which can be written as[64]:

$$\lambda_k^*(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t, \text{failure from cause } k \mid T \geq t \cup (T \leq t \cap \text{not cause } k))}{\Delta t}$$

The value of the subdistribution hazard represents the instantaneous probability of failure from cause k at time t given that either no failure has occurred before time t if a failure did occur, it was a failure from another cause[63]. Essentially, those individuals who experienced competing events remain in the risk set instead of being removed, even though they are technically no longer at risk.

The Fine and Gray model (also known as Fine-Gray Regression (FGR)) has a similar form to the Cox PH model, whereby it is semiparametric and leaves the baseline subdistribution hazard unspecified[63].

$$\lambda_k^*(t) = \lambda_{k0}^*(t)\exp(\Sigma\beta_p X_{pi})$$

The resulting exponentiated coefficients, $\exp(\beta)$, are interpreted as the subdistribution hazard ratios (SHR) and compare one level of a covariate to another while holding all else constant[63]. The direction of the SHR will specify the direction of the effect of the specific covariate on the $\mathrm{CIF_k}$, however does not give the exact magnitude of the effect of the covariate[62]. The equation for obtaining the predicted probability of an event at a specific time from the FGR model involves the estimated coefficients from the FGR model as well as the baseline cumulative incidence function ($\mathrm{CIF_0}$)[62]:

$$1 - \mathrm{CIF_k}(t) = \left(1 - \mathrm{CIF_{k,0}}(t)\right)^{\exp(X\beta)}$$

Depending on the study objective, the ideal method to implement for modelling competing risks can differ[62]. If the question is focused on answering etiologic questions about the exposure and outcome relationship, then the CSH model is fine to use as the regression is solely assessing the influence of covariates on a specific event type[62]. When the focus is on prediction, and determining the absolute incidence of an event occurring, then using FGR is recommended as it models the effect of covariates on the cumulative incidence while taking into account competing risk events[62].

Random Survival Forests

Tree-based methods aim to divide the covariate space into k non-overlapping regions using recursive binary partitioning to partition the outcome[65,66]. Although traditionally applied to classification and regression problems, tree based methods have been expanded to solve prediction problems for survival data, and the steps for the latter follow roughly the same format as the former. Random survival forests (RSF) aim to divide the covariate space into groups with similar time-to-event outcomes[67]. Benefits of the survival tree methods include their flexibility and ability to handle high dimensional covariates; however their potential to favour variables with multiple split points can result in bias[67].

A decision tree is developed based on the set of observations, a node splitting rule, and a stopping rule[65]. The node splitting rule determines how the observations are partitioned (or "divided") at each node. The rule is chosen such that the split results in the greatest reduction in node impurity, or so the observations within each partition are more homogeneous[66]. A stopping rule is assigned to determine how large the tree will grow; ideally such that each resulting partition contains at least some predetermined minimum number of observations[67]. The *terminal nodes* (nodes which do not divide further) are referred to as leaves. All of the observations within a terminal node will be assigned the same predicted value for the response (in the case of survival data, this is the same estimate of time-to-event).

Bootstrap sampling has been shown to increase the stability of predictors and reduce variance in model building for a wide variety of modelling approaches[65]. In general, bootstrap sampling is performed by repeatedly sampling N observations from the original data set *with* replacement

(once the observations are sampled, they are put back into the sample) in order to generate the set of bootstrap data sets[68]. B bootstrapped samples of size N are generated from the original dataset[68]. A survival tree is then built on the bth dataset to obtain prediction estimates[68]. Subsequently, an average is taken of the B prediction estimates, a process known as *bootstrap aggregation* or "bagging"[65,68].

The method of random forest is similar to bagging, although introduces a step which decorrelates the developed trees from one another[68,69]. Similarly, B bootstrap samples are drawn from the original data set and a tree is grown for each bootstrap sample. However, at each node, only $\sqrt{p}$ variables from the available set of p covariates are selected for inclusion as split candidates[68]. This allows the generated bootstrapped trees to be distinct, as not all variables are evaluated at every split[68,69]. This is particularly beneficial should a few variables be particularly strong predictors and therefore prominent in the majority of splits[68]. The process of only choosing a subset of $\sqrt{p}$ variables for consideration at each split point allows other variables to be considered when the strong predictors are excluded, which ultimately results in reduced correlation between the ensuing trees[68,69]. Results from random forest have been shown to have reduced bias and variance, and produce a smaller prediction error than other procedures, including bagging alone[69].

Unlike regression and classification trees which have an established measure of node impurity to use for splitting the nodes, there is no such measure universally acknowledged for survival data, providing a challenge for the development of survival forests[66,67]. The most frequently used splitting rule for survival data with censoring is the logrank test statistic, which compares a

weighted version of the Nelson-Aalen cumulative hazard estimator between daughter nodes[70]. The Nelson-Aalen cumulative hazard estimator ($\widehat{H}(t)$) is a non-parametric estimate of the hazard function of an event at time t, which is defined as[57]:

$$\widehat{H}(t) = \int_0^t \frac{\sum_i \mathrm{I}(\mathrm{T} = \mathrm{s}, \delta_i = 1)}{\sum_i \mathrm{I}(\mathrm{T} \geq \mathrm{s})} \mathrm{ds}$$

The logrank test statistic is subsequently computed by weighting the difference between the Nelson-Aalen cumulative hazard estimators obtained from each daughter node by the number of events observed in each node and dividing by the variance of the weighted difference measure[70]. The best split is chosen as the one that maximizes the logrank test statistic, or results in the greatest difference between cumulative hazards estimators[70]. For trees with survival data incorporating competing risks, the superior splits are determined by maximizing the weighted difference between the cumulative incidences of events between daughter nodes (Gray's test statistic) which assesses the direct effect of covariates on the cumulative incidence of the event of interest, and is appropriate for studies interested in predicting the probabilities of events[71].

1.2.3 Time-specific Logistic Regression

Logistic regression is a regression technique for binary outcome variables, which indicate yes or no if the event of interest occurred[68]. The logistic function modelled by logistic regression is:

$$\Pr(Y \mid X_i) = \frac{\exp(\beta_0 + \sum \beta_p X_{pi})}{1 + \exp(\beta_0 + \sum \beta_p X_{pi})}$$

where Y is the binary variable representing the occurrence of the event, X_{pi} is a vector of covariate values for the ith individual, and β_p are the coefficients corresponding to the covariates[68]. The logistic regression equation can be presented in many forms; however the most common is the log odds equation[68]:

$$\text{logit}\big(\Pr(Y|X_i)\big) = \log\left(\frac{\Pr(Y|X_i)}{1-\Pr(Y|X_i)}\right) = \beta_0 + \Sigma\beta_p X_{pi}$$

The log odds form represents one of the assumptions of logistic regression, which is that there exists a linear relationship between the covariates and the log odds of the event. Additional assumptions for using logistic regression include ensuring that covariates are not highly correlated and that the observations are independent (no repeated measures).

The measure of association estimated through logistic regression is the OR, which is given by $\exp(\beta_p)$. The odds of the occurrence of the event will be increased by $\exp(\beta_p)$ for a one unit increase in the corresponding covariate value (X_p), while holding all other covariates constant[68]. Although not directly estimated, the RR can be estimated from the OR provided by logistic regression assuming that the rare disease assumption is valid as described previously. Obtaining predicted probabilities of event occurrence using logistic regression involves rearranging the log odds equation back to the original logistic form and plugging in the estimated coefficients and individual covariate values.

Logistic regression is unlike the aforementioned survival models, as it does not deal with time-to-event data. Logistic regression can be used to provide an estimate of the odds of the event at a

specific time of interest and the resulting predicted probability can subsequently be computed. Time-specific logistic regression (TLR) estimates can be computed for each of the pre-specified time periods required. To account for censoring of event time which can occur due to LTFU or an unobserved event during the study period, the method of inverse probability-of-censoring weights can be applied (described in detail in Chapters 3 and 4).

1.2.4 Model Performance and Accuracy Assessments

Following the development of prediction models, it is important to evaluate model performance to aid in determining the superior model.

The Receiver Operating Characteristic Curve and the Area under the Curve

The receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) are methods used to assess the discrimination of a classifier (either a model or a diagnostic test)[72]. The discrimination refers to the ability of the classifier to correctly classify the outcome as positive (event observed) or negative (no event observed)[73]. The ROC curve plots the *sensitivity* against 1 - the *specificity* of a diagnostic test for different threshold values[72]. The sensitivity of a diagnostic test measures how well the test is able to correctly predict a positive outcome when a positive outcome is present (a "true" positive (TP))[72]. It uses the ratio of TP to the total number of observations with a positive outcome (TP + false negatives (FN)) which gives the true positive rate[72]. In contrast, the specificity measures how well the diagnostic test is able to correctly identify a negative outcome when a negative outcome is present (the ratio of true negatives (TN) to all observations classified as negative (TN + false positive (FP))[72].

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Visually, the closer the plot is to a straight line through the origin (45-degree angle), the less ability the classifier has to discriminate between positive and negative outcomes[72]. A straight line through the origin indicates that the test does no better than selecting each outcome by chance[72]. Optimally, the curve will be high in the upper left corner[72].

The AUC value summarizes the discrimination ability of a classifier based on the ROC curve. It is the area under the ROC curve, and represents the probability that a randomly chosen observation with a positive outcome will be ranked higher than a randomly chosen observation with a negative outcome[72]. Values for the AUC range from 0.5 to 1.0, with 0.5 indicating no discrimination, and corresponding to an ROC curve with a straight line through the origin[72]. Values greater than 0.5 indicate the model has some ability to distinguish between positive and negative outcomes, and ideally the value will be close to 1.0[72].

The Precision Recall Curve and the Area under the Curve

Although discrimination has consistently been used in clinical studies to evaluate diagnostic tests, evidence has shown that the AUC measure is inappropriate for evaluating the accuracy of *prospective* risk prediction models, as the AUC is comprised of two *retrospective* measures (sensitivity and specificity)[74]. Therefore, the positive predictive value (PPV) has been proposed

as a more suitable evaluator of prospective risk prediction models as it is calculated using data from a prospective cohort[74]. The PPV of a test corresponds to the proportion of individuals classified with a positive outcome who actually have a positive outcome[72,75].

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

The precision-recall curve (PR) was developed to summarize the PPV of a test and used to assess its' accuracy[74,75]. The PR curve plots the PPV (also known as *precision*) against the sensitivity (also known as *recall*) of a diagnostic test for various threshold values[75]. A non-informative PR curve will be a horizontal line intersecting the y-axis at the value of the event rate in the population[74].

The PR curve has been shown to provide a more honest estimate of model performance when the outcome of interest is of low prevalence in the target population[75]. It does not take into account the number of individuals correctly identified with a negative outcome (the TNs), which is high in a population with a low event rate, and would overinflate the estimate of the performance of a classifier[75,76]. The PR curve can be summarized by the area under the precision recall curve (AUCPR)[75,76]. The AUCPR value is denoted as the average positive predictive value (AP) with a range between the prevalence of the event in the population and 1.0, with 1.0 occurring when positive outcomes are *always* assigned a higher value than negative outcomes[74,75]. Values greater than the prevalence of the outcome in the population indicate that the test provides some discriminatory ability. To evaluate models for incidence over time, the time-dependent AP can be used to assess the model predictive accuracy.

Calibration Curves

Calibration is a measure of the agreement between observed outcomes and the predicted probabilities from a model, and indicates the reliability of the resulting predictions[73,77]. A model is said to be well calibrated if for a subpopulation assigned a predicted risk of p, the observed proportion of individuals with the event is close to p[77]. For example, assume a group of 100 individuals were each given a predicted probability of 0.20 for developing NSPM. The model would be deemed reliable if on average 20% (or 20 out of the 100 individuals) developed the event. Frequently, the results will be presented through a calibration curve, which gives a graphical assessment of the calibration[73]. A calibration curve representing perfect agreement between predicted probability and observed probability will be a 45-degree line[73].

1.3 References

1. Ries LAG, Eisner MP, Kosary CL, et al. Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975-1995, National Cancer Institute, SEER Program. *National Institutes of Health Publications.* 1999:No. 99-4649.

2. Robison LL, Hudson MM. Survivors of childhood and adolescent cancer: life-long risks and responsibilities. *Nature Reviews Cancer.* 2014;14(1):61-70.

3. Oeffinger KC, Mertens AC, Sklar CA, et al. Chronic Health Conditions in Adult Survivors of Childhood Cancer. *The New England Journal of Medicine.* 2006;355(15):1572-1582.

4. Gnaneswaran S, Deans R, Cohn RJ. Reproductive Late Effects in Female Survivors of Childhood Cancer. *Obstetrics and Gynecology International.* 2012;2012:1-7.

5. Bhakta N, et al. The cumulative burden of surviving childhood cancer: an initial report from the St Jude Lifetime Cohort Study (SJLIFE). *The Lancet.* 2017;390(10112):2569-2582.

6. Chemaitilly W, Mertens AC, Mitby P, et al. Acute Ovarian Failure in the Childhood Cancer Survivor Study. *The Journal of Endocrinology and Metabolism.* 2006;91(5):1723-1728.

7. Sklar CA. Maintenance of Ovarian Function and Risk of Premature Menopause Related to Cancer Treatment. *Journal of National Cancer Institute Monographs.* 2005;34:25-27.

8. Torrealday S, Pal L. Premature Menopause. *Endocrinology and Metabolism Clinics of North America.* 2015;44:543-557.

9. Levine JM, et al. Nonsurgical premature menopause and reproductive implications in survivors of childhood cancer: A report from the Childhood Cancer Survivor Study. *Cancer.* 2018;124(5):1044-1052.

10. Burger HG. Physiology and endocrinology of the menopause. *Medicine.* 2006;34(1):27-30.

11. Burger HG, Hale GE, Robertson DM, Dennerstein L. A review of hormonal changes during the menopausal transition: focus on findings from the Melbourne Women's Midlife Health Project. *Human Reproduction Update.* 2007;13(6):559-565.

12. Faubion SS, Kuhle CL, Shuster LT, Rocca WA. Long-term health consequences of premature or early menopause and considerations for management. *Climacteric.* 2015;18(4): 483-491.

13. Barbieri RL. The Endocrinology of the Menstrual Cycle. In: Rosenwaks Z, Wassarman PM, eds. *Human Fertility: Methods and Protocols.* 1st ed. Humana Press; 2014:145-169.

14. Alberts B, Johnson A, Lewis J, et al. Eggs. In: *Molecular Biology of the Cell.* 4th ed. New York: Garland Science; 2002.

15. Jamnongjit M, Hammes SR. Oocyte Maturation: The Coming of Age of a Germ Cell. *Seminars in Reproductive Medicine.* 2005;23(3):234–241.

16. Johnston RJ, Wallace WH. Normal ovarian function and assessment of ovarian reserve in the survivor of childhood cancer. *Pediatric Blood and Cancer.* 2009;53(2):296-302.

17. Shuster LT, Rhodes DJ, Gostout BS, Grossardt BR, Rocca WA. Premature menopause or early menopause: Long-term health consequences. *Maturitas.* 2010;65(2):161.

18. Muka T, Oliver-Williams C, Kunuscor S, et al. Association of Age at Onset of Menopause and Time Since Onset of Menopause with Cardiovascular Outcomes, Intermediate Vascular Traits, and All-Cause Mortality: A Systematic Review and Meta-Analysis. *The Journal of the American Medical Association Cardiology.* 2016;1(7):769-776.

19. Singer D, Mann E, Hunter MS, Pitkin J, Panay N. The silent grief: psychosocial aspects of premature ovarian failure. *Climacteric.* 2011;14(4):428-237.

20. Cousineau TM, Domar AD. Psychological impact of infertility. *Best Practice & Research Clinical Obstetrics and Gynaecology.* 2007;21(2):293-308.

21. Benetti-Pinto CL, de Almeida DMB, Makuch MY. Quality of life of women with premature ovarian failure. *Gynecological Endocrinology.* 2011;27(9):645-649.

22. Massarotti C, Scaruffi P, Lambertini M, Remorgida V, Del Mastro L, Anserini P. State of the art on oocyte cryopreservation in female cancer patients: A critical review of the literature. *Cancer Treatment Reviews.* 2017;57:50-57.

23. Anderson RA, Weddell A, Spoudeas HA, et al. Do doctors discuss fertility issues before they treat young patients with cancer? *Human Reproduction.* 2008;23(10):2246-2251.

24. Domingo J, Garcia-Velasco JA. Oocyte cryopreservation for fertility preservation in women with cancer. *Reproductive Endocrinology.* 2016;23:1-5.

25. Yu J, Huang J, Rosenwaks Z. Assisted Reproductive Techniques. In: Rosenwaks Z, Wassarman PM, eds. *Human Fertility: Methods and Protocols.* 1st ed. Humana Press; 2014: 171-231.

26. Paramanantham J, Talmor AJ, Osianlis T, Weston GC. Cryopreserved Oocytes: Update on Clinical Applications and Success Rates. *Obstetrical and Gynecological Survey.* 2015;70(2): 97-114.

27. Levine JM, Canada A, Stern CJ. Fertility Preservation in Adolescents and Young Adults with Cancer. *Journal of Clinical Oncology.* 2010;28(32):4831-4841.

28. Oktay K, Okem O. Fertility preservation medicine: A new field in the care of young cancer survivors. *Pediatric Blood and Cancer.* 2009;53:267-273.

29. Ladanyi C, Mor A, Christianson MS, Dhillon N, Segars JH. Recent advances in the field of ovarian tissue cryopreservation and opportunities for research. *Journal of Assisted Reproduction and Genetics.* 2017;34(6):709-722.

30. Abir R, Ben-Aharon I, Garor R, et al. Cryopreservation of *in vitro* matured oocytes in addition to ovarian tissue freezing for fertility preservation in paediatric female cancer patients before and after cancer therapy. *Human Reproduction.* 2016;31(4):750-762.

31. Resetkova N, Hayashi M, Kolp LA, Christianson MS. Fertility Preservation for Prepubertal Girls: Update and Current Challenges. *Current Obstetrics and Gynecology Reports.* 2013; 2(4):218-225.

32. Sun L, Tan L, Yang F, et al. Meta-analysis suggests that smoking is associated with an increased risk of early natural menopause. *Menopause: The Journal of The North American Menopause Society.* 2012;19(2):126-132.

33. Hyland A, Piazza K, Hovey KM, et al. Associations between lifetime tobacco exposure with infertility and age at natural menopause: The Women's Health Initiative Observational Study. *Tobacco Control.* 2015;0:1-9.

34. Emmons KM, Butterfield RM, Park ER, et al. Smoking Among Participants in the Childhood Cancer Survivors Cohort: The Partnership for Health Study. *Journal of Clinical Oncology.* 2003;21(1):189-196.

35. Gibson TM, Liu W, Armstrong GT, et al. Longitudinal smoking patterns in survivors of childhood cancer: An update from the Childhood Cancer Survivor Study. *Cancer.* 2015; 121(22):4035-4043.

36. National Cancer Institute: Cancer Survivors and Smoking. Cancer Trends Progress Report. Accessed June 2017, from: https://www.progressreport.cancer.gov/after/smoking.

37. DeLellis Henderson K, Bernstein L, Henderson B, Kolonel L, Pike MC. Predictors of the Timing of Natural Menopause in the Multiethnic Cohort Study. *American Journal of Epidemiology.* 2008;167(11):1287-1294.

38. Luborsky JL, Meyer P, Sowers MF, Gold EB, Santoro N. Premature menopause in a multi-ethnic population study of the menopause transition. *Human Reproduction.* 2002;18(1):199-206.

39. Green DM, Sklar CA, Boice JD, et al. Ovarian Failure and Reproductive Outcomes After Childhood Cancer Treatment: Results from the Childhood Cancer Survivor Study. *Journal of Clinical Oncology.* 2009;27(14):2374-2381.

40. Oktem O, Kim SS, Selek U, Schatmann G, Urman B. Ovarian and Uterine Functions in Female Survivors of Childhood Cancers. *The Oncologist.* 2018;23(214-224).

41. Overbeek A, et al. Chemotherapy-related late adverse effects on ovarian function in female survivors of childhood and young adult cancer: A systematic review. *Cancer Treatment Reviews.* 2017;53:10-24.

42. Grant CH, Gourley C. Chapter 2: Relevant Cancer Diagnoses, Commonly Used Chemotherapy Agents and Their Biochemical Mechanisms of Action. In: Anderson RA, Spears N, eds. *Cancer Treatment and the Ovary: Clinical and Laboratory Analysis of Ovarian Toxicity.* 1st ed. Elsevier Science; 2015.

43. Thomas-Teinturier C, et al. Ovarian reserve after treatment with alkylating agents during childhood. *Human Reproduction.* 2014;30(6):1437-1446.

44. 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. *Pediatric Blood and Cancer.* 2014;61:53-67.

45. Sklar CA, Mertens AC, Mitby P, et al. Premature Menopause in Survivors of Childhood Cancer: A Report From the Childhood Cancer Survivor Study. *Journal of the National Cancer Institute.* 2006;98(13):890-896.

46. Sklar CA, Mostoufi-Moab S. Discussion regarding analysis of the Childhood Cancer Survivor Study dataset (Personal Communication). March 13th, 2018.

47. Baskar R, Lee KA, Yeo R, Yeoh K. Cancer and Radiation Therapy: Current Advances and Future Directions. *International Journal of Medical Sciences.* 2012;9(3):193-199.

48. Green DM, Kawashima T, Stovall M, et al. Fertility of Female Survivors of Childhood Cancer: A Report From the Childhood Cancer Survivor Study. *Journal of Clinical Oncology.* 2009; 27(16):2677-2685.

49. Gao W, Liang J, Yan Q. Exposure to radiation therapy is associated with female reproductive health among childhood cancer survivors: a meta-analysis study. *Journal of Assisted Reproduction and Genetics.* 2015;32:1179-1186.

50. Bath LE, Wallace WHB, Critchley HOD. Late effects of the treatment of childhood cancer on the female reproductive system and the potential for fertility preservation. *British Journal of Obstetrics and Gynaecology.* 2002;109(2):107-114.

51. Vern-Gross TZ, Bradley JA, Rotondo RL, Indelicato DJ. Fertility in childhood cancer survivors following cranial irradiation for primary central nervous system and skull base tumors. *Radiotherapy and Oncology.* 2015;117:195-205.

52. Letourneau JM, Ebbel EE, Katz PP, et al. Acute ovarian failure underestimates age-specific reproductive impairment for young women undergoing chemotherapy for cancer. *Cancer.* 2012;118(7):1933-1939.

53. Gordis L. *Epidemiology.* 5th ed. Elsevier Canada; 2013.

54. Koepsell TD, Weiss NS. *Epidemiologic Methods: Studying the Occurrence of Illness.* Oxford University Press; 2003.

55. Austin PC. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. *Journal of Clinical Epidemiology.* 2010;63(1):2-6.

56. Robison LL, Mertens AC, Boice JD, et al. Study design and cohort characteristics of the Childhood Cancer Survivor Study: a multi-institutional collaborative project. *Medical and Pediatric Oncology.* 2002;38:229-239.

57. Lawless J. *Statistical Models and Methods for Lifetime Data.* 2nd ed. Wiley; 2003.

58. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Society Interface. Series B (Methodological).* 1972;34(2):187-220.

59. Therneau T, Grambsch P, eds. *Modeling Survival Data: Extending the Cox Model.* 1st ed. Springer; 2000.

60. Hosmer DW, Lemeshow S, May S. Chapter 6: Assessment of Model Accuracy. In: *Applied Survival Analysis: Regression Modeling of Time-to-Event Data.* 2nd ed. New Jersey: John Wiley & Sons, Inc.; 2008:169.

61. Hosmer DW, Lemeshow S, May S. Chapter 5: Model Development. In: *Applied Survival Analysis: Regression Modeling of Time-to-Event Data.* 2nd ed. New Jersey: John Wiley & Sons, Inc.; 2008:132.

62. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risks. *Statistics in Medicine.* 2017;36:4391-4400.

63. Dignam JJ, Zhang Q, Kocherginsky MN. The Use and Interpretation of Competing Risks Regression Models. *Clinical Cancer Research.* 2012;18(8):2301-2308.

64. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association.* 1999;94(446):496-509.

65. Hothorn T, Lausen B, Benner A, et al. Bagging Survival Trees. *Statistics in Medicine.* 2004; 23(1):77-91.

66. Zhou Y, McArdle J. Rationale and Applications of Survival Tree and Survival Ensemble Methods. *Psychometrika.* 2015;80(3):811-833.

67. Nasejje JB, Mwambi H, Dheda K, Lesosky M. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Medical Research Methodology.* 2017; 17(115).

68. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: with Applications in R.* Springer; 2013.

69. Ishwaran H, Kogalur UB. Consistency of Random Survival Forests. *Statistics and Probability Letters.* 2010;80:1056-1064.

70. LeBlanc M, Crowley J. Survival Trees by Goodness of Split. *Journal of the American Statistical Association.* 1993;88(422):456-467.

71. Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. *Biostatistics.* 2014;15(4):757-773.

72. Zhou X, Obuchowski NA, McClish DK. Chapter 2: Measures of Diagnostic Accuracy. In: *Statistical Methods in Diagnostic Medicine.* 2nd ed. John Wiley & Sons, Inc.; 2011:13-56.

73. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology.* 2010;21(1):128-138.

74. Yuan Y, Zhou QM, Li B, Cai H, Chow EJ, Armstrong GT. A threshold-free summary index of prediction accuracy for censored time to event data. *Statistics in Medicine*. 2018;37(10): 1671-1681.

75. Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*. 2015;68(8):855-859.

76. Boyd K, Eng KH, Page CD. Area Under the Precision-Recall Curve: Point Estimates and Confidence Intervals. ECML PKDD 2013: Machine learning and knowledge discovery in databases. 2013;8190:451-466.

77. Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. *Statistics in Medicine.* 2014;33(18):3191-203.

2 Exploratory Data Analysis

2.1 Background

Childhood cancer survival has been steadily increasing over the latter half of the 20th century[1]. Based on estimates from the Surveillance, Epidemiology, and End Results (SEER) Program, the probability of surviving a childhood cancer diagnosis during the period of 1975-1979 was 67.8%[2]. This probability increased to almost 80% by 1995-1999, surpassed 81% during the first 5 years of the 21st century, and is currently over this value[2,3]. Due to the increased quantity of patients surviving childhood cancer, and the high likelihood of negative conditions developing during survival, there is an essential need to assess the impact of cancer treatment on the development of chronic conditions[4].

Consequently, the CCSS, a multi-institutional retrospective cohort study with a study population of over 24,000 cancer survivors was developed[4]. The CCSS uses questionnaires to follow-up childhood cancer survivors and assess the impact of their cancer and its treatment on the development of many conditions, including chronic physical and mental health disorders, second cancers, organ dysfunction and early death[4]. The study began in 1994 with an original cohort of over 14,000 individuals who were diagnosed with a malignant neoplasm between January 1, 1970 and December 31, 1986 before age 21[4]. The project was subsequently expanded in 2007 to include an additional 10,000 more survivors diagnosed between 1987 and 1999 comprising the expansion cohort[5].

Population Selection

Participants were selected for the CCSS cohorts from participating institutions across North America[4]. Individuals were eligible for the cohort if they were diagnosed and treated within one of the above specified time periods for leukemia, a central nervous system (CNS) malignancy (excluding meningioma and craniopharyngioma), Hodgkin lymphoma, non-Hodgkin lymphoma, neuroblastoma, soft tissue sarcoma, kidney cancer or bone cancer, and were 5 year survivors from the date of their initial diagnosis[4]. Out of 20,276 individuals eligible for the original cohort, and 14,962 eligible for the expansion cohort, 69.4% (n = 14,361) and 66.8% (n = 10,002) respectively were enrolled in the study.

Selection of individuals for the expansion cohort was performed through stratified random sampling. In an effort to expand research on low prevalence diagnoses of interest[6,7], individuals diagnosed with acute lymphoblastic leukemia (the most common childhood cancer) were undersampled[8]. Therefore, analysis using data from the combination of both cohorts requires weighting individuals within the expansion cohort with sampling weights to be analogous to the original cohort.

Survey Content

Initially, participants completed a baseline survey, which requested information regarding various demographic and lifestyle characteristics. These included questions regarding ethnicity, health habits (such as physical activity levels, smoking and alcohol consumption), education and employment history, medical conditions, prescribed medications, offspring and pregnancy history, as well as a family history[9,10]. Medical charts were reviewed by the CCSS centres to

obtain treatment data for each patient, specifically with respect to radiation and chemotherapy exposures. Information was collected regarding the specific chemotherapy agents and their respective doses, as well as radiation dose and location. Following the baseline survey, follow-up surveys were periodically administered to obtain updated information regarding health and well-being and to monitor the development of any adverse events. Five follow-up questionnaires (follow-up surveys 1-5) were released to the original cohort, and one follow-up questionnaire was released to the expansion cohort (identical to the follow-up 5 survey for the original cohort). A copy of the survey questions specific to determining ovarian status is included in Appendix A. Table 2.1 details when the surveys were released, as well as which surveys provided information sufficient to determine ovarian status.

Table 2.1 CCSS surveys released to the original and expansion cohorts

Survey	Years Released	Information Sufficient for Determining Ovarian Status
Original Cohort		
Original Baseline	10/1992 – 12/2002	No
Follow-up 1 (2000)	02/2000 – 12/2002	**Yes**
Follow-up 2 (2003)	11/2002 – 04/2005	No
Follow-up 3 (2005)	04/2005 – 11/2006	No
Follow-up 4 (2007)	07/2007 – 11/2009	**Yes**
Follow-up 5 (2014)	2014 – 2016	**Yes**
Expansion Cohort		
Expansion Baseline	05/2008 - present	**Yes**
Follow-up 5 (2014)	2014 – 2016	**Yes**

CCSS Study Sample Eligibility

For the purpose of developing risk prediction models for NSPM in childhood cancer survivors, data from the female survivors in the CCSS cohort was obtained. Initial exclusion criteria included failure to participate in a follow-up survey with sufficient menstrual history (MH) information to determine ovarian status, survey completion through a proxy (for individuals who were less than 18 at the latest follow-up survey completion, or those who were deceased following the 5 year survival window), as well as the absence of key MH information to determine ovarian status.

Individuals who overlapped with the St Jude Lifetime Cohort Study (SJLIFE) were set aside for external validation. Individuals with cranial or pituitary radiation greater than 30 Gy were excluded, as targeted high dose radiation to the brain can influence the timing of the release of important ovarian hormones as described in Chapter 1. Those with a SMN within the first 5 years of the primary cancer diagnosis were excluded, as consistent treatment exposure information was not uniformly collected. As individuals with AOF are no longer at risk of developing NSPM, those individuals were excluded during the analysis. Exclusions were also made for individuals with missing treatment exposure records for radiation and chemotherapy.

CCSS Study Sample Exclusions

Of the 11,336 females who completed the baseline survey, data was received for 8,770 individuals (77.4%). From the 2,566 excluded individuals, 1,774 individuals were excluded for failing to participate on a follow-up survey with sufficient MH information, 766 were excluded due to survey completion through a proxy, and 26 were excluded due to missing key MH

information. Within the remaining 8,770 individuals, 932 overlapped with the SJLIFE cohort, 808 individuals were exposed to cranial radiation greater than 30 Gy or had suspected pituitary dysfunction, the age at menopause was unable to be determined for 73 individuals, and 9 individuals had a SMN within 5 years. 1,086 individuals with missing treatment exposure records and 354 individuals with AOF were excluded during model development. A summary of the exclusions made from the total eligible sample is provided in Table 2.2.

The total study sample was comprised of 5,508 individuals, of which 4,054 (73.6%) were designated as training data for model development, and the remaining 1,454 (26.4%) were test data for internal validation. After accounting for sampling weights, the total study sample size was 6,252, with a training set of 4,644 (74.3%) and a test set of 1,608 (25.7%). All reported frequencies, population measures, and figures for the remainder of Chapter 2 are weighted with sampling weights.

Table 2.2 CCSS study sample exclusions

	Number of Observations	
Total Number of CCSS Female Participants	**11,336**	
No participation on survey with MH information		*1,774*
Proxy provided MH information		*766*
Excluded for missing key MH information		*26*
Data Received from CCSS	**8,770**	
Overlap with SJLIFE Cohort		*932*
Cranial or pituitary radiation > 30 Gy		*808*
Inability to determine age at menopause		*73*
SMN within 5 years of primary cancer diagnosis		9
Interim Total	**6,948**	
Missing key treatment data		1,086
Diagnosis of AOF		354
Training Data	4,054	
Test Data	1,454	
Total (unweighted)	**5,508**	

MH is menstrual health, CCSS is the Childhood Cancer Survivor Study, SJLIFE is the St Jude Lifetime Cohort Study, SMN is second malignant neoplasm and AOF is acute ovarian failure.

2.2 Baseline Characteristics

Diagnosis and Treatment Period

The years that an individual could be diagnosed with and treated for cancer to be eligible to participate in the CCSS were from 1970-1999. In order to observe the distribution of participants within this time period, the 30 years were divided into 6 categories (called "treatment periods"), which each category representing 5 consecutive years. Assessing the distribution of risk factors for NSPM by treatment period is important to observe how treatment exposures may differ and contribute to the increase in childhood cancer survivorship.

Age at Cancer Diagnosis

As described in Chapter 1, an increased age at cancer diagnosis is a risk factor for NSPM. In order to ensure consistency across treatment periods, the distribution of age at cancer diagnosis was examined. The average age at cancer diagnosis was 7.90 years, ranging from 0 to 21.0 with a median age of 6.3 (Table 2.3). Visually, the distribution of age at cancer diagnosis follows the same pattern in all the treatment periods, whereby there is an initial spike of diagnoses, followed by a decline (Figure 2.1). The percent of cancer diagnoses increases to a smaller peak around approximately age 15. The most recent treatment period has a lower percentage of individuals in the youngest age group (diagnosis age 0 and 1) compared to the remaining treatment periods, as many individuals diagnosed between those ages in that time period would not be eligible for the study based on their current attained age.

Table 2.3 Age at cancer diagnosis by treatment period

Frequency	Diagnosis and Treatment Period						Overall
	1970-1974	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	
***n* (%)**	365 (7.9)	658 (14.2)	912 (19.6)	893 (19.3)	957 (20.6)	859 (18.5)	4,644
Median Age	6.8	7.5	5.7	5.9	5.1	8.1	6.3
Mean (95% CI)	8.5 (7.8, 9.1)	8.7 (8.2, 9.1)	7.7 (7.3, 8.1)	7.4 (7.0, 7.8)	6.9 (6.5, 7.3)	8.9 (8.4, 9.4)	7.9 (7.7, 8.1)

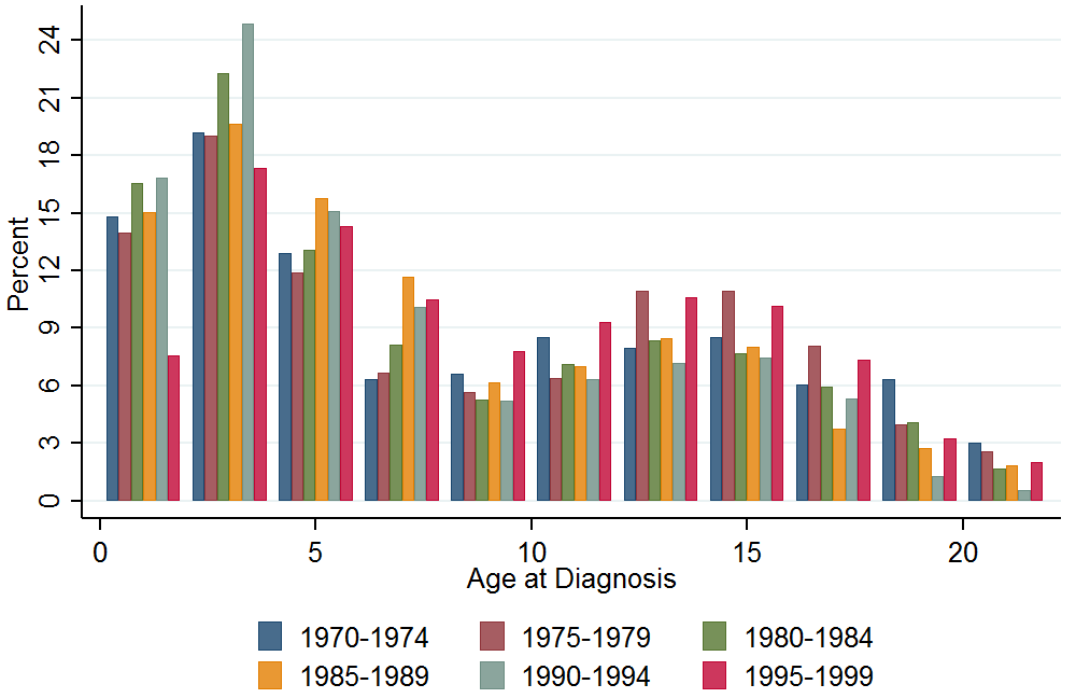


Figure 2.1 Age at cancer diagnosis by treatment period

Age at Menarche

131 individuals did not provide information on their age at menarche and 21 individuals were excluded due to suspected reporting or data entry error. Therefore, there were 4,446 individuals who provided accurate information regarding their age at menarche. Overall, the median age at menarche was 12 years. Visually, there is a large peak in the age at menarche in all cohorts at approximately 12 years old, with the majority of ages at menarche occurring between ages 10 and 15 (Figure 2.2).

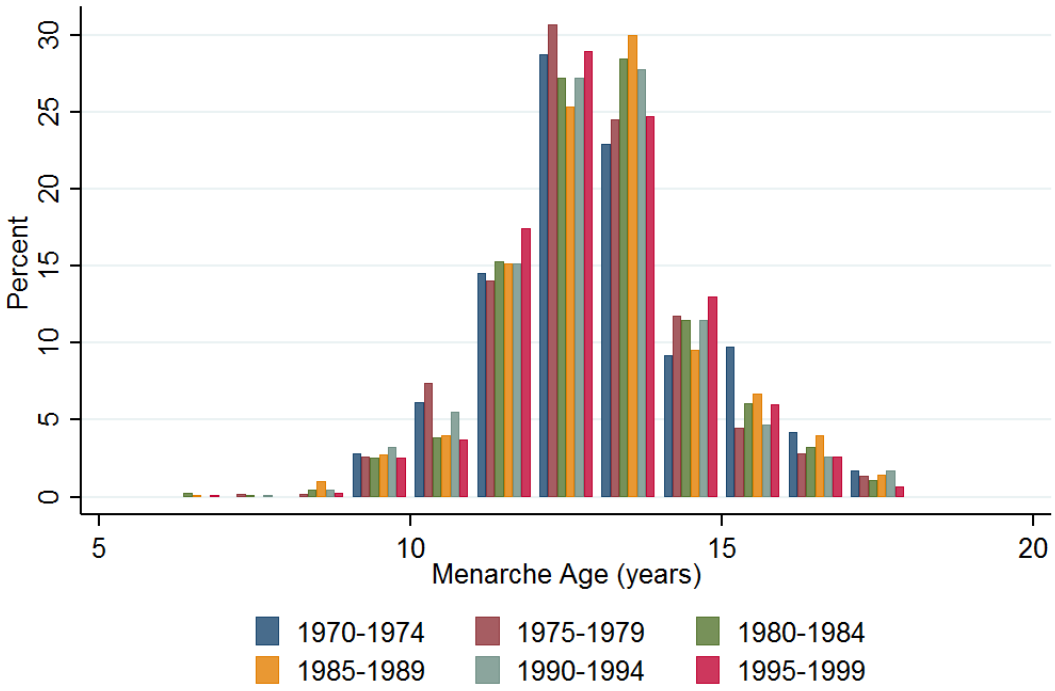


Figure 2.2 Age at menarche by treatment period

Primary Cancer Diagnosis

Patients were classified into 13 categories based on their primary cancer diagnosis group according to the International Classification of Diseases for Oncology (ICD-O)[4]. Specific cancer diagnosis codes which were eligible for inclusion in the study can be accessed online at https://ccss.stjude.org/content/dam/en_US/shared/ccss/documents/icdocodes.pdf.

The cancer diagnosis group with the highest proportion of individuals was acute lymphoblastic leukemia (ALL, *n* = 1,712) accounting for 36.9% of all diagnoses. The cancer diagnosis group with the lowest proportion of individuals was the group of medulloblastoma, primitive neuroectodermal tumor (PNET, *n* = 11; 0.2%), although "other bone tumors" and "other leukemia" had percentages less than 1% (Table 2.4).

Table 2.4 Frequency and percent of primary cancer diagnoses

Cancer Diagnosis Group	Total *n* (%)
ALL	1,712 (36.9)
Acute myeloid leukemia	151 (3.3)
Astrocytomas	357 (7.7)
Ewings sarcoma	126 (2.7)
Hodgkin lymphoma	540 (11.6)
Kidney tumors	519 (11.2)
Medulloblastoma, PNET	11 (0.2)
Neuroblastoma	337 (7.3)
Non-Hodgkin lymphoma	262 (5.6)
Osteosarcoma	283 (6.1)
Other bone tumors	26 (0.6)
Other CNS tumors	70 (1.5)
Other leukemia	29 (0.6)
Soft tissue sarcoma	221 (4.8)
Total	**4,644**

Percent represents column percent; PNET is a primitive neuroectodermal tumor, CNS is the central nervous system.

Race and Ethnicity

Individuals self-reported their race and ethnicity as White, Black, American Indian or Alaskan native and Asian or Pacific Islander, and Hispanic (Yes/No). Their responses were compared to data from the United States of America (USA) census from the years 1970, 1980 and 1990[11]. 4,034 individuals self-reported their race as White, which accounted for 86.9% of the total cohort (Table 2.5). Compared to the USA census data, the percent of White individuals during each treatment decade is slightly higher (93% versus 87.5%, 87.7% versus 83.1%, and 82.2% versus 80.3% respectively).

The percent of self-reported Black individuals was 4.6% overall. This is lower than the percent of Black individuals in the USA during 1970-1990, which ranged from 11.1% to 12.1% respectively (Table 2.5). There were a lower percentage of individuals identifying as Asian or Pacific Islander for all treatment periods within the study sample compared to the USA population (1.42% versus 1.73%). There were no individuals identifying as American Indian or Alaskan Native in the earliest treatment period (1970-1974), however by 1995-1999, they comprised 0.8% of the study population which aligns with the USA census proportion from the year 1990.

88.8% of the study sample overall reported to not be of Hispanic origin (Table 2.5). This is slightly lower than the overall population percent in the USA during the study period which ranged from 95.5% to 91%. 403 individuals identified as Hispanic, representing 8.7%, and falls within the proportion of Hispanic individuals in the USA during the study period, which ranged

from an estimated 4.5% in 1970 to 9.0% in 1990. Hispanic origin for 116 individuals is unknown.

Table 2.5 Self-reported race and ethnicity by treatment period

	Diagnosis and Treatment Period						Total
	1970-1974 *n* (%)	**1975-1979** *n* (%)	**1980-1984** *n* (%)	**1985-1989** *n* (%)	**1990-1994** *n* (%)	**1995-1999** *n* (%)	*n* (%)
Race							
American Indian or Alaskan Native	0	1 (0.2)	4 (0.4)	5 (0.6)	1 (0.1)	7 (0.8)	18 (0.4)
Asian or Pacific Islander	1 (0.3)	5 (0.8)	11 (1.2)	14 (1.6)	9 (0.9)	26 (3.0)	66 (1.4)
Black	15 (4.1)	12 (1.8)	29 (3.2)	43 (4.8)	60 (6.2)	55 (6.4)	213 (4.6)
Mixed Race	5 (1.4)	0	14 (1.5)	2 (0.2)	0	0	21 (0.5)
Unknown	9 (2.5)	20 (3.0)	35 (3.8)	65 (7.3)	77 (8.1)	86 (10.0)	292 (6.3)
White	335 (91.8)	620 (94.2)	819 (89.8)	764 (85.5)	810 (84.6)	686 (79.8)	4,034 (86.9)
Hispanic Origin							
No	331 (90.7)	604 (91.8)	826 (90.6)	781 (87.5)	847 (88.6)	736 (85.7)	4,126 (88.8)
Yes	15 (4.1)	34 (5.2)	49 (5.4)	87 (9.8)	99 (10.4)	118 (13.8)	403 (8.7)
Unknown	19 (5.2)	20 (3.0)	37 (4.1)	25 (2.8)	10 (1.1)	5 (0.6)	116 (2.5)

Percent represents column percent

Second Malignant Neoplasms

Due to the lack of availability of consistent treatment data, any individual diagnosed with a SMN *within* 5 years of their primary cancer diagnosis was excluded from the study sample. 195 individuals reported having a SMN after 5 years following their primary cancer diagnosis accounting for 4.2% of the total study population (Table 2.6). The ovarian status category with the highest proportion of survivors developing a SMN was NSPM, with 10.4% of the individuals diagnosed.

Table 2.6 Second malignant neoplasms by ovarian status

SMN	**Ovarian Status**			**Total** *n* (%)
	Normal *n* (%)	**SPM** *n* (%)	**NSPM** *n* (%)	
No	4,081 (91.7)	204 (4.6)	164 (3.7)	4,449 (95.8)
Yes	160 (82.2)	16 (8.2)	19 (9.5)	195 (4.2)

Percent represents row percent; Normal represents normal ovarian status at last MH survey completion before SMN diagnosis. SMN is second malignant neoplasm, SPM is surgical premature menopause, and NSPM is nonsurgical premature menopause.

2.3 Exposure Characteristics

Specific covariates examined as potential predictors involved the risk factors for NSPM, including age at cancer diagnosis and treatment exposures. Factors regarding chemotherapy assessed during model development included overall chemotherapy exposure (Yes/No), procarbazine exposure (Yes/No), and the cyclophosphamide equivalent dose (CED)[12]. Radiation therapy exposures included overall radiation exposure, maximum prescribed radiation dose to the abdomen, pelvis, total body, and minimum radiation dose to the ovaries. Collaboration with oncologists and researchers helped to ensure that all biologically significant factors and interactions were assessed during model development. Unless otherwise stated, table percents represent column percent.

Radiation Therapy Exposure

In the earlier periods of diagnosis, over 60% of survivors were exposed to radiation (Table 2.7). However, beginning in 1985, there was a reduction in the use of radiation and the percent of individuals exposed to radiation within each treatment period decreased. In the most recent treatment period, only 24% of survivors had exposure to radiation. Regions where radiation exposure is relevant for the development of NSPM include abdominal, pelvic, and ovarian radiation.

Table 2.7 Frequency of radiation exposure by treatment period

Received Radiation	**Diagnosis and Treatment Period**						**Total** *n* (%)
	1970-1974 *n* (%)	**1975-1979** *n* (%)	**1980-1984** *n* (%)	**1985-1989** *n* (%)	**1990-1994** *n* (%)	**1995-1999** *n* (%)	
No	122 (33.4)	225 (34.2)	438 (48.0)	598 (66.9)	737 (77.0)	653 (76.0)	2,773 (59.7)
Yes	243 (66.6)	433 (65.8)	474 (52.0)	296 (33.1)	220 (23.0)	206 (24.0)	1,872 (40.3)

Maximum Abdominal Radiation Dose

1,872 individuals were exposed to abdominal radiation. Except for the earliest treatment period (1970-1974) which had a median maximum abdominal radiation dose of 12 Gy, the median maximum dose of abdominal radiation was 2 Gy in all treatment periods (Figure 2.3). The earliest treatment period also had the largest maximum abdominal radiation dose of 60 Gy and the largest mean exposure of 14.9 Gy (95% CI = (12.9, 16.8)). Although the range of magnitude of exposure to abdominal radiation was similar across the treatment periods, the mean maximum doses are decreasing.

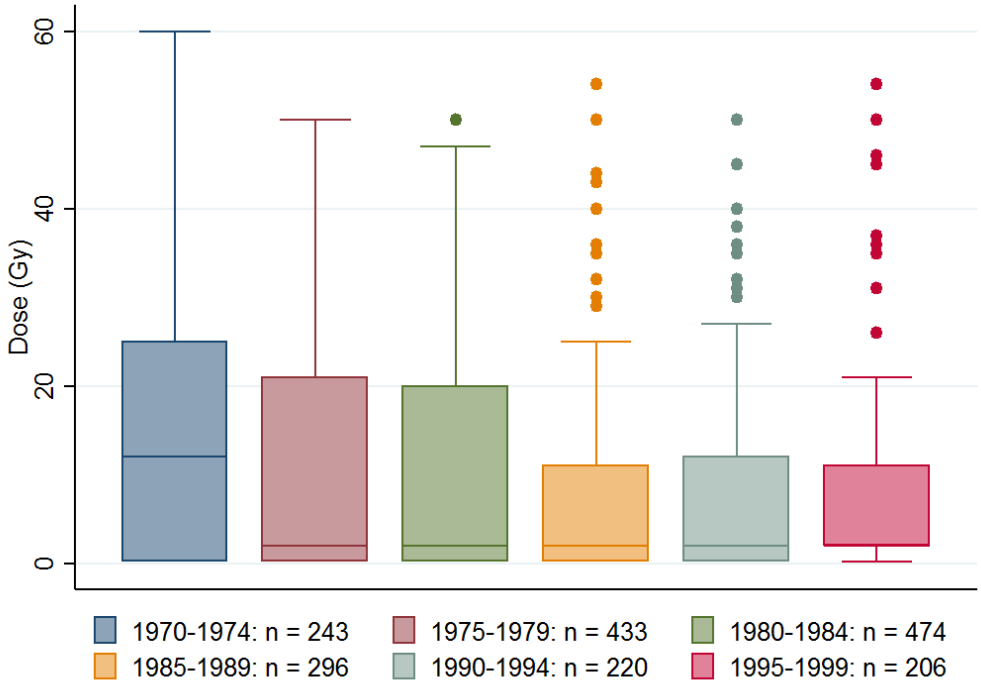


Figure 2.3 Maximum abdominal radiation dose by treatment period

Maximum Pelvic Radiation Dose

1,872 individuals had exposure to pelvic radiation. All treatment periods had a median and minimum dose of maximum pelvic radiation of 0.2 Gy and overall, a decreasing trend is seen in the magnitude as treatment period increases (Figure 2.4). All individuals with exposure to pelvic radiation also had exposure to abdominal radiation.

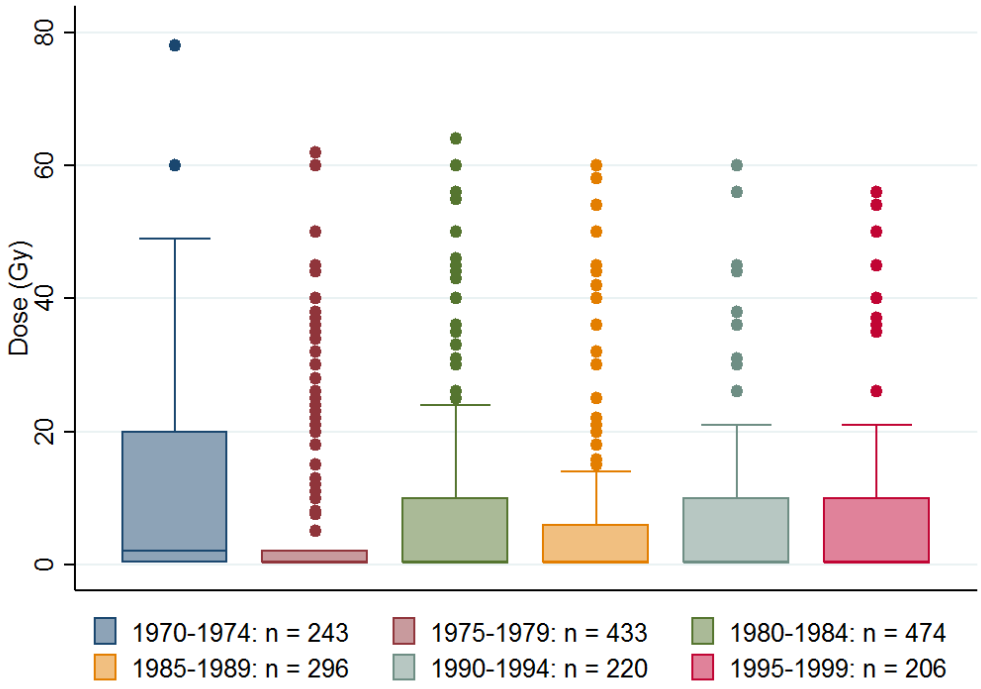


Figure 2.4 Maximum pelvic radiation dose by treatment period

Pituitary Radiation

Only participants with < 30 Gy of pituitary radiation are included in the analyses, as targeted cranial and pituitary radiation greater than 30 Gy can influence the timing of the release of important reproductive hormones as previously discussed. 11 participants had no information for their pituitary radiation exposure, and 1,865 individuals had exposure to pituitary radiation < 30 Gy. For individuals with exposure to pituitary radiation, the median pituitary radiation dose overall was 1.3 Gy, ranging from a minimum of 0.002 Gy to a maximum of 28.6 Gy (Figure 2.5). Exposure to pituitary radiation decreased over time, with the lowest range of dose exposure observed in the most recent treatment period. All individuals with exposure to abdominal, pelvic and ovarian radiation had exposure to pituitary radiation.

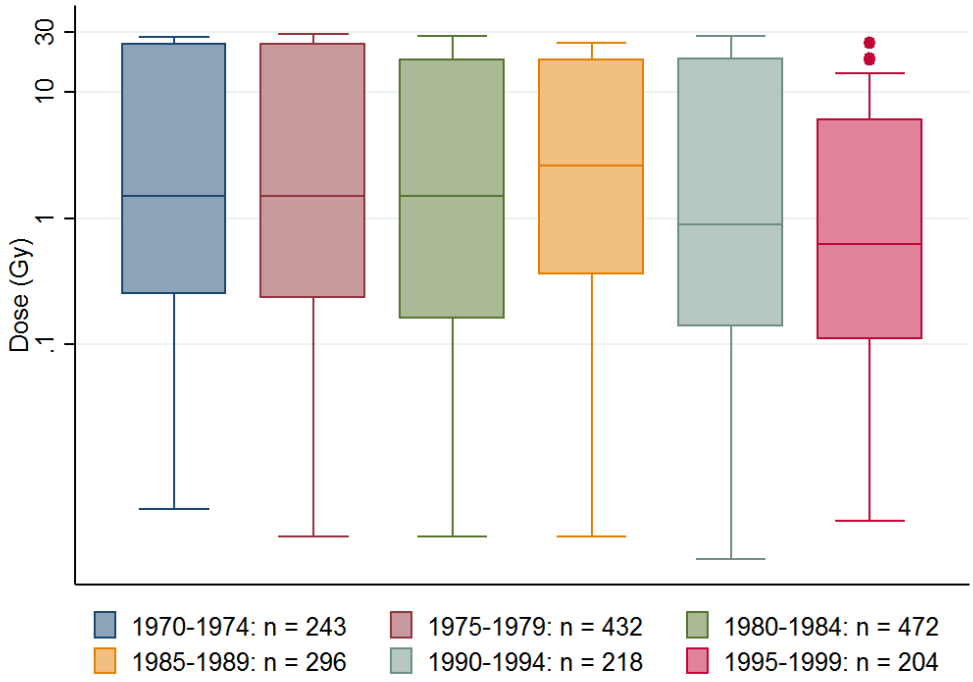


Figure 2.5 Pituitary radiation dose by treatment period

Minimum Ovarian Radiation

The specific magnitude of radiation exposure to the ovaries was calculated based on the dose received through abdominal, pelvic and total body irradiation. 1,871 individuals had exposure to ovarian radiation. The largest median dose of ovarian radiation was reported in the earliest treatment period (1970-1974) with a dose of 0.58 Gy. Overall, a decreasing trend is seen in the magnitude of ovarian radiation dose as the treatment period increases (Figure 2.6). The largest maximum ovarian radiation dose was 45.5 Gy in the most recent treatment period (1995-1999). The range of exposure to ovarian radiation is similar after excluding the earliest treatment period.

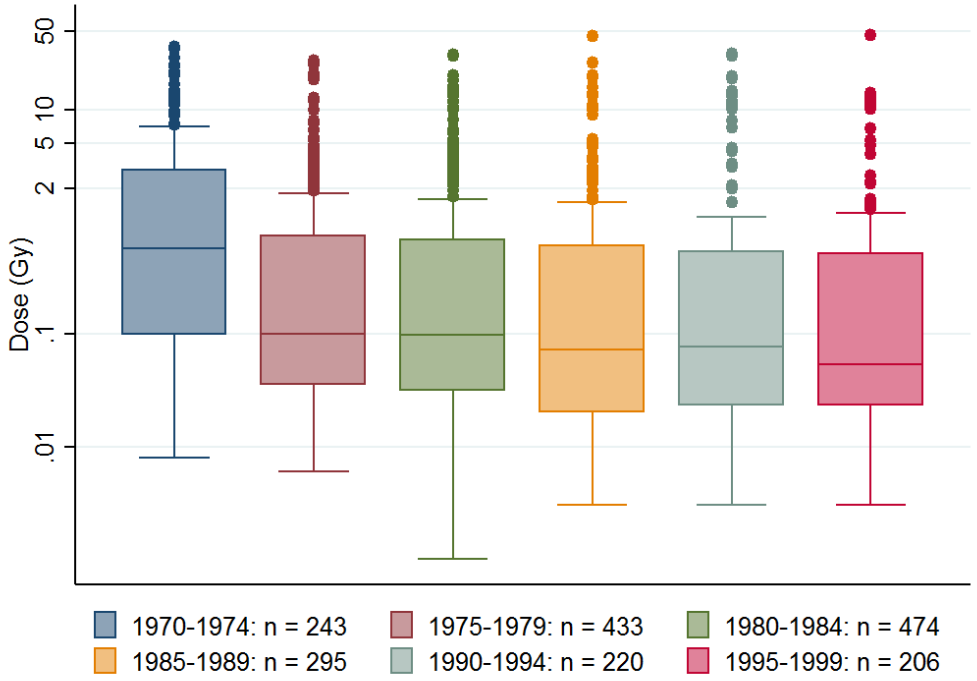


Figure 2.6 Minimum ovarian radiation dose by treatment period

Chemotherapy Exposure

Over all treatment periods, more individuals are exposed o chemotherapy than not (Table 2.8). The lowest proportion of chemotherapy exposure was during the first treatment period, where only 71.8% of the 365 individuals were exposed. For the remaining treatment periods, this value was above 80%.

Table 2.8 Frequency of chemotherapy exposure by treatment period

Received Chemotherapy	**Diagnosis and Treatment Period**						**Total** *n* (%)
	1970-1974 *n* (%)	**1975-1979** *n* (%)	**1980-1984** *n* (%)	**1985-1989** *n* (%)	**1990-1994** *n* (%)	**1995-1999** *n* (%)	
No	103 (28.2)	131 (19.9)	163 (17.9)	120 (13.4)	160 (16.7)	130 (15.1)	807 (17.4)
Yes	262 (71.8)	527 (80.1)	749 (82.1)	773 (86.6)	797 (83.3)	729 (84.9)	3,837 (82.6)

Cyclophosphamide Equivalent Dose

The cyclophosphamide equivalent dose (CED) is a measure of the cumulative alkylating agent exposure calculated through the standardization of 10 common alkylating agents (cyclophosphamide included) to the units of cyclophosphamide[12]. The equation for the calculation of the CED value is as follows:

CED (mg/m^2) = 1.0 (cumulative cyclophosphamide dose (mg/m^2))

+ 0.244 (cumulative ifosfamide dose (mg/m^2))

+ 0.857 (cumulative procarbazine dose (mg/m^2))

+ 14.286 (cumulative chlorambucil dose (mg/m^2))

+ 15.0 (cumulative carmustine dose (mg/m^2))

+ 16.0 (cumulative lomustine dose (mg/m^2))

+ 40.0 (cumulative melphalan dose (mg/m^2))

+ 50.0 (cumulative thiotepa dose (mg/m^2))

+ 100.0 (cumulative nitrogen mustard dose (mg/m^2))

+ 8.823 (cumulative busulfan dose (mg/m^2))

Cyclophosphamide Equivalent Dose Distribution

2,198 individuals were exposed to an alkylating agent included within the equation for calculating the CED value. CED values over all treatment periods ranged from 0.001 g/m^2 to 74.1 g/m^2 with a median of 6.2 g/m^2. The earliest treatment period had the largest mean and median of exposure and a slight decreasing trend is observed as the years increase (Figure 2.7).

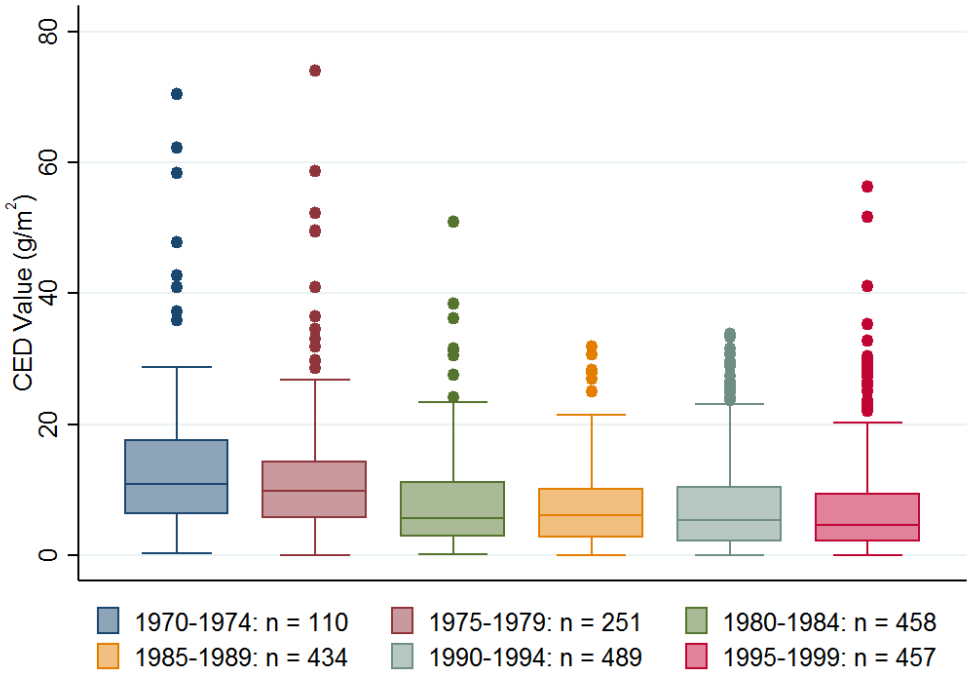


Figure 2.7 Distribution of CED values by treatment period

CED is the cyclophosphamide equivalent dose

Procarbazine Dose

337 individuals had exposure to procarbazine during their treatment. 95% (n = 320) of these individuals were diagnosed with Hodgkin lymphoma, with diagnoses including astrocytomas, medulloblastomas, non-Hodgkin lymphoma and other CNS tumours accounting for the remaining 5%. 5 individuals exposed to procarbazine were missing doses, leaving 332 individuals with values for the dose of procarbazine. The 1990-1994 treatment period had the largest maximum procarbazine dose of 17.5 g/m^2; but overall had a much narrower distribution of values and smaller median and mean exposure doses compared to the remaining treatment periods (Figure 2.8).

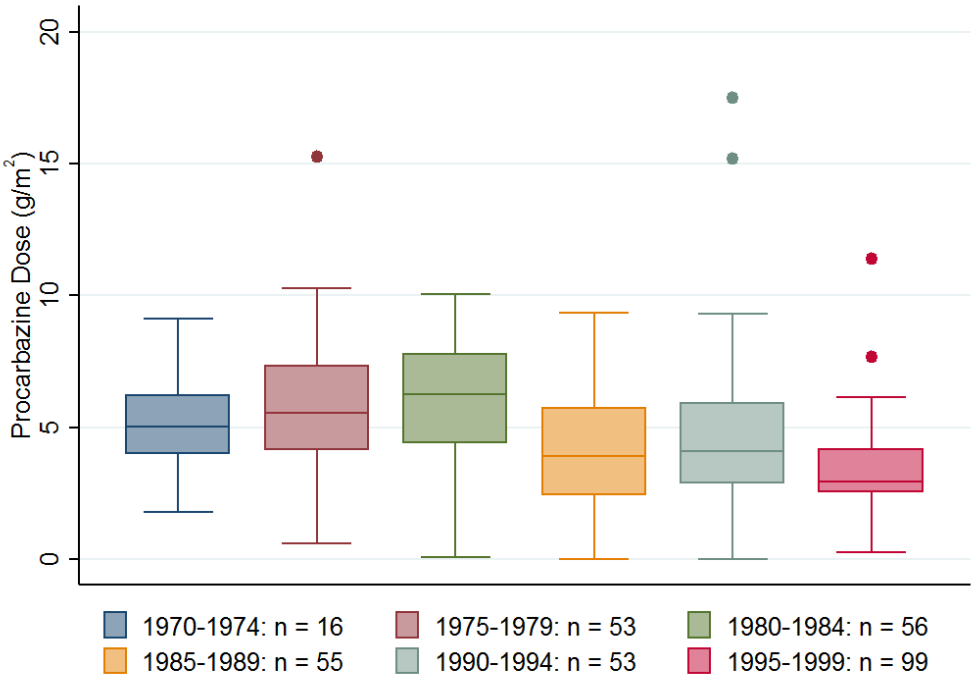


Figure 2.8 Procarbazine dose by treatment period

Bone Marrow Transplant

Specific BMT data was not uniformly collected for patients in the original cohort, but the procedure was generally preceded by TBI. Therefore, exposure to TBI was used as a proxy for indicating a bone marrow transplant in these individuals. In the expansion cohort, high doses of busulfan and cyclophosphamide generally replaced the preceding exposure of TBI for a BMT, but overall, more accurate data was collected for these participants. All individuals within the expansion cohort who had TBI exposure subsequently underwent a BMT. Overall, 158 individuals underwent a BMT, representing 3.4% of the total cohort (Table 2.9). However, the proportion of individuals within each treatment period who underwent this procedure varied. No individual underwent a BMT in the earliest treatment period, and less than 1% of individuals in the following two treatment periods were exposed. In the last 3 treatment periods, over 5% individuals within each category underwent a BMT.

Table 2.9 Frequency of BMT by treatment period

BMT	Diagnosis and Treatment Period						Total *n* (%)
	1970-1974 *n* (%)	**1975-1979** *n* (%)	**1980-1984** *n* (%)	**1985-1989** *n* (%)	**1990-1994** *n* (%)	**1995-1999** *n* (%)	
Yes	-	3 (0.5)	8 (0.9)	50 (5.6)	50 (5.3)	47 (5.5)	158 (3.4)
No	365 (100)	655 (99.5)	904 (99.1)	843 (94.4)	906 (94.7)	812 (94.5)	4,486 (96.6)

BMT is a bone marrow transplant

Treatment Combinations

There were 15 different combinations of relevant treatment exposures identified in the study sample. 1,509 individuals (32.5%) did not have any exposure to the primary treatment exposures assessed in the analysis. 23.6% of individuals had chemotherapy treatment (through an alkylating agent included in the CED value) as their only exposure. 19.4% of individuals were exposed to abdominal, pelvic, ovarian and pituitary radiation during their treatment without any additional treatment exposures. 19% of individuals were exposed to the above radiation treatments as well as an alkylating agent through the CED value. 34 individuals had exposure to the above radiation treatments, TBI before a BMT, as well as chemotherapy treatment with an alkylating agent. The remaining distribution of treatment combinations can be found in Table 2.10.

Table 2.10 Combinations of treatment exposures

Abdominal RT	Pelvic RT	Ovarian RT	Pituitary RT	TBI	BMT	Alkylating Agent	Total *n* (%)
X	X	X	X	X	X	X	34 (0.73)
X	X	X	X	X	X		9 (0.19)
X	X	X	X				901 (19.40)
X	X	X	X		X	X	36 (0.78)
X	X	X	X		X		8 (0.18)
X	X	X	X			X	883 (19.01)
X	X		X			X	1 (0.02)
					X	X	56 (1.21)
			X				4 (0.08)
					X		16 (0.33)
						X	1,189 (25.60)
							4,644

X represents whether the specific treatment was used in the regime, total number indicates the number of individuals who received the treatment regime. RT is radiation therapy, TBI is total body irradiation and BMT is bone marrow transplant

2.4 Outcome Characteristics

Age at Last Menstrual History Survey Completion

The age at which the individual completed the most recent survey with MH information indicates who remains at risk, as individuals are at risk for NSPM until they reach age 40 (or experience a competing event). The median age at last MH contact overall was 32 (minimum age = 18, maximum age 65), however this ranged from age 46 (minimum = 26, maximum = 65) in the earliest treatment period, to only 24 (minimum = 18, maximum = 41) in the most recent (Table 2.11). In the most recent treatment period, the vast majority of individuals have not reached age 40 yet, and are therefore still at risk for NSPM. This is in contrast to individuals from the earliest treatment periods, where the median age at last MH survey completion is greater than age 40. The difference in median age at last MH survey completion indicates a discrepancy in the proportion of individuals still at risk between the treatment periods.

Table 2.11 Age at last menstrual health contact by treatment period

Last MH Contact Age (years)	Diagnosis and Treatment Period						Total
	1970-1974	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	
Minimum	26	20	18	18	18	18	18
Median	46	42	35	32	27	24	32
Maximum	65	60	55	50	45	41	65
Mean (95% CI)	45.6 (44.8, 46.4)	42.0 (41.4, 42.7)	35.6 (35.0, 36.1)	32.7 (32.1, 33.0)	27.7 (27.2, 28.1)	25.4 (24.9, 25.8)	33.2 (32.9, 33.5)

MH is menstrual health

Ovarian Status

Data from responses to the survey questions assisted in classifying patients as having AOF, NSPM, SPM or neither. Since AOF is defined as amenorrhea within 5 years of diagnosis[13], a positive diagnosis can be determined by comparing the date of diagnosis with answers and dates indicating a loss of menstruation. A diagnosis of NSPM can be determined by assessing those individuals who retained normal ovarian function for at least 5 years after treatment but have indicated a lack of menstruation for at least 6 months before age 40.

Under the heading "Other Medical Conditions" in the follow-up 1 survey completed by the original cohort, an individual could indicate a hysterectomy or oophorectomy. In the remaining surveys released to the original and expansion cohorts, participants were asked to indicate if any of the following surgical procedures had been completed and the age at occurrence: removal of one ovary, removal of both ovaries or removal of uterus. After a surgical event such as a hysterectomy or bilateral oophorectomy, a patient is classified with SPM.

220 individuals (4.5%) underwent a surgical procedure initiating menopause prior to age 40, however the stratum-specific percentages of surgical premature menopause differ between treatment periods (Table 2.12). Over 8% of individuals diagnosed and treated within the first two treatment periods (during the 1970s) were classified as having SPM. In the most recent treatment periods, only 1.4% and 0.5% of individuals were classified with SPM respectively. This contrast can in part be attributed to the difference in median follow-up time between the various treatment periods, and the age specific probability of SPM which is higher with an increased

attained age. Therefore, the vast majority of individuals are still at risk for SPM in the most recent treatment periods.

Overall, the majority of individuals had normal ovarian function at the time of latest assessment (Table 2.12). Out of 4,918 individuals, 4,242 were classified as normal accounting for 86.2%. 5.6% of individuals (n = 274) were diagnosed with AOF and 3.7% of individuals (n = 183) were diagnosed with NSPM. The percent of cases within each treatment period does not differ considerably and range from a low of 2.8% during 1980-1984 to a high of 4.9% in 1970-1974. The slight differences are once again attributable to the difference in follow-up length between the various periods of diagnosis and treatment, and these percentages may adjust as individuals age and are diagnosed.

Table 2.12 Ovarian status stratified by treatment period

Ovarian Status	Diagnosis and Treatment Period						Total *n* (%)
	1970-1974 *n* (%)	**1975-1979** *n* (%)	**1980-1984** *n* (%)	**1985-1989** *n* (%)	**1990-1994** *n* (%)	**1995-1999** *n* (%)	
Normal	310 (76.4)	578 (82.5)	828 (87.7)	800 (86.8)	900 (89.2)	825 (88.1)	4,242 (86.2)
AOF	41 (10.1)	43 (6.1)	32 (3.4)	29 (3.1)	53 (5.2)	77 (8.2)	274 (5.6)
SPM	35 (8.6)	57 (8.1)	58 (6.1)	50 (5.5)	14 (1.4)	5 (0.5)	220 (4.5)
NSPM	20 (4.9)	23 (3.3)	26 (2.8)	43 (4.6)	42 (4.2)	29 (3.1)	183 (3.7)
Total	**406**	**701**	**944**	**922**	**1,009**	**936**	**4,918**

AOF is acute ovarian failure, SPM is surgical premature menopause and NSPM is nonsurgical premature menopause. Individuals with AOF are included in this table for descriptive purposes but are not included during model development.

Cumulative Incidence Curves

The curves in Figure 2.9 demonstrate the cumulative incidence trends for NSPM and SPM respectively. While the cumulative incidence of NSPM linearly increases with attained age, the cumulative incidence of SPM appears to follow an exponential distribution similar to the distribution of SPM observed for women in the general population of the USA[14]. Prevalence and incidence measures begin at age 26, as that is the latest age that an individual could have entered the CCSS cohort. At age 26, the cumulative incidence of NSPM was higher than the cumulative incidence of SPM by approximately 2.5%. With its exponential increase however, the cumulative incidence of SPM surpasses that of NSPM just after age 35. Stratified by treatment decade, the cumulative incidence curves for NSPM and SPM follow the same trend (Figure 2.10). The large jumps in the latest treatment decade are indicative of the lack of individuals with the events during that period.

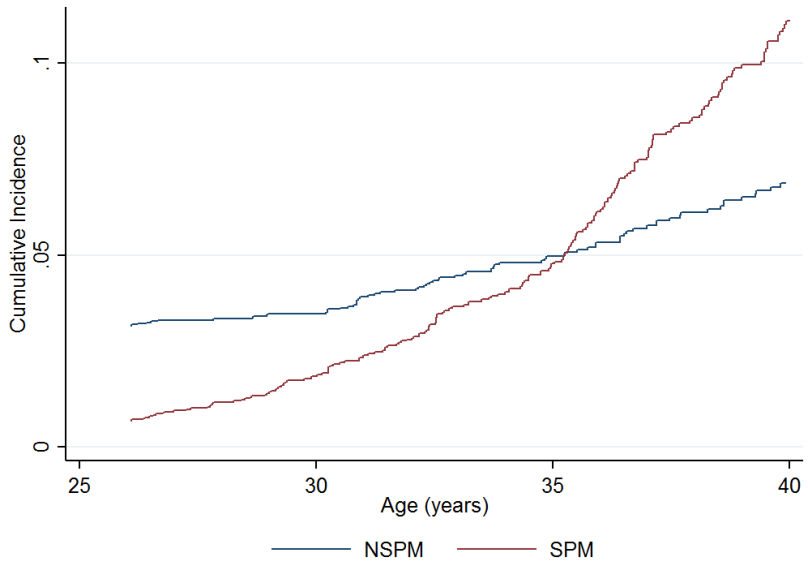
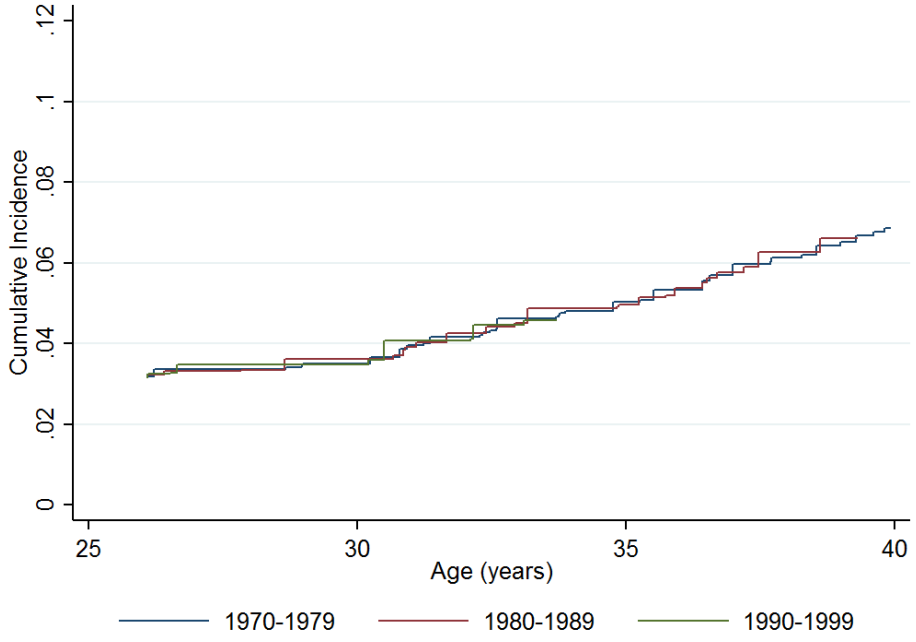


Figure 2.9 Cumulative incidence of NSPM and SPM

NSPM is nonsurgical premature menopause, SPM is surgical premature menopause

Nonsurgical Premature Menopause



Surgical Premature Menopause

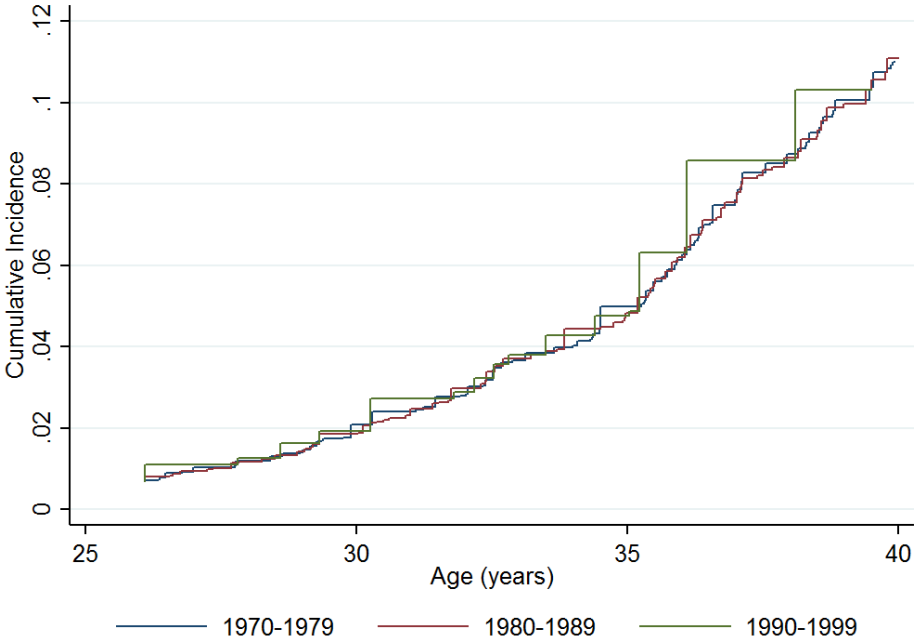


Figure 2.10 Cumulative incidence curves for NSPM and SPM by treatment decade

Prevalence

The prevalence at age 26 is a point prevalence value; however the remaining prevalence measures are calculated by using the period prevalence equation as in Chapter 1. The numerator includes those cases of NSPM or SPM that developed before age 26, as well as those individuals who developed the event during the time period. The denominator is the average population, which is taken by averaging the population size for each study year within the specific time period of interest.

The prevalence for NSPM was 3.21% (95% CI = (2.63%, 3.79%)) at age 26 (Table 2.13). By age 30, the period prevalence was 4.04% (95% CI = (3.36%, 4.73%)) and increased to 5.83% (95% CI = (4.96%, 6.71%)) by age 35. By age 40, the maximum age for which NSPM could be observed, the period prevalence was 7.81% (95% CI = (6.72, 8.90)). The prevalence values for SPM display a different trend compared to those for NSPM, with a larger increase in prevalence for each subsequent period. The prevalence of SPM at age 26 was 0.91%, increased to 2.22% by age 30, 5.33% by age 35, and 9.40% by age 40, which surpasses the prevalence of NSPM.

Table 2.13 Period prevalence values for NSPM and SPM

Time Period	Population Size at Specific Age	Average Period Population Size	NSPM *n*	NSPM % (95% CI)	SPM *n*	SPM % (95% CI)
By Age 26	3,525	3,525	113	3.21 (2.63, 3.79)	32	0.91 (0.60, 1.22)
By Age 30	2,814	3,185	129	4.04 (3.36, 4.73)	71	2.22 (1.71, 2.74)
By Age 35	1,951	2,743	160	5.83 (4.96, 6.71)	146	5.33 (4.49, 6.17)
By Age 40	1,311	2,343	183	7.81 (6.72, 8.90)	220	9.38 (8.20, 10.56)

NSPM is nonsurgical premature menopause and SPM is surgical premature menopause. The prevalence estimate calculated at age 26 is a point prevalence value.

Incidence Rates

Incidence rates for NSPM and SPM were calculated using Poisson regression for 10,000 PY at risk. Overall, the cumulative incidence rate for NSPM development by age 26, 30, 35 and 40 was similar (within 17.7 - 21.5 events per 10,000 PY). For the development of NSPM between ages of interest, the lowest incidence rate (7.4 events per 10,000 PY) occurred between ages 26-30, increased to 32.3 events per 10,000 PY for ages 30-35, and reached a maximum of 44.4 events per 10,000 PY between ages 35-40 (Table 2.14). No NSPM events were recorded between ages 35-40 for individuals diagnosed from 1990 onwards due to the low number of individuals having reached that age range. For SPM, the incidence rate increases with an increase in age instead of remaining similar. A similar increasing trend was observed for incidence rates calculated between ages of interest.

Table 2.14 Incidence rates for NSPM and SPM

Time	Diagnosis Year						Total
	1970-1974 IR (95% CI)	**1975-1979** IR (95% CI)	**1980-1984** IR (95% CI)	**1985-1989** IR (95% CI)	**1990-1994** IR (95% CI)	**1995-1999** IR (95% CI)	**Total** IR (95% CI)
NSPM							
By Age 26	13.3 (6.0, 29.5)	5.0 (1.9, 13.4)	10.4 (5.9, 18.2)	23.2 (13.5, 39.9)	29.9 (19.3, 46.4)	32.2 (18.4, 56.3)	19.9 (15.5, 25.5)
Between Ages 26-30	14.5 (3.6, 58.2)	8.5 (2,1, 34.0)	3.4 (0.5, 24.1)	7.5 (1.6,29.9)	6.2 (0.9, 44.4)	8.5 (1.2, 60.7)	7.4 (3.9, 14.2)
Between Ages 30-35	32.7 (13.6, 78.9)	37.8 (20.3, 70.2)	13.8 (5.2, 36.8)	43.6 (16.5, 115.0)	49.2 (20.4, 118.8)	30.0 (7.5, 120.7)	32.3 (22.2, 47.1)
Between Ages 35-40	55.0 (26.3, 115.1)	33.8 (16.1, 71.0)	57.2 (29.7, 110.0)	53.3 (22.1, 128.3)	-	-	44.4 (30.7, 64.4)
SPM							
By Age 26	2.2 (0.3, 15.8)	5.0 (1.9, 13.3)	8.6 (4.5, 16.0)	4.3 (1.8, 10.4)	1.8 (0.5, 7.4)	1.2 (0.2, 8.8)	4.2 (2.8, 6.3)
Between Ages 26-30	50.9 (24.3, 106.6)	46.6 (25.8, 84.2)	23.8 (11.3, 49.8)	29.8 (14.9, 59.8)	12.5 (3.1, 49.9)	8.5 (1.2, 60.5)	29.6 (21.4, 41.1)
Between Ages 30-35	52.3 (26.2, 104.6)	56.6 (34.2, 93.9)	62.2 (39.2, 98.6)	94.3 (49.2, 180.6)	37.8 (14.1, 100.9)	45.0 (14.5, 139.2)	61.9 (47.1, 81.3)
Between Ages 35-40	149.3 (95.3, 233.9)	130.5 (89.4, 190.5)	146.1 (96.9, 220.4)	198.6 (114.5, 344.5)	177.2 (78.7, 399.0)	-	149.0 (120.5, 184.2)

IR is the incidence rate, NSPM is nonsurgical premature menopause and SPM is surgical premature menopause.

2.5 References

1. Ries LAG, Eisner MP, Kosary CL, et al. Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975-1995, National Cancer Institute, SEER Program. *National Institutes of Health Publications.* 1999:No. 99-4649.

2. Noone AM, Howlader N, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review (CSR) 1975-2015: SEER Relative Survival (percent) by Year of Diagnosis, All Races, Males and Females, Ages 0-19.

3. Robison LL, Hudson MM. Survivors of childhood and adolescent cancer: life-long risks and responsibilities. *Nature Reviews Cancer.* 2014;14(1):61-70.

4. Robison LL, Mertens AC, Boice JD, et al. Study design and cohort characteristics of the Childhood Cancer Survivor Study: A Multi-Institutional Collaborative Project. *Medical and Pediatric Oncology.* 2002;38:229-239.

5. Childhood Cancer Survivor Study. Treatment Exposure Status: Expansion cohort as of December 2015, Original Cohort and Overall Cohort. Accessed February 2017, from https://ccss.stjude.org/content/dam/en_US/shared/ccss/documents/data/treatment-exposure-tables.pdf. 1995.

6. Armstrong GT. Childhood Cancer Survivor Study. *NIH U.S. National Library of Medicine.* Accessed August 2018, from: https://clinicaltrials.gov/ct2/show/NCT01120353. 1995.

7. Robison LL, Armstrong GT, Boice JD, et al. The Childhood Cancer Survivor Study: A National Cancer Institute-supported resource for outcome and intervention research. *Journal of Clinical Oncology.* 2009;27(14):2308-2318.

8. Turcotte LM, Liu Q, Yasui Y, et al. Temporal Trends in Treatment and Subsequent Neoplasm Risk Among 5-Year Survivors of Childhood Cancer, 1970-2015. *The Journal of the American Medical Association.* 2017;317(8):814-824.

9. Childhood Cancer Survivor Study. Original Cohort Baseline Survey: Long-Term Follow-Up Study of Individuals Treated for Cancer, Leukemia, Tumor or Similar Illness. Accessed August 2018, from: https://ccss.stjude.org/content/dam/en_US/shared/ccss/documents/survey/survey-baseline.pdf. 1992.

10. Childhood Cancer Survivor Study. Expansion Cohort Baseline Survey: Long-Term Follow-Up Study of Individuals Treated for Cancer, Leukemia, Tumor or Similar Illness. Accessed August 2018, from: https://ccss.stjude.org/content/dam/en_US/shared/ccss/documents/survey/survey-baseline-exp.pdf. 2008.

11. Gibson C, Jung K. Historical Census Statistics on Population Totals by Race, 1790 to 1990, and by Hispanic origin from 1970-1990 for the United States, Regions, Divisions and States. 2002; Working Paper No. 56.

12. 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. *Pediatric Blood and Cancer.* 2014;61:53-67.

13. Chemaitilly W, Mertens AC, Mitby P, et al. Acute Ovarian Failure in the Childhood Cancer Survivor Study. *The Journal of Endocrinology and Metabolism.* 2006;91(5):1723-1728.

14. Merrill R. Hysterectomy Surveillance in the United States, 1997 through 2005. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research.* 2008;14(1):CR24-31.

3 Evaluating Model Accuracy under Sampling Frame for Time-to-Event Data

Abstract

Background

Both sampling design and loss to follow-up of participants can impact the analysis of cohort studies. In order to obtain consistent parameter estimates, weights are required to account for these features in model development and evaluation. We considered several weight estimators that utilize sampling and censoring weights in the estimation of model accuracy measures.

Methods

We designed and implemented four simulation settings, varying the relationship between sampling design, censoring distribution and risk score distribution. Within each simulation, we assessed weighting scenarios with distinct weight estimators.

Results

Depending on the relationship between sampling design, censoring distribution and risk score distribution, one or more weighting estimators gave consistent estimates of the true accuracy assessment values.

Conclusions

Ignoring or inadequately accounting for weights can result in biased estimates of accuracy measures and mislead investigators regarding the accuracy of the developed models. When assessing risk prediction models developed using data from cohort studies, investigators need to evaluate how sampling design and censoring of participants are related to the covariates and the implications these may have on their model development and evaluation.

3.1 Introduction

The prominent feature of time-to-event data is censoring, which motivated the development of survival analysis methodology such as Kaplan-Meier's product limit estimator and Cox's proportional hazards regression[1,2]. The evaluation of corresponding risk prediction models developed using time-to-event data typically requires weighting observations with inverse probability-of-censoring weights to account for censoring, e.g., the time-specific area under the ROC curve (AUC_t)[3], the time-specific average positive predictive value (AP_t)[4], and the Brier score[5]. Additionally, cohort studies may involve a sampling design that enriches certain groups of participants. For example, survivors with the most frequently observed diagnosis of childhood cancer were undersampled during recruitment for the expansion cohort of the Childhood Cancer Survivor Study, which allowed for a larger contribution of rarer cancers in order to aid in research[6]. Hence, study design should also be considered during analysis and assessment of models developed from such populations.

While there exists abundant literature describing methods for accounting for design weights and censoring in model building[7-10], there remains a lack of investigations on how to account for these features during model accuracy assessment using appropriate weights. This paper aims to provide insight into this area through simulation studies.

Model Accuracy Assessment

Following the development of risk prediction models, it is essential to assess their performance. Risk prediction models are often assessed on their ability to correctly discriminate the event status at a future time t as positive (event observed) or negative (no event observed)[5]. Overall

model performance can be measured through the amount of explained variation (R^2) by the model, as well as by the Brier score, which computes the squared difference between predicted probabilities and true event statuses for binary outcomes[5,11]. How well the resulting predicted probabilities compare to the observed event probabilities is evaluated using calibration curves, which provide an indication of model reliability[5]. Although a variety of methods for assessing model performance have been described, we chose to focus on the estimation of AUC_t and AP_t values. The AUC_t is a popular summary measure of discrimination which is widely used in medical literature. Model predictive power can be assessed using AP_t, which has shown to be a more suitable performance measure when the event rate is low in the population[4].

Inverse Probability-of-Censoring Weights

Censoring occurs if individuals do not experience the event of interest during the study period or if they are lost to follow-up. The critical assumption for modelling censored time-to-event data is that the censoring process is at least independent, or ideally, non-informative[12]. We will focus on these two situations (non-informative and independent censoring) for this paper. For independent censoring, the event process and censoring process are assumed to be independent conditional on (a subset of) the covariates in the event time model. When censoring is non-informative, or "completely at random", it assumes that there is no relationship whatsoever between the event and censoring time distributions (i.e. individuals who are censored have the same event risk as those who remain uncensored)[9].

For model evaluation at a specific time point $t_0 > 0$, only a subset of the original study sample contributes full information, i.e., those individuals who have either not been censored by time t_0

or those who had the event on or before t_0[9,13]. Inverse probability-of-censoring weights (IPCW) are applied to the subset of individuals who contribute full information to account for the unknown event status of individuals censored before time t_0[9]. The IPCW method allows the resulting estimates to be unbiased as long as the censoring distribution is specified correctly[9,13,14]. For non-informative censoring, the censoring distribution can be estimated using the Kaplan-Meier method, and for independent censoring, the correct model-based estimate of the censoring distribution must be used. Individuals censored before time t_0 are given a weight of 0 and only indirectly contribute information through their contribution to the estimation of the censoring distribution[9].

Sampling Weights

Researchers may use one of a variety of sampling designs during participant enrollment depending on the aim of the study. Stratified random sampling occurs when individuals are divided into mutually exclusive sampling categories based on a characteristic (e.g., ethnicity or age group) and randomly sampled for the study from within each stratum[15]. By applying this sampling technique, investigators can enrich data from a particular subgroup that is of a lower proportion in the target population by oversampling the lower proportion subgroup and undersampling the other, more prevalent subgroups[15]. However, differential sampling causes the study sample to not be representative of the population of interest[16]. To draw appropriate inferences for the target population from the developed models, the sampling frame needs to be considered during analysis.

To account for the sampling design, observations are typically weighted by their inverse probability of being sampled, known as a sampling weight[8,17]. It can be shown that if the sampling scheme is unrelated to the outcome and the event time model is correctly specified, excluding sampling weights will not impact model parameter estimates[17]. However, ignoring sampling weights when the sampling category is related to the outcome (such as in informative sampling) can introduce bias into the parameter estimates, even after accounting for the sampling categories as a variable within the model[8,17]. Sampling weights may lead to an increase in the variance of parameter estimates; particularly if the variances within each sampling category are equal, the sample population size is small, or if there are substantial differences in the sampling probabilities between groups[18,19]. If the increase in variance from incorporating sampling weights is larger than the reduction of bias, the overall model error may increase when sampling weights are used. Therefore, the inclusion of sampling weights deserves careful consideration.

Objectives

We aim to investigate ways to combine both sampling and censoring weights in the estimation of model accuracy measures, and to answer two questions:

1. Should sampling weights be accounted for while estimating the censoring distribution for the inverse probability-of-censoring weights?
2. How can sampling and censoring weights be appropriately accounted for under various sampling and censoring settings?

The remainder of this paper is structured as follows. Section 2 introduces the notation and weighting scenarios. Section 3 presents four simulation settings where the weighting scenarios

are evaluated. Results are presented in Section 4, and findings are discussed and summarized in Section 5.

3.2 Methods

Notation

Sampling weights:

$$p_i = \frac{1}{s_i} \tag{1}$$

p_i is the sampling weight for individuals from the *i*th sampling category, and s_i is the probability an individual in sampling category i is selected for the study from the target population.

Following standard survival analysis notation, let T_j and C_j be the event and censoring times, respectively, for the jth individual. $X_j = \min(T_j, C_j)$ is the observed time for the jth individual, and $\delta_j = I(T_j < C_j)$ is the event status indicator[12]; 1 for individuals who experienced the event, and 0 for censoring.

The IPCW are given by [4]:

$$\hat{c}_{t,j} = \frac{I(X_j < t)\delta_j}{\hat{G}(X_j)} + \frac{I(X_j \geq t)}{\hat{G}(t)} \tag{2}$$

$\hat{G}(\cdot)$ is an estimator of the probability of remaining uncensored.

$\hat{G}(\cdot)$ can be estimated using the Kaplan-Meier (KM) method if the censoring distribution is non-informative. If the censoring distribution depends on a covariate included in the event time

model, then model-based estimates can be used for $\hat{G}(\cdot)$. If the censoring distribution is dependent on a covariate that is *not* a risk factor in the event time model, then the covariate can be ignored without jeopardizing the estimation of censoring weights. The problem arises if the censoring and event times both depend on an unmeasured covariate(s), leading to informative censoring (not discussed here).

The non-parametric estimators for $\mathrm{AUC_t}$ and $\mathrm{AP_t}$ are[4]:

$$\widehat{AUC_t} = \frac{\sum_{j=1}^{n} \hat{w}_{t,j} I(X_j > t) \widehat{\mathrm{TPF}}_t(Z_j)}{\sum_{j=1}^{n} \hat{w}_{t,j} I(X_j > t)} \tag{3}$$

$$\widehat{\mathrm{AP}}_t = \frac{\sum_{j=1}^{n} \hat{w}_{t,j} I(X_j \leq t) \widehat{\mathrm{PPV}}_t(Z_j)}{\sum_{j=1}^{n} \hat{w}_{t,j} I(X_j \leq t)} \tag{4}$$

$\hat{w}_{t,j}$ is the weight at time t for individual j, Z_j is the risk score value for subject j typically obtained from a model. $\widehat{\mathrm{TPF}}_t$ is the estimated true positive fraction at t (5) and $\widehat{\mathrm{PPV}}_t$ is the estimated positive predictive value at t (6).

The non-parametric estimators for $\mathrm{TPF_t}$ and $\mathrm{PPV_t}$ are[4]:

$$\widehat{\mathrm{TPF}}_t(Z_j) = \frac{\sum_{k=1}^{n} \hat{w}_{t,k} I(X_k \leq t) I(Z_k \geq Z_j)}{\sum_{k=1}^{n} \hat{w}_{t,k} I(X_k \leq t)} \tag{5}$$

$$\widehat{\mathrm{PPV}}_t(Z_j) = \frac{\sum_{k=1}^{n} \hat{w}_{t,k} I(X_k \leq t) I(Z_k \geq Z_j)}{\sum_{k=1}^{n} \hat{w}_{t,k} I(Z_k \geq Z_j)} \tag{6}$$

Weighting Scenarios

Various weighting scenarios were examined to assess the appropriate combination of censoring and sampling weights for $\widehat{w}_{t,j}$, the weight variable in (3 – 6).

1. Unweighted. Individuals who were censored before time t are ignored and given a weight of 0. Individuals with the event at or before t, or those that remained at risk at t, were given a weight of 1.

$$\widehat{w}_{t,j} = I(X_j \geq t) + I(X_j < t)\delta_j$$

2. Only sampling weights are included. Similar to scenario 1, except instead of the weight equal to 1 for individuals who have either not been censored by t or those who have had the event on or before t, it is equal to their corresponding sampling weight, p_i.

$$\widehat{w}_{t,j} = p_i\left[I(X_j \geq t) + I(X_j < t)\delta_j\right]$$

Scenarios 3 and 4 only include IPCW, but assess whether the sampling design should be accounted for while estimating the censoring distribution.

3. Only censoring weights included. The weights are estimated from the censoring distribution $\hat{G}(\cdot)$ which <u>does not</u> account for the sampling design (i.e. uses unweighted observations).

$$\text{IPCW}^1\text{: } \widehat{w}_{t,j} = \hat{c}^1_{t,j} \stackrel{\text{def}}{=} \hat{c}_{t,j}\left(\hat{G}(\cdot)\right)$$

4. Only censoring weights included. The weights are estimated from the censoring distribution $\hat{G}^{p_i}(\cdot)$ which does account for the sampling design in (2) (i.e., uses weighted observations in (2)).

$$\text{IPCW}^{p_i}: \widehat{w}_{t,j} = \hat{c}_{t,j}^{p_i} \stackrel{\text{def}}{=} \hat{c}_{t,j}\left(\hat{G}^{p_i}(\cdot)\right)$$

The following two scenarios are designed as double inverse probability weights[20]. They combine both censoring and sampling weights multiplicatively, and also assess whether the censoring distribution should account for the sampling design as in scenarios 3 and 4. The weights are given by:

$$\widehat{w}_{t,i} = (\text{sampling weight}) \times (\text{censoring weight})$$

5.

$$\widehat{w}_{t,j} = p_i \hat{c}_{t,j}^{1}$$

6.

$$\widehat{w}_{t,j} = p_i \hat{c}_{t,j}^{p_i}$$

Table 3.1 summarizes these six scenarios.

Table 3.1 Weights for estimating AUC_t, AP_t, and the event rate

Weights		Description	Weight Equation
Sampling	*Censoring*		
-	-	Unweighted	-
Yes	-	Only sampling weights	$\widehat{w}_{t,j} = p_i$
-	IPCW^1	$\hat{c}^1_{t,j}$ = IPCW estimated from the censoring distribution on unweighted samples	$\widehat{w}_{t,j} = \hat{c}^1_{t,j}$
-	IPCW^{p_i}	$\hat{c}^{p_i}_{t,j}$ = IPCW estimated from the censoring distribution on weighted samples	$\widehat{w}_{t,j} = \hat{c}^{p_i}_{t,j}$
Yes	IPCW^1	$\hat{c}^1_{t,j}$ multiplied by sampling weights	$\widehat{w}_{t,j} = p_i \hat{c}^1_{t,j}$
Yes	IPCW^{p_i}	$\hat{c}^{p_i}_{t,j}$ multiplied by sampling weights	$\widehat{w}_{t,j} = p_i \hat{c}^{p_i}_{t,j}$

$IPCW^1$ = Inverse probability-of-censoring weights – unweighted observations used

$IPCW^{p_i}$ = Inverse probability-of-censoring weights – weighted observations used

3.3 Simulation Studies

The appropriate weight to use during model evaluation will depend on whether the relationship between sampling design and censoring distribution with the event risk is non-informative or independent. Therefore, we designed simulation settings to assess the four combinations of non-informative and independent censoring and sampling. In the first setting, we assumed that the sampling design and censoring were both non-informative (i.e., unrelated to the distribution of event risk in the population). The second setting assumed the sampling design was independent but censoring was non-informative, and the third setting assumed non-informative sampling, but independent censoring. The last setting assumed both sampling design and censoring distributions were independent to the event risk distribution.

We generated four distinct fixed populations of 500,000 random risk scores, Z. Z represents the combined effect of multiple risk factors for an outcome, such as the combination of smoking status, age, and sex in the risk prediction of heart disease. The distribution of Z within each population is referred to as the risk score distribution. Within all populations, observations were randomly assigned to category 1 (40% of total population) or category 2 (60% of total population), to represent sampling design through a "sampling category" variable.

To simulate non-informative sampling in simulation settings i) and iii), risk scores for all observations were generated from one right skewed beta distribution ($Z \sim \text{beta}(\text{a} = 0.8, \text{b} = 3)$, mean = 0.210, standard deviation (SD) = 0.185). To represent independent sampling in simulation settings ii) and iv), risk scores were generated from distribution 1 ($Z \sim \text{beta}(\text{a} = 0.8, \text{b} = 3)$, mean = 0.210, SD = 0.185) if sampling category = 1, and from distribution 2

($Z \sim \text{beta}(\text{a} = 0.6, \text{b} = 4)$, mean = 0.131, SD = 0.142) if sampling category = 2, which was less variable than distribution 1. When the risk scores are obtained from different distributions for each sampling category, it follows that the sampling category is independent of the event time distribution conditional on Z. Z was then scaled by the standard deviation of distribution 1, defined as $Z_s = \frac{Z}{\text{SD}_1}$.

In all simulations, the true event time was generated through a Weibull distribution ($T \sim \text{Weibull}(3\text{e}^{-0.56Z_s}, 2)$, with a corresponding hazard function of $\lambda(t) = \frac{2}{9}t \cdot \exp(1.12Z_s)$. Non-informative censoring for settings i) and ii) was generated by obtaining a censoring time from a uniform distribution. For independent censoring in simulation settings iii) and iv), censoring time was generated from the following lognormal distribution, where $\varepsilon \sim N(0,1)$:

$$\log(C) = 1.4 - \beta Z_s + \varepsilon$$

β was set equal to 0.93 for setting iii), and 1.02 for setting iv) in order to obtain similar censoring rates at the pre-specified time t_0. The value of t_0 was chosen such that the corresponding event rate in the simulated population was approximately 10%. A summary of simulation settings is presented in Table 3.2.

Table 3.2 Summary of simulation settings

Setting	Type of Sampling	Type of Censoring	True Event Time	Risk Score	Censoring Time	t_0	Censoring Rate	
							Overall	*At* t_0
i)	Non-informative	Non-informative	$T \sim \text{Weibull}(3e^{-0.56Z_s}, 2)$	Z ~ beta(a = 0.8, b = 3)	C ~ uniform(0, 2.4)	0.38	58.2	15.2
ii)	Independent	Non-informative		Category 1: Z ~ beta(a = 0.8, b = 3) Category 2: Z ~ beta(a = 0.6, b = 4)	C ~ uniform(0, 3)	0.47	55.3	15.1
iii)	Non-informative	Independent		Z ~ beta(a = 0.8, b = 3)	$\log(C) = 1.4 - 0.93Z_s + \varepsilon$ $\varepsilon \sim N(0,1)$	0.38	45.2	15.3
iv)	Independent	Independent		Category 1: Z ~ beta(a = 0.8, b = 3) Category 2: Z ~ beta(a = 0.6, b = 4)	$\log(C) = 1.4 - 1.02Z_s + \varepsilon$ $\varepsilon \sim N(0,1)$	0.47	44.7	15.6

$Z_s = \frac{Z}{SD_1}$, *where* SD_1 *= 0.185 (the SD from distribution 1).*

The overall population AUC_t and AP_t values were obtained for risk score Z_s evaluated at t_0 using the "APBinary" function from the <APTools> package in R[21]. Each simulation was repeated 500 times. Within each repetition, a stratified random sample of n observations was selected without replacement from the respective fixed population. We examined two sample sizes, $n = 800$ and $n = 3000$, to assess the consistency of the estimators. Individuals in sampling category 1 were undersampled with a sampling probability of 0.2, leading to a sampling weight of 5. All individuals in category 2 were sampled proportionally, with a sampling weight of 1.

The Cox proportional hazards model was fit to the weighted and unweighted study sample:

$$\lambda(t) = \lambda_0(t)\exp(\beta Z_s)$$

$\widehat{w}_{t,j}$ were computed according to each weighting scenario and used to obtain estimates of AUC_t, AP_t and the event rate. For scenarios which included IPCW, three methods were used to estimate the censoring distribution in order to illustrate the effect of model misspecification: the Kaplan-Meier method ($IPCW_{KM}$), the Cox proportional hazards model ($IPCW_{Cox}$), and a lognormal model ($IPCW_{LN}$).

3.4 Results

Table 3.3 and Table 3.4 provide AUC_t, AP_t and event rate estimates using the various estimates of $\widehat{w}_{t,j}$ shown in Table 3.1 for sample size 800 and 3000 respectively. In all studies, the estimates of the accuracy measures and event rate were identical regardless of whether the censoring distribution was estimated using unweighted or weighted observations, due to the sampling design and censoring time being either unrelated or independent given the risk score. Estimates were also the same regardless of whether weighted or unweighted observations were used during model development with Cox proportional hazards regression, as the coefficient estimates were very similar (Appendix B). This was expected as the sampling design and event time were designed to be either unrelated or independent given the risk score. Therefore, only results from $IPCW^1$ scenarios are reported for all methods of estimating the censoring distribution.

In all studies, censoring weights were necessary to include to produce consistent estimates, regardless of whether the censoring distribution was non-informative or independent. When censoring was non-informative, i.e. settings i) and ii), there were no differences in the point estimates and their SDs when either the KM method ($IPCW_{KM}$) or the Cox proportional hazards model ($IPCW_{Cox}$) was used to estimate the censoring distribution. Modelling the censoring distribution with a lognormal distribution ($IPCW_{LN}$) slightly biased the AP_t and event rate estimates.

When censoring was independent, i.e. settings iii) and iv), using $IPCW_{KM}$ to estimate the censoring probability produced biased AUC_t, AP_t and event rate estimates. When the censoring distribution was correctly specified using the lognormal distribution ($IPCW_{LN}$), consistent

estimates were obtained with or without sampling weights. Estimates of the AUC_t, AP_t and event rate were not significantly different from the population quantities when $IPCW_{Cox}$ was combined with sampling weights multiplicatively; however, resulted in large empirical standard deviations.

When sampling was non-informative (in settings i) and iii)), including sampling weights multiplicatively was not required to produce consistent accuracy estimates. However, under the independent sampling of settings ii) and iv), the influence of sampling weights was observed. Both sampling and censoring weights were needed to give consistent estimates of the population quantities and excluding sampling weights produced biased estimates of the AUC_t and AP_t as well as the event rate.

Table 3.3 Simulation results; $n = 800$

Scenario		**Setting i)** *Non-informative Sampling, Non-informative Censoring*			**Setting ii)** *Independent Sampling, Non-informative Censoring*			**Setting iii)** *Non-informative Sampling, Independent Censoring*			**Setting iv)** *Independent Sampling, Independent Censoring*		
		$\widehat{\mathrm{AUC}}_t$ Mean (SD)	$\widehat{\mathrm{AP}}_t$ Mean (SD)	Event Rate (SD)	$\widehat{\mathrm{AUC}}_t$ Mean (SD)	$\widehat{\mathrm{AP}}_t$ Mean (SD)	Event Rate (SD)	$\widehat{\mathrm{AUC}}_t$ Mean (SD)	$\widehat{\mathrm{AP}}_t$ Mean (SD)	Event Rate (SD)	$\widehat{\mathrm{AUC}}_t$ Mean (SD)	$\widehat{\mathrm{AP}}_t$ Mean (SD)	Event Rate (SD)
Population Values		0.832	0.487	0.100	0.810	0.478	0.102	0.832	0.487	0.100	0.810	0.478	0.102
Weights													
Sampling	*Censoring*												
-	-	0.831 (0.027)	0.502 (0.061)	0.107 (0.012)	0.793 (0.035)	0.432 (0.067)	0.091 (0.011)	0.808 (0.034)	0.423 (0.068)	0.077 (0.010)	0.750 (0.041)	0.326 (0.066)	0.069 (0.010)
Yes	-	0.829 (0.037)	0.508 (0.078)	0.107 (0.016)	0.813 (0.040)	0.504 (0.091)	0.109 (0.016)	0.803 (0.046)	0.421 (0.088)	0.077 (0.013)	0.765 (0.049)	0.377 (0.096)	0.078 (0.014)
-	$\mathrm{IPCW}_{\mathrm{KM}}$	0.830 (0.027)	0.487 (0.061)	0.101 (0.011)	0.792 (0.036)	0.418 (0.067)	0.086 (0.010)	0.807 (0.035)	0.407 (0.067)	0.073 (0.009)	0.748 (0.041)	0.314 (0.066)	0.066 (0.009)
Yes	$\mathrm{IPCW}_{\mathrm{KM}}$	0.829 (0.037)	0.493 (0.078)	0.101 (0.015)	0.811 (0.040)	0.490 (0.091)	0.103 (0.015)	0.802 (0.046)	0.406 (0.087)	0.072 (0.012)	0.763 (0.079)	0.364 (0.095)	0.074 (0.014)
-	$\mathrm{IPCW}_{\mathrm{Cox}}$	0.830 (0.028)	0.487 (0.062)	0.101 (0.011)	0.792 (0.036)	0.418 (0.067)	0.086 (0.010)	0.831 (0.044)	0.503 (0.161)	0.103 (0.024)	0.782 (0.069)	0.420 (0.199)	0.097 (0.068)
Yes	$\mathrm{IPCW}_{\mathrm{Cox}}$	0.829 (0.037)	0.493 (0.078)	0.101 (0.015)	0.812 (0.040)	0.490 (0.091)	0.103 (0.015)	0.825 (0.053)	0.498 (0.168)	0.102 (0.028)	0.794 (0.080)	0.468 (0.220)	0.114 (0.082)
-	$\mathrm{IPCW}_{\mathrm{LN}}$	0.830 (0.028)	0.479 (0.061)	0.096 (0.011)	0.792 (0.036)	0.410 (0.066)	0.082 (0.010)	0.830 (0.034)	0.504 (0.115)	0.100 (0.014)	0.781 (0.046)	0.416 (0.137)	0.086 (0.016)
Yes	$\mathrm{IPCW}_{\mathrm{LN}}$	0.828 (0.037)	0.485 (0.078)	0.096 (0.014)	0.811 (0.040)	0.481 (0.090)	0.098 (0.015)	0.825 (0.045)	0.501 (0.133)	0.100 (0.019)	0.795 (0.059)	0.470 (0.172)	0.101 (0.028)

SD is the empirical standard deviation of the mean

Table 3.4 Simulation results; $n = 3000$

Scenario		**Setting i)** *Non-informative Sampling, Non-informative Censoring*			**Setting ii)** *Independent Sampling, Non-informative Censoring*			**Setting iii)** *Non-informative Sampling, Independent Censoring*			**Setting iv)** *Independent Sampling, Independent Censoring*		
		$\widehat{\text{AUC}}_t$ Mean (SD)	$\widehat{\text{AP}}_t$ Mean (SD)	Event Rate (SD)	$\widehat{\text{AUC}}_t$ Mean (SD)	$\widehat{\text{AP}}_t$ Mean (SD)	Event Rate (SD)	$\widehat{\text{AUC}}_t$ Mean (SD)	$\widehat{\text{AP}}_t$ Mean (SD)	Event Rate (SD)	$\widehat{\text{AUC}}_t$ Mean (SD)	$\widehat{\text{AP}}_t$ Mean (SD)	Event Rate (SD)
Population Values		0.832	0.487	0.100	0.810	0.478	0.102	0.832	0.487	0.100	0.810	0.478	0.102
Weights													
Sampling	*Censoring*												
-	-	0.830 (0.014)	0.500 (0.030)	0.108 (0.006)	0.791 (0.017)	0.424 (0.033)	0.091 (0.006)	0.810 (0.017)	0.421 (0.034)	0.078 (0.005)	0.754 (0.021)	0.326 (0.036)	0.069 (0.005)
Yes	-	0.831 (0.018)	0.501 (0.039)	0.107 (0.008)	0.812 (0.020)	0.496 (0.047)	0.109 (0.009)	0.807 (0.023)	0.418 (0.046)	0.078 (0.007)	0.771 (0.025)	0.376 (0.053)	0.078 (0.007)
-	IPCW_{KM}	0.829 (0.014)	0.485 (0.030)	0.101 (0.005)	0.790 (0.017)	0.410 (0.033)	0.086 (0.005)	0.808 (0.017)	0.404 (0.034)	0.073 (0.005)	0.753 (0.021)	0.314 (0.036)	0.066 (0.004)
Yes	IPCW_{KM}	0.830 (0.018)	0.486 (0.039)	0.101 (0.007)	0.811 (0.020)	0.482 (0.047)	0.103 (0.008)	0.805 (0.024)	0.402 (0.046)	0.073 (0.007)	0.769 (0.025)	0.362 (0.053)	0.074 (0.007)
-	IPCW_{Cox}	0.830 (0.014)	0.485 (0.030)	0.101 (0.005)	0.790 (0.017)	0.410 (0.033)	0.086 (0.005)	0.840 (0.031)	0.523 (0.131)	0.111 (0.051)	0.805 (0.065)	0.476 (0.200)	0.109 (0.074)
Yes	IPCW_{Cox}	0.830 (0.018)	0.486 (0.039)	0.101 (0.007)	0.811 (0.020)	0.482 (0.047)	0.103 (0.008)	0.836 (0.037)	0.516 (0.143)	0.111 (0.048)	0.819 (0.068)	0.528 (0.210)	0.130 (0.091)
-	IPCW_{LN}	0.829 (0.014)	0.476 (0.030)	0.097 (0.005)	0.789 (0.017)	0.401 (0.033)	0.082 (0.005)	0.834 (0.017)	0.505 (0.062)	0.101 (0.008)	0.790 (0.022)	0.423 (0.077)	0.086 (0.007)
Yes	IPCW_{LN}	0.830 (0.018)	0.477 (0.039)	0.096 (0.007)	0.810 (0.020)	0.473 (0.047)	0.099 (0.008)	0.831 (0.023)	0.501 (0.080)	0.101 (0.010)	0.807 (0.030)	0.486 (0.107)	0.102 (0.014)

SD is the empirical standard deviation of the mean

3.5 Discussion

Current literature highlights the importance of including sampling weights during model development, and discusses procedures for applying IPCW to account for censoring of observations for the estimation of unbiased model parameters[7-10]. We designed and conducted simulation studies to evaluate weighting methods in order to determine the appropriate weights to use when evaluating survival model performance under a sampling design. As the assumption of at least independent censoring is critical for the analysis of censored time-to-event data, we did not assess scenarios with informative censoring, and instead focused on non-informative and independent censoring situations. A non-informative sampling or censoring relationship was simulated by independently generating the event time, censoring time and sampling category variables. Independent sampling was simulated by generating risk scores based on sampling category variable. Independent censoring was simulated by computing the censoring time from the risk scores.

In all simulations, we observed that identical results were produced regardless if sampling design was accounted for while estimating the censoring distribution. That is, identical estimates were obtained if weighted or unweighted observations were used to estimate the censoring distribution. The implications of this finding are applicable to modelling the censoring distribution in order to obtain IPCW – if sampling and censoring are at least independent, then an unweighted sample can be used to estimate the censoring distribution.

Our simulations provide empirical evidence that IPCW to account for censoring must be included when performance metrics are estimated for survival algorithms. When sampling is

independent of the event time conditional on covariates, sampling weights are to be included multiplicatively to produce unbiased estimates. If censoring is non-informative, the censoring distribution for calculating IPCW can be estimated with the Kaplan-Meier method ($IPCW_{KM}$).

For situations with independent censoring, if the cohort sampling design is non-informative and the censoring distribution is correctly specified ($IPCW_{LN}$ in our case), then sampling design is ignorable, and including only $IPCW_{LN}$ produces unbiased accuracy estimates. However, if the censoring distribution is correctly specified but the sampling design is independent, sampling weights are non-ignorable and must be included multiplicatively in order to produce unbiased estimates. Study results for the specific situations that we assessed are summarized in Table 3.5.

Table 3.5 Summary of findings

Setting	Type of Sampling	Type of Censoring	Weights	
			Sampling	*Censoring*
i)	Non-informative	Non-informative	No/Yes	$IPCW_{KM}$ or $IPCW_{cox}$
ii)	Independent	Non-informative	Yes	$IPCW_{KM}$ or $IPCW_{cox}$
iii)	Non-informative	Independent	No/Yes	$IPCW_{LN}$
iv)	Independent	Independent	Yes	$IPCW_{LN}$

Our study confirms the theoretical result that in order to use IPCW to consistently estimate the required measures, the censoring distribution must be correctly specified in the calculation[14]. When settings iii) and iv) were simulated with independent censoring, only the weighting scenarios which correctly specified the censoring distribution, i.e., using $IPCW_{LN}$, produced

consistent estimates for all required measures. When a non-informative censoring distribution was assumed ($IPCW_{KM}$) or when the censoring distribution was modelled incorrectly ($IPCW_{Cox}$), biased estimates were produced. Additionally, estimates of the AP_t from scenarios where $IPCW_{Cox}$ was used were highly variable compared to those with $IPCW_{LN}$ or $IPCW_{KM}$, indicating the consequences of model misspecification for the censoring distribution. This study was performed in a hypothetical situation where the model-based estimator of the censoring distribution is known. Therefore it may be difficult to ascertain what the correct model-based estimator is when using data from a real population.

We focused on determining the correct method for combining sampling and censoring weights for model accuracy estimates, as the incorporation of weights during model development has been thoroughly evaluated elsewhere[7,8,10,14]. We evaluated a situation where sampling design does not have an effect on the coefficient estimates of the model, and is therefore ignorable during model development. Therefore, our findings may not extend to other circumstances where the sampling design impacts model coefficient estimation, such as under informative sampling.

When assessing risk prediction models developed using data from cohort studies, investigators should pay close attention to sampling design and carefully model the censoring distribution. Inadequately accounting for weights can result in accuracy estimates which do not reflect the model performance in the target population of their risk prediction models. In particular, an assessment of how the risk score distribution is related to both the sampling design and censoring distribution should be undertaken in order to ensure that the correct weights are used when model performance is evaluated.

3.6 References

1. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*. 1958;53(282):457-481.

2. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Society Interface. Series B (Methodological)*. 1972;34(2):187-220.

3. Heagerty PJ, Lumley T, Pepe MS. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics*. 2000;56:337-344.

4. Yuan Y, Zhou QM, Li B, Cai H, Chow EJ, Armstrong GT. A threshold-free summary index of prediction accuracy for censored time to event data. *Statistics in Medicine*. 2018;37(10):1671-1681.

5. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology*. 2010;21(1):128-138.

6. Turcotte LM, Liu Q, Yasui Y, et al. Temporal Trends in Treatment and Subsequent Neoplasm Risk Among 5-Year Survivors of Childhood Cancer, 1970-2015. *The Journal of the American Medical Association*. 2017;317(8):814-824.

7. Boudreau C, Lawless JF. Survival Analysis Based on the Proportional Hazards Model and Survey Data. *The Canadian Journal of Statistics*. 2006;34(2):203-216.

8. Cesar CC, Carvalho MS. Stratified sampling design and loss to follow-up in survival models: evaluation of efficiency and bias. *BMC Medical Research Methodology*. 2011;11:99.

9. Willems S, Schat A, van Noorden M, Fiocco M. Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. *Statistical Methods in Medical Research*. 2018;27(2):323-335.

10. Spittal MJ, Carlin JB, Currier D, et al. The Australian Longitudinal Study on Male Health sampling design and survey weighting: implications for analysis and interpretation of clustered data. *BMC Public Health*. 2016;16(Suppl 3):1062.

11. Gerds TA, Cai T, Schumacher M. The Performance of Risk Prediction Models. *Biometrical Journal*. 2008;50(4):457-479.

12. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data.* 2nd ed. John Wiley & Sons; 2002.

13. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*. 2011;22(3):278–295.

14. Hajducek DM, Lawless JF. Estimation of finite population duration distributions from longitudinal survey panels with intermittent followup. *Lifetime Data Analysis*. 2013;19:371-392.

15. Kalsbeek W, Heiss G. Building Bridges Between Populations and Samples in Epidemiological Studies. *Annual Review of Public Health*. 2000;21(1):147-169.

16. Gordis L. *Epidemiology.* 5th ed. Elsevier Canada; 2013.

17. Lavallée P, Beaumont J. Why We Should Put Some Weight on Weights. *Survey Methods: Insights from the Field*. 2015; Weighting: Practical Issues and 'How to' Approach (Invited Article).

18. Elliott MR. Bayesian weight trimming for generalized linear regression models. *Survey Methodology*. 2007;33(1):23-34.

19. United Nations. Chapter 6: Construction and use of sample weights. In: *Designing Household Survey Samples: Practical Guidelines.* New York: United Nations Publication; 2008:109.

20. Zhou QM, Zheng Y, Chibnik LB, Karlson EW, Cai T. Assessing incremental value of biomarkers with multi-phase nested case-control studies. *Biometrics*. 2015;71:1139-1149.

21. Cran.R-project.org. APtools: Average positive predictive values (AP) for binary outcomes and censored event times. Accessed August 2018, from: https://cran.r-project.org/web/packages/APtools/index.html.

4 Risk Prediction for Nonsurgical Premature Menopause in Childhood Cancer Survivors

Abstract

Introduction

Female childhood cancer survivors are at an increased risk of developing nonsurgical premature menopause (NSPM) due to toxicities from their treatment. For NSPM to occur, menopause must develop non-surgically before age 40, with ovarian function having been retained for at least 5 years following a cancer diagnosis. Such a condition can negatively impact quality of life and reduce potential reproductive years. Risk factors for NSPM in the literature include an older age at cancer diagnosis and treatment with high doses of chemotherapy and radiation. In order to facilitate informed discussions between physicians, patients, and their families regarding the need for fertility preservation interventions, we aimed to develop prediction algorithms to estimate the risk of patients developing NSPM.

Methods

Data was acquired for 5,508 female participants of the Childhood Cancer Survivor Study. Candidate models were developed on a training set of 4,054 observations using the time-specific logistic regression model with competing risks (TLR-CR), the Fine-Gray regression model (FGR), and the random survival forest method with competing risks (RSF-CR). Model performance and accuracy were measured using the time-specific area under the ROC curve (AUC_t), the time-specific average positive predictive value (AP_t), and calibration curves on both the training set and a test set of 1,454 observations for internal validation.

Results

Following model development, final predictor variables consisted of various risk factors including minimum ovarian radiation dose, cumulative chemotherapy exposure, bone marrow transplant, and age at cancer diagnosis. During model evaluation, the TLR-CR, FGR, and RSF-CR models performed similarly on the training set. At 15 years post cancer diagnosis, AUC_t values were between 0.72-0.77, and AP_t values were larger than the event rate of 1.7% (AP_t = 5.0-6.0%) indicating adequate model performance. All AP_tvalues remained larger than the event rate of 2.5% (AP_t= 8.4-9.9%), and the ratio between AP_t and the event rate increased for the RSF-CR and TLR-CR models. AUC_t and AP_t values from the test set over 10-20 years post cancer diagnosis showed similar findings. The models were well calibrated on both datasets, especially for low risk patients, but only the TLR-CR model was consistently well calibrated for high risk patients.

Conclusions

Obtaining risk estimates for NSPM has the potential to improve the lives of childhood cancer survivors by providing information to enhance discussions of fertility preservation. Overall, the TLR-CR model performed consistently with good calibration from the training to test set, and was the best model of the three. Moving forward, generalizability will be assessed through validation on an external cohort.

4.1 Introduction

The increased survival rate following a childhood cancer diagnosis over the previous few decades has consequently increased the number of survivors who are at risk of developing chronic conditions later in life due to the toxicities of their treatment[1]. Ovarian dysfunction is a primary concern of female childhood cancer survivors after completing chemotherapy or radiation treatment[2,3]. Nonsurgical premature menopause (NSPM), a specific form of ovarian dysfunction, occurs when ovarian function is maintained for at least 5 years after diagnosis with cancer, but menopause develops naturally before age 40[4]. NSPM can severely limit potential reproductive years, reduce quality of life, and increase anxiety and depressive feelings[2,3,5-7]. Female childhood cancer survivors have a significantly increased risk of developing NSPM compared to females in the general population[8], and are 10.5 times more likely to develop the condition compared to their otherwise healthy siblings[8].

Risk factors for NSPM have been well established, and include exposure to high doses of certain chemotherapy agents, targeted ovarian radiation, as well as an older age at cancer diagnosis[8,9]. It was reported that exposure to greater than 5 Gy of ovarian radiation significantly increased the odds of developing NSPM by 700%, and that individuals receiving a dose of procarbazine greater than or equal to 4000 mg/m^2 had 9 times the odds of developing NSPM compared to individuals with no exposure[8].

Cancer patients are also at risk to develop acute ovarian failure (AOF), which occurs when a patient either fails to achieve menarche by 18 years of age or stops menstruating within 5 years of their cancer diagnosis[10]. In 2006, Chemaitilly et al. estimated that 6.3% of female childhood

cancer survivors developed AOF following treatment[10]. Additionally, surgical premature menopause (SPM) can occur, where hysterectomies or bilateral oophorectomies before age 40 result in the onset of menopause.

Interventions to preserve future reproductive opportunities, such as oocyte and ovarian tissue cryopreservation, can be performed prior to or shortly following cancer treatment. However, these options may be time-sensitive, invasive, and traumatic to young girls[11-13]. In order to inform discussions of fertility preservation, the primary objective of this research was to develop a risk prediction model to estimate the individual absolute risk of a childhood cancer survivor developing NSPM following cancer treatment.

4.2 Methods

Study Population

The Childhood Cancer Survivor Study (CCSS), a multi-institutional cohort study of over 24,000 childhood cancer survivors from across North America, was the primary source of data for this project[14]. The study is composed of an original cohort of cancer survivors diagnosed between 1970 and 1986, and an expansion cohort of survivors diagnosed between 1987 and 1999. Eligibility criteria included being diagnosed and treated for leukemia, a central nervous system (CNS) malignancy, Hodgkin lymphoma, non-Hodgkin lymphoma, neuroblastoma, soft tissue sarcoma, kidney cancer, or bone cancer before age 21[14]. Participants had to be 5-year survivors from the date of their initial diagnosis[14].

Data from the female survivors in the CCSS was used to develop risk prediction models for NSPM. Follow-up surveys 1, 4 and 5 from the original cohort, and the baseline and follow-up survey from the expansion cohort provided adequate information to determine ovarian status. Of the 11,336 females who completed a baseline survey, 8,770 (77.4%) were eligible for inclusion (Table 4.1). Reasons for excluding the 2,566 participants at this stage included insufficient information to determine ovarian status (n = 1,774), survey completion through a proxy (n = 766, representing participants less than age 18 at latest follow-up survey completion or those who were deceased following the 5 year survival mark), or the absence of necessary menstrual history information (n = 26).

Individuals within the CCSS who overlapped with the St Jude Lifetime Cohort Study (SJLIFE, n=932) were set aside for external validation (Table 4.1). Further exclusions were made if participants experienced cranial radiation exposure greater than 30 Gy and/or had suspected pituitary dysfunction (n = 808), lack of/no information on age at menopause (n = 73), or a second malignancy within 5 years of the primary cancer diagnosis (n=9). Finally, 1,086 individuals were excluded due to missing treatment exposure information.

Participants were classified into 4 ovarian status categories; normal function, AOF, SPM or NSPM. Individuals diagnosed with AOF (n = 354, Table 4.1) were excluded from NSPM model development as AOF individuals are not at risk of developing NSPM. Individuals who underwent SPM cannot subsequently develop NSPM, and therefore SPM is a competing risk event for NSPM. SPM events will be treated as competing risk events during model development

and analysis. A method for graphically assessing the independence of competing events is described in Appendix E.

The total unweighted study sample was comprised of 5,508 individuals, of which 4,054 (73.6%) were designated as training data for model development, and the remaining 1,454 (26.4%) as test data for internal validation.

Table 4.1 Study sample exclusions

	Number of Observations
Total Number of CCSS Female Participants	**11,336**
Invalid or Missing Menstrual History Information	*2,566*
Data Received from CCSS	**8,770**
Overlap with External Validation Cohort	*932*
Cranial or Pituitary Radiation > 30 Gy	*808*
Missing Age at Menopause	*73*
Second Malignancy within 5 Years of Primary Cancer	*9*
Interim Total	**6,948**
Missing Key Treatment Information	1,086
Diagnosis of AOF	354
Training Data	4,054 (73.6%)
Test Data	1,454 (26.4%)
Total	**5,508**

CCSS is the Childhood Cancer Survivor Study, and AOF is acute ovarian failure

Statistical Analysis

Potential predictor variables assessed during model development included age at cancer diagnosis, age at menarche, cancer type, and treatment exposure and doses (for chemotherapy and radiation exposure). A treatment period variable (composed of 6 categories, each representing 5 consecutive years from 1970-1999) was included to assess an influence of treatment year on the risk of NSPM for the regression models.

Individuals were considered at risk for NSPM from the time they entered the study (5 years after their diagnosis date) until they either had NSPM, SPM, reached age 40, or were lost to follow-up. The self-reported age at last menstrual period was used as a proxy for the time of entering menopause, and women who survived past age 40 without entering menopause were censored at age 40. Individuals who died during the study (before reaching menopause or age 40) were censored at the age of last survey completion before their death, due to the I nability to determine ovarian status at time of death. Methods to estimate the risk of NSPM included the Fine-Gray regression (FGR) model[15,16], the time-specific logistic regression with competing risks (TLR-CR) model[17], and the random survival forest method with competing risks (RSF-CR) with a minimum node size of 10 observations[18,19].

During recruitment of participants for the expansion cohort, individuals diagnosed with the most common childhood cancer, acute lymphoblastic leukemia (ALL), were undersampled through stratified random sampling in order to allow for more intensive research on rare cancers[20]. Thus, individuals in the expansion cohort were weighted with sampling weights, defined as the inverse of the probability that they were selected for the study, in order to be representative of the target

population. Those diagnosed with ALL when they were less than 1 year of age or greater than 10 years of age were assigned a sampling weight of 1.21, and those diagnosed between age 1 and 10 were assigned a sampling weight of 3.63[21]. Sampling weights were incorporated in all reported descriptive characteristics and analyses.

A double inverse probability weight was applied when using the TLR-CR model to model the NSPM status (Appendix C) and all model accuracy assessment evaluations in order to account for both sampling design and censoring under competing risks. This weight was chosen based on our previous research into the correct combination of sampling and censoring weights for assessing model performance[22]. The censoring distribution was estimated assuming independent censoring with the Cox proportional hazards model, a decision which is examined further in the discussion.

Model accuracy was assessed using the area under the time-specific receiver operating characteristic (ROC_t) curve (AUC_t, measuring discrimination) and the time-specific average positive predictive value, which corresponds to the area under the time-specific precision-recall curve (AP_t, measuring predictive accuracy). The "APBinary" function from the <APTools> package in R was used to compute the estimates at specific follow-up times[23,24]. Performance was also assessed through calibration curves, where well calibrated models can be expected to produce reliable predictions[25]. As the models included competing risk events, calibration plots were generated by grouping observations and plotting mean predicted probabilities for NSPM against the corresponding observed cumulative incidence (Appendix D)[26]. Analysis was performed using Stata version 14.2, R version 3.4.3, and SAS version 14.1.

4.3 Results

Model Development

Of the 4,644 patients (accounting for sampling weights) in the training set with complete data, 183 developed NSPM, and 219 had SPM. At 15 years post cancer diagnosis, the event rate for NSPM was 0.017. Demographic and treatment characteristics of the survivors are presented in Table 4.2 and are similar between the training and test datasets. Model development, along with the variable effect sizes in the final models, is presented Appendix F. The final regression models included minimum ovarian radiation dose, the cyclophosphamide equivalent dose value (CED value, a measure of cumulative chemotherapy exposure), a bone marrow transplant (BMT) indicator, age at cancer diagnosis, and a treatment period variable. Interactions between clinical variables were examined but not included in order to obtain the most parsimonious models, and supported by a recent study which identified no significant interactions[8]. In addition to the treatment exposure variables in the regression models, the RSF-CR model included age at menarche, maximum abdominal and pelvic radiation doses, specific year of diagnosis (instead of treatment period), and cancer type for prediction.

Table 4.2 Characteristics of the CCSS study sample

Characteristic	CCSS Training Set n = 4,644*	CCSS Test Set n = 1,608*
	n (%)	n (%)
Age at Cancer Diagnosis		
< 5	1,958 (42.2)	683 (42.5)
5 – 9	1,023 (22.0)	327 (20.3)
10 – 14	950 (20.5)	311 (19.3)
≥ 15	713 (15.4)	287 (17.8)
Cancer Diagnosis		
Bone cancer	435 (9.4)	167 (10.4)
Central nervous system	438 (9.4)	145 (9.0)
Hodgkin lymphoma	540 (11.6)	210 (13.1)
Kidney tumors	519 (11.2)	183 (11.4)
Leukemia	1,892 (40.7)	606 (37.7)
Non-Hodgkin lymphoma	262 (5.6)	95 (5.9)
Neuroblastoma	337 (7.3)	123 (7.7)
Soft tissue sarcoma	221 (4.8)	79 (4.9)
Cyclophosphamide Equivalent Dose, mg/m^2		
None	2,446 (52.7)	873 (54.3)
<4000	801 (17.2)	286 (17.8)
4000 - 7999	550 (11.8)	169 (10.5)
≥ 8000	847 (18.2)	279 (17.4)
Ovarian Radiation Dose, Gy		
None	2,774 (59.7)	912 (56.7)
<5	1,698 (36.6)	626 (38.9)
5-9	48 (1.0)	27 (1.7)
10 - 14	85 (1.8)	30 (1.9)
15 - 19	15 (0.3)	6 (0.4)
≥ 20	24 (0.5)	7 (0.4)
Bone Marrow Transplant		
Yes	158 (3.4)	57 (3.5)
No	4,486 (96.6)	1,551 (96.5)
Treatment Period		
1970-1974	365 (7.9)	151 (9.4)
1975-1979	658 (14.2)	244 (15.2)
1980-1984	912 (19.6)	328 (20.4)
1985-1989	893 (19.3)	307 (19.1)
1990-1994	957 (20.6)	302 (18.8)
1995-1999	859 (18.5)	275 (17.1)

**Frequencies adjusted for sampling weights*

Model Evaluation

AUC_t and AP_t values were computed on both the training and test sets; estimates for 15 years post cancer diagnosis are presented in Table 4.3. On the training set, the values were similar for all models, and AUC_t estimates ranged from 0.75 to 0.77 indicating adequate performance. A similar pattern was observed for the AP_t estimates which ranged from 0.058 to 0.069, with all estimates considerably larger than the population event rate of 0.017. The NSPM event rate was 0.025 in the test set. AUC_t estimates were lower on the test set than their training set counterparts, with the largest decrease observed in the FGR model (from 0.76 to 0.59). The AP_t / Event Rate ratio increased for the TLR-CR model, implying a relative increase in predictive ability by the model. Although the AP_t point estimates increased, the AP_t / Event Rate ratio decreased from 3.41 to 3.36 for the FGR model and from 4.06 to 3.60 for the RSF-CR model.

ROC_t and PR_t curves for both the training and test set at 15 years post cancer diagnosis are provided in Figure 4.1. Visually, the ROC_t curves on the test set are less concave than those on the training set, and lie closer to the centre diagonal, indicating a reduction in the discriminatory ability of the model. In contrast, the PR_t curves on the test set show an improved positive predictive value for low true positive rates compared to the training set.

On the test set, AUC_t and AP_t values were computed every 6 months from 10 to 20 years post cancer diagnosis, and plotted in Figure 4.2. The values were not computed for less than 10 years post cancer diagnosis, as the event rate in the population was essentially 0. The performance of the three models over this time period was consistent with their point estimates at 15 years post cancer diagnosis. The AUC_t estimates for each model remained similar during the time frame,

with those from the FGR model lower than those from the TLR-CR and RSF-CR models. Throughout the entire time frame, all AP_t estimates were larger than the corresponding population event rate. The AP_t estimates for the FGR and TLR-CR models essentially identical until 14 years post cancer diagnosis, and were similar throughout the remainder of the time period. Although the AP_t values from the RSF-CR model were lower initially, they steadily increased and surpassed the other two models for the majority of the time frame.

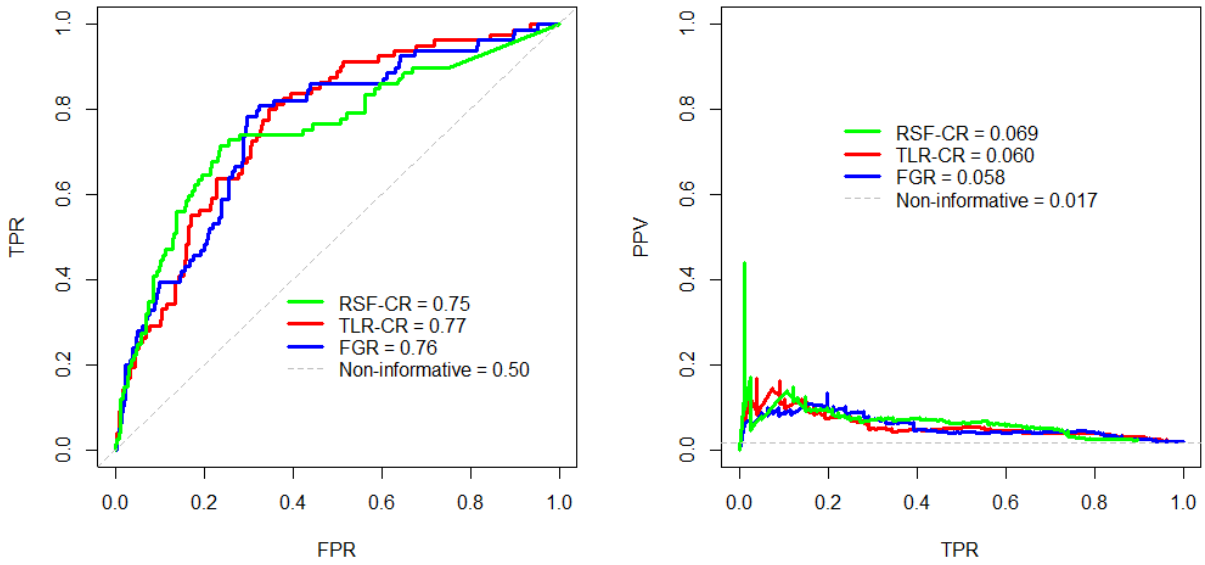
The calibration of the developed models was assessed using the calibration curves. At 15 years post cancer diagnosis, both regression models performed well for all predicted probabilities on the training set, illustrated by the blue and red lines following closely to the centre diagonal (Figure 4.3). The RSF-CR model performs well initially, however overestimates the actual observed risk for participants at a higher risk. When model calibration was assessed on the test dataset, both regression models continued to perform well at 15 years post cancer diagnosis; the TLR-CR model in particular improved its performance for larger predicted probabilities. The RSF-CR model continued to overestimate the actual observed risk. Calibration curves using the test set for 12 and 18 years post cancer diagnosis are included in Appendix G, and reflect similar conclusions made from 15 years post cancer diagnosis.

Table 4.3 Model performance and accuracy assessment values

	FGR Model	TLR-CR Model	RSF-CR Model
AUC_t Value (95% CI)			
Training Set	0.76 (0.69, 0.82)	0.77 (0.71, 0.83)	0.75 (0.64, 0.81)
Test Set	0.59 (0.49, 0.72)	0.66 (0.56, 0.77)	0.73 (0.61, 0.82)
AP_t Value (95% CI)			
Training Set	0.058 (0.040, 0.118)	0.060 (0.040, 0.129)	0.069 (0.040, 0.147)
Test Set	0.084 (0.032, 0.208)	0.099 (0.045, 0.245)	0.090 (0.051, 0.239)
AP_t / Event Rate Ratio			
Training Set *(Event rate = 0.017)*	3.41	3.53	4.06
Test Set *(Event rate = 0.025)*	3.36	3.96	3.60

Values are computed for 15 years post cancer diagnosis. FGR is the Fine-Gray regression model, TLR-CR is the time-specific logistic regression model with competing risks, RSF-CR is the random survival forest model with competing risks, AUC_t is the time-specific area under the receiver operating characteristic curve, AP_t is the time-specific average positive predictive value.

Training Set



Test Set

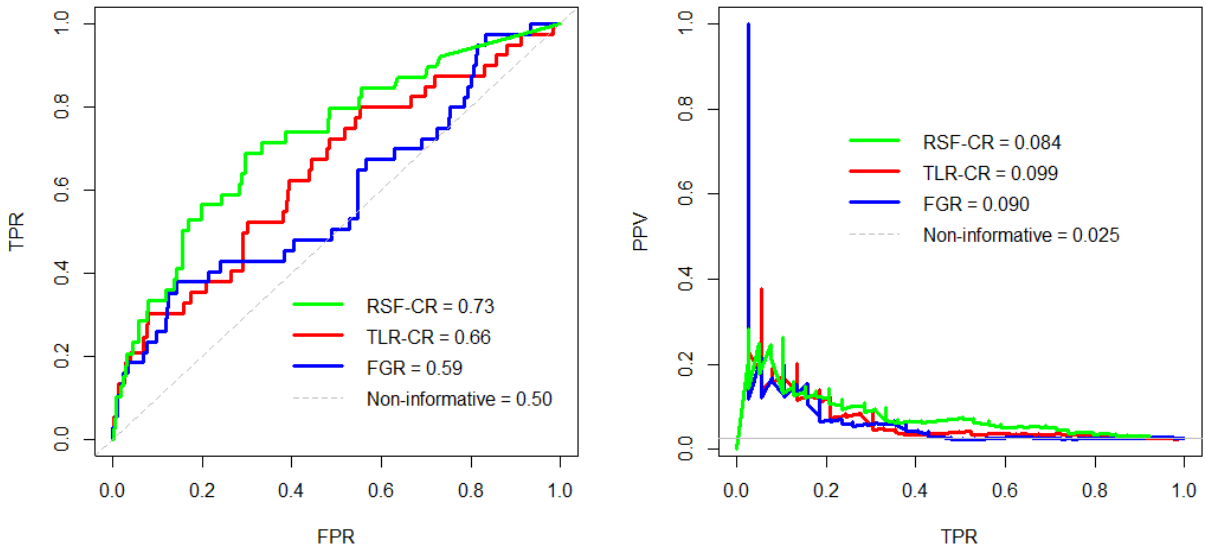


Figure 4.1 ROC_t and PR_t curves

Curves are computed for 15 years post cancer diagnosis. FGR is the Fine-Gray regression model, RSF-CR is the random survival forest model with competing risks, TLR-CR is the time-specific logistic regression model with competing risks, TPR is the true positive rate, FPR is the false positive rate and PPV is the positive predictive value.

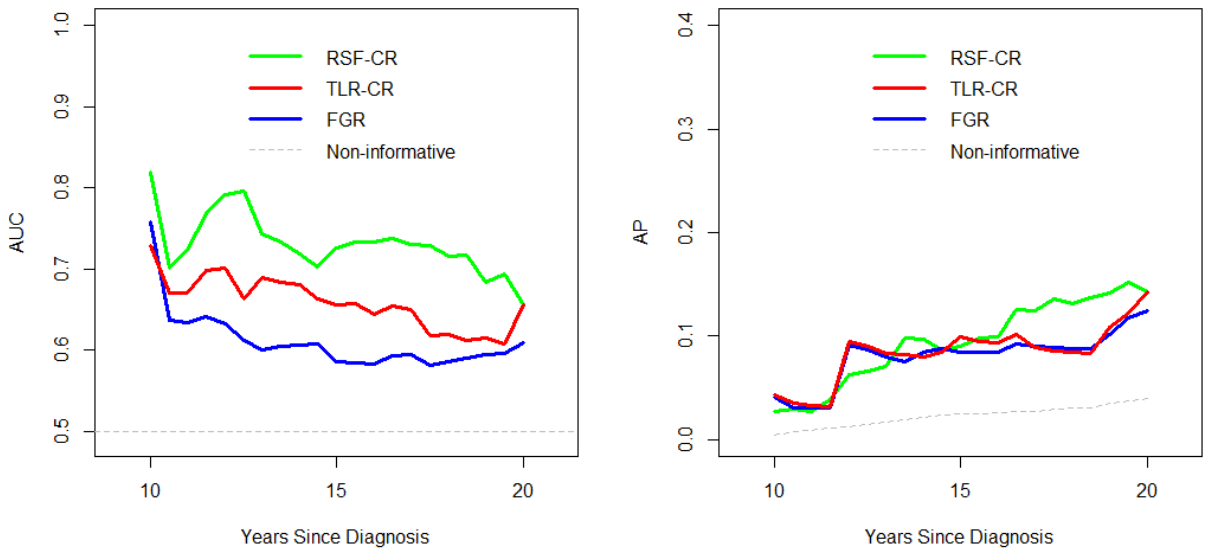


Figure 4.2 AUC_t and AP_t values over time on the test set

FGR is the Fine-Gray regression model, RSF-CR is the random survival forest model with competing risks, TLR-CR is the time-specific logistic regression model with competing risks.

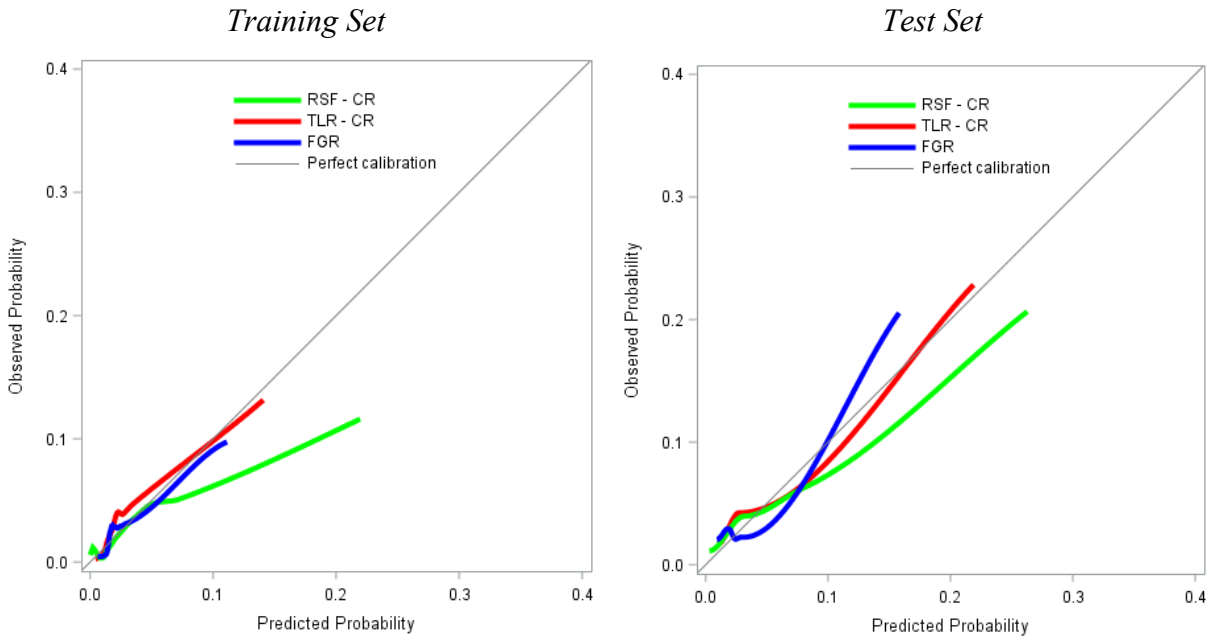


Figure 4.3 Calibration curves

Curves are computed for 15 years post cancer diagnosis. FGR is the Fine-Gray regression model, RSF-CR is the random survival forest model with competing risks, TLR-CR is the time-specific logistic regression model with competing risks.

4.4 Discussion

This research represents is the first step toward developing a risk prediction algorithm for clinicians to utilize during fertility discussions with their patients. Three risk prediction models were developed and evaluated in order to model the risk of childhood cancer survivors developing NSPM; a Fine-Gray regression model, a time-specific logistic regression model with competing risks, and a random survival forest model with competing risks.

After accounting for all assessment measures, the TLR-CR model provided a better performance than the other models. Recent evidence has shown that time-to-event model assessment using AUC_t values can be incomplete[24] and therefore, more consideration was placed on the results obtained from the AP_tvalues, AP_t / Event Rate ratios, and calibration curves when comparing model performance. Although the AUC_t value decreased for the TLR-CR model, internal validation demonstrated that it was very well calibrated. The test set event rate at 15 years post cancer diagnosis was 0.025, but the TLR-CR model was able to accurately predict the observed probability for up to 0.25. During clinical applications, this improved accuracy would provide patients at the highest risk with a more personalized prognosis than would otherwise be obtained from assessment of the average event rate at 15 years post cancer diagnosis alone. Likewise, this model serves to help reassure patients at a low risk.

While the RSF-CR model had the largest test set AUC_t value, the calibration of the model was lacking. Consistently, the RSF-CR model predicted a probability that was larger than the observed probability, particularly for high risk patients. Obtaining a risk estimate larger than the actual risk would be problematic for clinical applications, as we aim to provide reliable risk

predictions to patients, especially if the information given is to help inform decisions surrounding surgical interventions. If patients were assigned a significantly larger predicted probability compared to their actual risk, it may result in unnecessary interventions, potentially causing unintentional psychological distress, adverse health outcomes, and financial burdens from procedures. To mitigate the poor calibration issue, individuals with a predicted risk great than 0.10 could be stratified into a "high risk" patient group. Individuals would not be provided with specific prediction estimates, and this would avoid the negative implications with the lack of calibration for larger values.

Weights

As censoring can influence the estimation and evaluation of risk prediction models, weighting was required during analysis to account for missing information from those censored before their event time. Inverse probability-of-censoring weights (IPCW) were calculated by taking the inverse of the censoring distribution – the probability that an individual remains uncensored at a specific time point. Using the IPCW method to account for censoring is valid when the event and censoring processes are at least independent given the value of covariates. If censoring is assumed to be non-informative (analogous to "missing completely at random"), then the censoring distribution can be estimated using the Kaplan-Meier method. If censoring is independent (analogous to "missing at random"), the censoring distribution must be correctly specified using a model-based estimator in order to obtain consistent estimates of AUC_t and AP_t.

For model development and analysis in our analyses, weights were developed to account for both censoring of individuals during follow-up and the stratified random sampling design. The format

of the weights was chosen from our previous research into how best to combine sampling and censoring weights for obtaining unbiased accuracy assessment values[22]. Censoring weights were computed using IPCW assuming that censoring was independent. The censoring distribution was estimated using the Cox proportional hazards model, with both age at cancer diagnosis and treatment period included as predictor variables. Sampling weights to account for the stratified sampling of the expansion cohort were subsequently included multiplicatively.

Treatment Period Effect

Cancer treatment methods have advanced over the years, reflecting an increase in knowledge and technology. Changes in treatment generally involved increasing the number of individuals who received chemotherapy (but with no substantial increases in dose) and decreasing the dose and number of individuals who received radiation[21,28]. However, even after accounting for treatment changes, cancer survivorship would not have increased to the magnitude that it is seen to, implying that some unmeasured effect is contributing to survivorship[28]. Although treatment exposure doses were included in the models, treatment period remained an independent variable which significantly contributed during model development and to the model AUC_t and AP_t. All known treatment exposure changes were accounted for, therefore, the treatment period variable in our models is acting as a proxy for some other effect, such as increases in the quality of supportive care or environmental factors. Regardless, including the treatment period variable consequently poses an issue for prediction.

Prediction models are developed for application to a set of current patients. However, the categories of the treatment period variable are specifically related to the sample population and

are therefore not relevant for the new set of cancer patients whose risks are to be predicted. By excluding the treatment period variable from modelling, predictions can be made for current patients, but the aspect that treatment period is accounting for is neglected. We propose to include the treatment period variable in the final models and give a range of predicted risks for prospective patients from all six treatment periods to enhance applicability to a new population.

There is evidence that the treatment period effect may be explained by a relationship with cancer diagnoses which had large improvements in 5-year survival. Acute lymphoblastic leukemia, acute myeloid leukemia, Ewings sarcoma, medulloblastoma, neuroblastoma, non-Hodgkin lymphoma, osteosarcoma, other bone tumors and other leukemia had increases in 5-year survivorship of over 23% [29]. Higher proportions of these individuals were exposed to BMT, with the percentage of exposure increasing over time, an aspect which may explain the increase in survivorship. The remaining cancer diagnoses, including astrocytomas, Hodgkin lymphoma, kidney tumors, other CNS tumors and soft tissue sarcoma, increased over this period as well, however their increases ranged from only 10-16%[29]. When stratified analysis was performed, the significance of the treatment period effect disappeared for those diagnoses that did not have as large an increase in cancer survivorship, and remained for those individuals who did (Appendix H). Therefore, the significant treatment period effect that was found during modelling may be attributable to the large increase in survivorship that is seen in those select diagnoses.

Follow-up Length

A limitation remains with the length of follow-up time contributed by individuals diagnosed in the expansion cohort period (1987-1999). Many individuals diagnosed during this period have

not been followed for long enough to observe the development of the event. The median age at NSPM for the entire cohort was 24; however the median age at last follow-up was only 28 in the expansion cohort (compared to age 37 in the original cohort). Furthermore, during the most recent menstrual health survey collection in 2014, individuals diagnosed in 1999 in infancy or young childhood would only be 15. In the dataset, all individuals must be over 18 to have their information included, which is another limitation of case ascertainment from self-reported data. Therefore, individuals from earlier treatment periods have a greater contribution to the model estimation and evaluation. As those individuals were treated a long time ago, they may not represent the characteristics and outcomes of the expansion cohort patients diagnosed most recently, which are likely to be more closely aligned with the characteristics of current patients. With continued follow-up and data collection, further information regarding individuals from the expansion cohort will be provided and allow for the increased contribution of their data in analysis.

Conclusions

Developing risk prediction models is the first step in assessing the risk prediction of NSPM, which involved developing models using self-reported menstrual history data. Future directions may involve investigating multiple imputation to consider information from the 1,086 individuals who were excluded due to missing treatment data. The next step moving forward involves confirming the results from self-reported menstrual history data with a group of individuals with clinically verified ovarian status in order to assess the external validity of the model. The SJLIFE cohort has clinically verified ovarian status classifications for its participants, which provides the ideal platform to assess model performance in the wider

population. Ultimately, the developed models will play a role in the risk assessment of NSPM, and help families by providing additional information during their decision making process. The practical application of risk estimates will ideally have a positive impact on the quality of life of survivors well into their adulthood.

4.5 References

1. Ries LAG, Eisner MP, Kosary CL, et al. Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975-1995, National Cancer Institute, SEER Program. *National Institutes of Health Publications.* 1999:No. 99-4649.

2. Singer D, Mann E, Hunter MS, Pitkin J, Panay N. The silent grief: psychosocial aspects of premature ovarian failure. *Climacteric.* 2011;14(4):428-237.

3. Benetti-Pinto CL, de Almeida DMB, Makuch MY. Quality of life of women with premature ovarian failure. *Gynecological Endocrinology.* 2011;27(9):645-649.

4. Sklar CA. Maintenance of Ovarian Function and Risk of Premature Menopause Related to Cancer Treatment. *Journal of National Cancer Institute Monographs.* 2005;34:25-27.

5. Shuster LT, Rhodes DJ, Gostout BS, Grossardt BR, Rocca WA. Premature menopause or early menopause: Long-term health consequences. *Maturitas.* 2010;65(2):161.

6. Torrealday S, Pal L. Premature Menopause. *Endocrinology and Metabolism Clinics of North America.* 2015;44:543-557.

7. Cousineau TM, Domar AD. Psychological impact of infertility. *Best Practice & Research Clinical Obstetrics and Gynaecology.* 2007;21(2):293-308.

8. Levine JM, et al. Nonsurgical premature menopause and reproductive implications in survivors of childhood cancer: A report from the Childhood Cancer Survivor Study. *Cancer.* 2018;124(5):1044-1052.

9. Overbeek A, et al. Chemotherapy-related late adverse effects on ovarian function in female survivors of childhood and young adult cancer: A systematic review. *Cancer Treatment Reviews.* 2017;53:10-24.

10. Chemaitilly W, Mertens AC, Mitby P, et al. Acute Ovarian Failure in the Childhood Cancer Survivor Study. *The Journal of Endocrinology and Metabolism.* 2006;91(5):1723-1728.

11. Domingo J, Garcia-Velasco JA. Oocyte cryopreservation for fertility preservation in women with cancer. *Reproductive Endocrinology.* 2016;23:1-5.

12. Yu J, Huang J, Rosenwaks Z. Assisted Reproductive Techniques. In: Rosenwaks Z, Wassarman PM, eds. *Human Fertility: Methods and Protocols.* 1st ed. Humana Press; 2014: 171-231.

13. Gnaneswaran S, Deans R, Cohn RJ. Reproductive Late Effects in Female Survivors of Childhood Cancer. *Obstetrics and Gynecology International.* 2012;2012:1-7.

14. Robison LL, Mertens AC, Boice JD, et al. Study design and cohort characteristics of the Childhood Cancer Survivor Study: a multi-institutional collaborative project. *Medical and Pediatric Oncology.* 2002;38:229-239.

15. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association.* 1999;94(446):496-509.

16. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risks. *Statistics in Medicine.* 2017;36:4391-4400.

17. Zhou QM, Zheng Y, Chibnik LB, Karlson EW, Cai T. Assessing incremental value of biomarkers with multi-phase nested case-control studies. *Biometrics.* 2015;71:1139-1149.

18. Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. *Biostatistics.* 2014;15(4):757-773.

19. Cran.R-project.org. randomForestSRC: Random forests for survival, regression, and classification (RF-SRC) Accessed August 2018, from: https://cran.r-project.org/web/packages/randomForestSRC/index.html.

20. Ness KK, Hudson MM, Jones KE, et al. Effect of Temporal Changes in Therapeutic Exposure on Self-Reported Health Status in Childhood Cancer Survivors. *Annals of Internal Medicine*. 2017;166:89-98.

21. Turcotte LM, Liu Q, Yasui Y, et al. Temporal Trends in Treatment and Subsequent Neoplasm Risk Among 5-Year Survivors of Childhood Cancer, 1970-2015. *The Journal of the American Medical Association.* 2017;317(8):814-824.

22. Clark RA. Risk Prediction for Nonsurgical Premature Menopause in Childhood Cancer Survivors. [Master of Science in Epidemiology]. Edmonton, Alberta: School of Public Health, University of Alberta; 2018.

23. Cran.R-project.org. APtools: Average positive predictive values (AP) for binary outcomes and censored event times. Accessed August 2018, from: https://cran.r-project.org/web/packages/APtools/index.html.

24. Yuan Y, Zhou QM, Li B, Cai H, Chow EJ, Armstrong GT. A threshold-free summary index of prediction accuracy for censored time to event data. *Statistics in Medicine.* 2018;37(10): 1671-1681.

25. Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. *Statistics in Medicine.* 2014;33(18):3191-203.

26. Wolbers M, Koller MT, Witteman JCM, Steyerberg EW. Prognostic Models with Competing Risks: Methods and Application to Coronary Risk Prediction. *Epidemiology.*2009;20(4):555-561.

27. Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation.* 2016;133:601-609.

28. Sklar CA, Mostoufi-Moab S. Discussion regarding analysis of the Childhood Cancer Survivor Study dataset (Personal Communication). March 13th, 2018.

29. Ward E, DeSantis C, Robbins A, Kohler B, Jemal A. Childhood and Adolescent Cancer Statistics, 2014. *CA: A Cancer Journal for Clinicians.* 2014;64:83-103.

5 Conclusions

5.1 Summary

Nonsurgical premature menopause has emerged as a frequently observed chronic condition in female CCSs following treatment, and obtaining predicted risk estimates for individual patients has the potential to improve long-term quality of life. Chapter 1 highlighted the research of risk factors for the development of NSPM, which primarily include high doses of chemotherapy, treatment with procarbazine, an older age at cancer diagnosis, preparation for a bone marrow transplant, and targeted radiation to the ovaries[1,2]. Modelling the risk of NSPM involves knowledge of epidemiological and statistical techniques, which were also described in Chapter 1, and an examination of the study sample characteristics described in Chapter 2.

In Chapter 3, I assessed the appropriate combination of sampling design and censoring weights for analysis and assessment of data from cohort studies using simulation studies, which was crucial to ensure the correct implementation of weights during model development and evaluation. Weighting scenarios were assessed on four settings, which differed based on the relationship of the censoring distribution and sampling design with the event risk distribution.

Regardless of the relationship, censoring weights were required in order to obtain unbiased estimates. Under non-informative sampling, scenarios which included sampling weights multiplicatively and those that did not produced identical estimates. However, when sampling was independent, it was necessary to include sampling weights multiplicatively in order to ensure the estimates were unbiased. The study confirmed that in order to obtain consistent estimates when censoring is independent, the model for the censoring distribution must be

specified correctly. Otherwise, the resulting estimates were biased and produced large empirical standard deviations. In summary, investigators must pay close attention to the influence of sampling design and censoring distribution on the evaluation of their risk prediction models. Inadequately modelling or neglecting to include censoring weights can result in biased estimates which do not accurately represent the model performance.

In Chapter 4, three models to predict the risk of NSPM were developed and evaluated. These included a time-specific logistic regression model with competing risks, a Fine-Gray regression model, and a random survival forest model with competing risks. I presented AUC_t values, AP_t values, and calibration curves at 15 years post cancer diagnosis as an example of model performance, and assessed these measures every half year from 10 to 20 years post cancer diagnosis on the test set of data.

Assessment of all three models on the training set was similar, but following internal validation the TLR-CR model provided the best overall performance. Although the AUC_t values decreased, the performance of the model evaluated using the AP_t / Event Rate ratio and calibration increased when it was assessed on the test set. The model was able to provide reliable predictions at 15 years post cancer diagnosis for patients with risk estimates of up to 0.25, compared to an event rate in the sample of 0.025 – results which were supported when the model was assessed at 12 and 18 years post cancer diagnosis. Moving forward, external validation with the SJLIFE cohort study will help to provide an accurate representation of how the model performs in general, as participants have clinically verified ovarian status as opposed to self-reported menstrual history.

5.2 Study Limitations

Measurement error in a study is broadly divided into two categories: random error and systematic error[3]. Random error is the presence of unpredictable statistical fluctuations in the estimates of the true population values[4]. Systematic error, also known as "bias", is observed through methodical errors in the study design or how the results are interpreted[3]. Although many measures were taken to reduce the potential for bias, completely eliminating all impacts and/or accounting for all bias during analysis is unrealistic[3,5]. The CCSS is a retrospective cohort study, implying that information was collected for a past experience, and is therefore susceptible to forms of bias based on the specific study design. Model development was performed using CCSS data, implying that the biases highlighted below may have arisen during the study.

Selection bias occurs when individuals who are selected to participate in the study differ from the target population, resulting in an apparent association between exposure and outcome[3,5]. Nonresponse bias, a form of selection bias, may arise in studies when individuals who completed surveys are systematically different from individuals who were contacted but did not complete survey[5]. Therefore, the study population may not be representative of all individuals contacted. Similar to nonresponse bias is volunteer bias, where there are systematic differences between the individuals who completed the survey and the target population as a whole[5].

As information from the CCSS cohort was obtained through self-administered surveys, there exists potential for both nonresponse bias and volunteer bias. Individuals had to be willing to complete the baseline survey (as well as follow-up surveys with information about menstrual history for the original cohort participants) to be considered for inclusion in the study population.

Should the individuals who did participate have different health outcomes and treatment exposures from the individuals who did not complete the survey or from the target population overall, it could impact the magnitude and direction of any observed associations which would influence the calculation of the predicted risk estimate for each individual.

Information bias (also known as measurement bias) occurs when exposure or outcome details are recorded incorrectly[3]. In terms of survey design, each CCSS survey contained numerous pages for survivors to complete. With such a long survey, survivors may have neglected to complete the questions as accurately as possible for reasons such as fatigue, failure to follow instructions, or inattention. Questions were also different (or phrased differently) between the surveys distributed to the original and the expansion cohorts which could lead to inconsistent information collected between cohorts. For example, the follow-up 1 survey released to the original cohort requested "Age at Last Natural Menstrual Period" in years of age, whereas the expansion baseline survey asked the same question, but specifically requested the age in years and months.

Although treatment exposure data was collected from medical records, baseline information and menstrual history was self-reported, implying that recall bias, a type of information bias, may have occurred[3]. Specifically, individuals may have reported inaccurate ages for when they began or finished menstruating depending on their ovarian status, which would influence the time at risk outcome used in model development[3]. There were also many individuals excluded based on missing crucial data components. As model development was primarily restricted to individuals with complete information on the variables of interest, their exposures may not be representative of all individuals, implying a lack of generalizability of the models.

Since data was collected from patients who were diagnosed between 1970 and 1999, a considerable amount of time has passed. In fact, the earliest diagnoses in the study occurred almost 50 years ago! The treatment methods and regimes used during those times are quite different from current standards, and the results may not be generalizable to the current population, even after accounting for differences. A temporal difference in treatment methods is supported by the inclusion of the treatment period variable in model development, which remained a significant contributor to model performance, specifically when modelling was performed on diagnoses with large increases in cancer survivorship. Prediction models are developed for application on a set of current patients; however, the categories of the treatment period variable are specifically related to the sample population and are therefore not relevant for a new population. Providing a range of predicted probabilities to patients using all 6 treatment period categories is one potential solution to enhance applicability to a new population.

Predicting the risk of NSPM at specific ages is the ultimate objective of the project. However, developing models at specific ages was hindered by large observation weights resulting from the estimation of the corresponding censoring distribution. For age-specific models, a unique time at risk is computed for each individual based on their age at cancer diagnosis. For those diagnosed in infancy or early childhood, the length of time at risk before ages 30 or 40 is quite large, and the corresponding probability of remaining uncensored is low. It follows that the inverse of a low probability is a very large weight. The current models can be used to obtain age-specific predictions by calculating the length of time since cancer diagnosis before a specific age and computing the risk estimate at that length of time. Future research in this area could involve assessing methods that account for large weights when developing age-specific models.

5.3 Recommendations for Future Directions and Applications

Having information on the risk of NSPM development appropriately disseminated to clinicians is crucial for the success and durability of this project. Following external validation, the risk estimates produced by the model will be translated into a risk scoring system, which will categorize risk estimates into levels. The ultimate goal of this modelling work is to generate age-specific risk estimates that can be offered to relevant clinicians in the format of a user-friendly decision making tool.

Collaboration with developers will help to create a web and mobile application, aligning well with current trends in medical and health research output. This application will be used by physicians to obtain an estimate of the risk their patient has of developing NSPM based on their specific proposed course of treatment. The application must be designed appropriately for its purpose with a simple interface to facilitate use. The physician will input patient characteristics and the proposed treatment plan (including age of cancer diagnosis and the magnitudes of planned treatment exposures) into the application to obtain a risk estimate for specific ages following treatment.

In order for this application to be successful, it requires interest and support from the intended audience. A proposed method of rolling out the application is by having small trials at specific hospitals, ensuring the staff is properly trained to use the application. These small trials will help to identify problems and allow for adjustments before the tool is implemented more broadly. A gradual roll out can help assess clinical outcomes of using the application: Does the application aid in the decision making process? Do fertility intervention decisions change when an informed

risk estimate is provided? A future study could investigate decision making based on model prediction of NSPM, including whether the risk estimate contributed to easier decisions. Overall, it will be crucial to determine if clinicians are using the application and if patients are finding it helpful during fertility preservation discussions.

Many other factors contribute to fertility preservation decisions, including the psychological toll and financial costs of procedures. The formation of a multidisciplinary team involving fertility experts, oncologists, and psychologists is key to address the many concerns that will no doubt be brought up during this time. It would be beneficial to have counselling provided to patients to discuss the emotional stresses of surgery. Support groups may be established for families and individuals who have gone through these procedures to create a sense of community. In addition to treatment costs, there is the cost of the fertility preservation itself to consider[6]. Costs associated with the harvesting of oocyte or ovarian tissue as well as any additional storage fees (which may need to be paid on a yearly basis until used) should be conveyed to the patient and their family during discussions, as medical insurance coverage may vary[6]. Families will need to balance the cost of surgery and storage with the potential for reduced reproduction.

Although fertility preservation has been advocated for by many researchers[7], the surgical procedures for oocyte and ovarian tissue cryopreservation are still experimental and accompanied by risks and complications, such as bleeding and infection[6,8]. With ovarian tissue cryopreservation, there is the potential to reintroduce the original cancer back into the individual when the ovarian tissue is replanted[8]. Physicians may not feel comfortable with the procedure being performed on young girls and may be less likely to discuss it as an option[9]. This would

prevent patients from being informed of the NSPM risk and fertility options. For individuals with a substantial risk of developing NSPM, not only should these procedures be offered by clinicians, but they should be advocated for and accessible to those who are interested. In situations where the physician is reluctant to discuss fertility procedures, patients should have access to an alternative medical professional who will ensure the patient is fully informed of their options.

The application of a risk prediction model will substantially benefit the quality of life of CCSs by increasing the likelihood of future reproduction. By incorporating knowledge of risk estimates appropriately into patient care, clinicians and oncologists can facilitate informed discussions around the need for fertility preservation services with patients and their families. An accurate risk estimate of NSPM development following cancer treatment is the first step towards developing an important clinical tool for improving care outcomes in childhood cancer survivors.

5.4 References

1. Levine JM, et al. Nonsurgical premature menopause and reproductive implications in survivors of childhood cancer: A report from the Childhood Cancer Survivor Study. *Cancer*. 2018;124(5):1044-1052.

2. Overbeek A, et al. Chemotherapy-related late adverse effects on ovarian function in female survivors of childhood and young adult cancer: A systematic review. *Cancer Treatment Reviews*. 2017;53:10-24.

3. Rothman KJ, Greenland S, Lash TL. Chapter 9: Validity in Epidemiologic Studies. In: *Modern Epidemiology.* 3rd ed. Lippincott Williams & Wilkins; 2008:128-147.

4. Rothman KJ, Greenland S, Lash TL. Chapter 10: Precision and Statistics in Epidemiologic Studies. In: *Modern Epidemiology.* 3rd ed. Lippincott Williams & Wilkins; 2008:148-167.

5. Gordis L. *Epidemiology.* 5th ed. Elsevier Canada; 2013.

6. Levine JM, Canada A, Stern CJ. Fertility preservation in adolescents and young adults with cancer. *Journal of Clinical Oncology*. 2010;28(32):4831-4841.

7. Oktay K, Okem O. Fertility preservation medicine: A new field in the care of young cancer survivors. *Pediatric Blood and Cancer*. 2009;53:267-273.

8. Resetkova N, Hayashi M, Kolp LA, Christianson MS. Fertility Preservation for Prepubertal Girls: Update and Current Challenges. *Current Obstetrics and Gynecology Reports*. 2013; 2(4):218-225.

9. Bortoletto P, Confino R, Smith BM, Woodruff TK, Pavone ME. Practices and Attitudes Regarding Women Undergoing Fertility Preservation: A Survey of the National Physicians Cooperative. *Journal of Adolescent and Young Adult Oncology*. 2017;6(3):444-449.

References

Abir R, Ben-Aharon I, Garor R, et al. Cryopreservation of *in vitro* matured oocytes in addition to ovarian tissue freezing for fertility preservation in paediatric female cancer patients before and after cancer therapy. *Human Reproduction.* 2016;31(4):750-762.

Alberts B, Johnson A, Lewis J, et al. Eggs. In: *Molecular Biology of the Cell.* 4th ed. New York: Garland Science; 2002.

Anderson RA, Weddell A, Spoudeas HA, et al. Do doctors discuss fertility issues before they treat young patients with cancer? *Human Reproduction.* 2008;23(10):2246-2251.

Armstrong GT. Childhood Cancer Survivor Study. *NIH U.S. National Library of Medicine.* Accessed August 2018, from: https://clinicaltrials.gov/ct2/show/NCT01120353. 1995.

Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risks. *Statistics in Medicine.* 2017;36:4391-4400.

Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation.* 2016;133:601-609.

Austin PC. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. *Journal of Clinical Epidemiology.* 2010;63(1):2-6.

Barbieri RL. The Endocrinology of the Menstrual Cycle. In: Rosenwaks Z, Wassarman PM, eds. *Human Fertility: Methods and Protocols.* 1st ed. Humana Press; 2014:145-169.

Baskar R, Lee KA, Yeo R, Yeoh K. Cancer and Radiation Therapy: Current Advances and Future Directions. *International Journal of Medical Sciences.* 2012;9(3):193-199.

Bath LE, Wallace WHB, Critchley HOD. Late effects of the treatment of childhood cancer on the female reproductive system and the potential for fertility preservation. *British Journal of Obstetrics and Gynaecology.* 2002;109(2):107-114.

Benetti-Pinto CL, de Almeida DMB, Makuch MY. Quality of life of women with premature ovarian failure. *Gynecological Endocrinology.* 2011;27(9):645-649.

Bhakta N, et al. The cumulative burden of surviving childhood cancer: an initial report from the St Jude Lifetime Cohort Study (SJLIFE). *The Lancet.* 2017;390(10112):2569-2582.

Bortoletto P, Confino R, Smith BM, Woodruff TK, Pavone ME. Practices and Attitudes Regarding Women Undergoing Fertility Preservation: A Survey of the National Physicians Cooperative. *Journal of Adolescent and Young Adult Oncology*. 2017;6(3):444-449.

Boudreau C, Lawless JF. Survival Analysis Based on the Proportional Hazards Model and Survey Data. *The Canadian Journal of Statistics*. 2006;34(2):203-216.

Boyd K, Eng KH, Page CD. Area Under the Precision-Recall Curve: Point Estimates and Confidence Intervals. ECML PKDD 2013: Machine learning and knowledge discovery in databases. 2013;8190:451-466.

Burger HG, Hale GE, Robertson DM, Dennerstein L. A review of hormonal changes during the menopausal transition: focus on findings from the Melbourne Women's Midlife Health Project. *Human Reproduction Update.* 2007;13(6):559-565.

Burger HG. Physiology and endocrinology of the menopause. *Medicine.* 2006;34(1):27-30.

Cesar CC, Carvalho MS. Stratified sampling design and loss to follow-up in survival models: evaluation of efficiency and bias. *BMC Medical Research Methodology*. 2011;11:99.

Chemaitilly W, Mertens AC, Mitby P, et al. Acute Ovarian Failure in the Childhood Cancer Survivor Study. *The Journal of Endocrinology and Metabolism.* 2006;91(5):1723-1728.

Childhood Cancer Survivor Study. Expansion Cohort Baseline Survey: Long-Term Follow-Up Study of Individuals Treated for Cancer, Leukemia, Tumor or Similar Illness. Accessed August 2018, from: https://ccss.stjude.org/content/dam/en_US/shared/ccss/documents/survey/survey-baseline-exp.pdf. 2008.

Childhood Cancer Survivor Study. Original Cohort Baseline Survey: Long-Term Follow-Up Study of Individuals Treated for Cancer, Leukemia, Tumor or Similar Illness. Accessed August 2018, from: https://ccss.stjude.org/content/dam/en_US/shared/ccss/documents/survey/survey-baseline.pdf. 1992.

Childhood Cancer Survivor Study. Treatment Exposure Status: Expansion cohort as of December 2015, Original Cohort and Overall Cohort. Accessed February 2017, from https://ccss.stjude.org/content/dam/en_US/shared/ccss/documents/data/treatment-exposure-tables.pdf. 1995.

Clark RA. Risk Prediction for Nonsurgical Premature Menopause in Childhood Cancer Survivors. [Master of Science in Epidemiology]. Edmonton, Alberta: School of Public Health, University of Alberta; 2018.

Cousineau TM, Domar AD. Psychological impact of infertility. *Best Practice & Research Clinical Obstetrics and Gynaecology.* 2007;21(2):293-308.

Cox DR. Regression Models and Life-Tables. *Journal of the Royal Society Interface. Series B (Methodological)*. 1972;34(2):187-220.

Cran.R-project.org. APtools: Average positive predictive values (AP) for binary outcomes and censored event times. Accessed August 2018, from: https://cran.r-project.org/web/packages/APtools/index.html.

Cran.R-project.org. randomForestSRC: Random forests for survival, regression, and classification (RF-SRC) Accessed August 2018, from: https://cran.r-project.org/web/packages/randomForestSRC/index.html.

DeLellis Henderson K, Bernstein L, Henderson B, Kolonel L, Pike MC. Predictors of the Timing of Natural Menopause in the Multiethnic Cohort Study. *American Journal of Epidemiology.* 2008;167(11):1287-1294.

Dignam JJ, Zhang Q, Kocherginsky MN. The Use and Interpretation of Competing Risks Regression Models. *Clinical Cancer Research.* 2012;18(8):2301-2308.

Domingo J, Garcia-Velasco JA. Oocyte cryopreservation for fertility preservation in women with cancer. *Reproductive Endocrinology.* 2016;23:1-5.

Elliott MR. Bayesian weight trimming for generalized linear regression models. *Survey Methodology*. 2007;33(1):23-34.

Emmons KM, Butterfield RM, Park ER, et al. Smoking Among Participants in the Childhood Cancer Survivors Cohort: The Partnership for Health Study. *Journal of Clinical Oncology.* 2003;21(1):189-196.

Faubion SS, Kuhle CL, Shuster LT, Rocca WA. Long-term health consequences of premature or early menopause and considerations for management. *Climacteric.* 2015;18(4): 483-491.

Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association.* 1999;94(446):496-509.

Gao W, Liang J, Yan Q. Exposure to radiation therapy is associated with female reproductive health among childhood cancer survivors: a meta-analysis study. *Journal of Assisted Reproduction and Genetics.* 2015;32:1179-1186.

Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. *Statistics in Medicine.* 2014;33(18):3191-203.

Gerds TA, Cai T, Schumacher M. The Performance of Risk Prediction Models. *Biometrical Journal*. 2008;50(4):457-479.

Gibson C, Jung K. Historical Census Statistics on Population Totals by Race, 1790 to 1990, and by Hispanic origin from 1970-1990 for the United States, Regions, Divisions and States. 2002; Working Paper No. 56.

Gibson TM, Liu W, Armstrong GT, et al. Longitudinal smoking patterns in survivors of childhood cancer: An update from the Childhood Cancer Survivor Study. *Cancer.* 2015; 121(22):4035-4043.

Gnaneswaran S, Deans R, Cohn RJ. Reproductive Late Effects in Female Survivors of Childhood Cancer. *Obstetrics and Gynecology International.* 2012;2012:1-7.

Gordis L. *Epidemiology.* 5th ed. Elsevier Canada; 2013.

Grant CH, Gourley C. Chapter 2: Relevant Cancer Diagnoses, Commonly Used Chemotherapy Agents and Their Biochemical Mechanisms of Action. In: Anderson RA, Spears N, eds. *Cancer Treatment and the Ovary: Clinical and Laboratory Analysis of Ovarian Toxicity.* 1st ed. Elsevier Science; 2015.

Green DM, Kawashima T, Stovall M, et al. Fertility of Female Survivors of Childhood Cancer: A Report From the Childhood Cancer Survivor Study. *Journal of Clinical Oncology.* 2009; 27(16):2677-2685.

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. *Pediatric Blood and Cancer.* 2014;61:53-67.

Green DM, Sklar CA, Boice JD, et al. Ovarian Failure and Reproductive Outcomes After Childhood Cancer Treatment: Results from the Childhood Cancer Survivor Study. *Journal of Clinical Oncology.* 2009;27(14):2374-2381.

Hajducek DM, Lawless JF. Estimation of finite population duration distributions from longitudinal survey panels with intermittent followup. *Lifetime Data Analysis*. 2013;19:371-392.

Heagerty PJ, Lumley T, Pepe MS. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics*. 2000;56:337-344.

Hosmer DW, Lemeshow S, May S. Chapter 5: Model Development. In: *Applied Survival Analysis: Regression Modeling of Time-to-Event Data.* 2nd ed. New Jersey: John Wiley & Sons, Inc.; 2008:132.

Hosmer DW, Lemeshow S, May S. Chapter 6: Assessment of Model Accuracy. In: *Applied Survival Analysis: Regression Modeling of Time-to-Event Data.* 2nd ed. New Jersey: John Wiley & Sons, Inc.; 2008:169.

Hothorn T, Lausen B, Benner A, et al. Bagging Survival Trees. *Statistics in Medicine.* 2004; 23(1):77-91.

Hyland A, Piazza K, Hovey KM, et al. Associations between lifetime tobacco exposure with infertility and age at natural menopause: The Women's Health Initiative Observational Study. *Tobacco Control.* 2015;0:1-9.

Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. *Biostatistics.* 2014;15(4):757-773.

Ishwaran H, Kogalur UB. Consistency of Random Survival Forests. *Statistics and Probability Letters.* 2010;80:1056-1064.

James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: with Applications in R.* Springer; 2013.

Jamnongjit M, Hammes SR. Oocyte Maturation: The Coming of Age of a Germ Cell. *Seminars in Reproductive Medicine.* 2005;23(3):234–241.

Johnston RJ, Wallace WH. Normal ovarian function and assessment of ovarian reserve in the survivor of childhood cancer. *Pediatric Blood and Cancer.* 2009;53(2):296-302.

Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data.* 2nd ed. John Wiley & Sons; 2002.

Kalsbeek W, Heiss G. Building Bridges Between Populations and Samples in Epidemiological Studies. *Annual Review of Public Health*. 2000;21(1):147-169.

Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*. 1958;53(282):457-481.

Koepsell TD, Weiss NS. *Epidemiologic Methods: Studying the Occurrence of Illness.* Oxford University Press; 2003.

Ladanyi C, Mor A, Christianson MS, Dhillon N, Segars JH. Recent advances in the field of ovarian tissue cryopreservation and opportunities for research. *Journal of Assisted Reproduction and Genetics.* 2017;34(6):709-722.

Lavallée P, Beaumont J. Why We Should Put Some Weight on Weights. *Survey Methods: Insights from the Field*. 2015; Weighting: Practical Issues and 'How to' Approach (Invited Article).

Lawless J. *Statistical Models and Methods for Lifetime Data.* 2nd ed. Wiley; 2003.

LeBlanc M, Crowley J. Survival Trees by Goodness of Split. *Journal of the American Statistical Association.* 1993;88(422):456-467.

Letourneau JM, Ebbel EE, Katz PP, et al. Acute ovarian failure underestimates age-specific reproductive impairment for young women undergoing chemotherapy for cancer. *Cancer.* 2012;118(7):1933-1939.

Levine JM, Canada A, Stern CJ. Fertility Preservation in Adolescents and Young Adults with Cancer. *Journal of Clinical Oncology.* 2010;28(32):4831-4841.

Levine JM, et al. Nonsurgical premature menopause and reproductive implications in survivors of childhood cancer: A report from the Childhood Cancer Survivor Study. *Cancer*. 2018;124(5):1044-1052.

Luborsky JL, Meyer P, Sowers MF, Gold EB, Santoro N. Premature menopause in a multi-ethnic population study of the menopause transition. *Human Reproduction.* 2002;18(1):199-206.

Massarotti C, Scaruffi P, Lambertini M, Remorgida V, Del Mastro L, Anserini P. State of the art on oocyte cryopreservation in female cancer patients: A critical review of the literature. *Cancer Treatment Reviews.* 2017;57:50-57.

Merrill R. Hysterectomy Surveillance in the United States, 1997 through 2005. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research.* 2008;14(1):CR24-31.

Muka T, Oliver-Williams C, Kunuscor S, et al. Association of Age at Onset of Menopause and Time Since Onset of Menopause with Cardiovascular Outcomes, Intermediate Vascular Traits, and All-Cause Mortality: A Systematic Review and Meta-Analysis. *The Journal of the American Medical Association Cardiology.* 2016;1(7):769-776.

Nasejje JB, Mwambi H, Dheda K, Lesosky M. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Medical Research Methodology.* 2017; 17(115).

National Cancer Institute: Cancer Survivors and Smoking. Cancer Trends Progress Report. Accessed June 2017, from: https://www.progressreport.cancer.gov/after/smoking.

Ness KK, Hudson MM, Jones KE, et al. Effect of Temporal Changes in Therapeutic Exposure on Self-Reported Health Status in Childhood Cancer Survivors. *Annals of Internal Medicine*. 2017;166:89-98.

Noone AM, Howlader N, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review (CSR) 1975-2015: SEER Relative Survival (percent) by Year of Diagnosis, All Races, Males and Females, Ages 0-19.

Oeffinger KC, Mertens AC, Sklar CA, et al. Chronic Health Conditions in Adult Survivors of Childhood Cancer. *The New England Journal of Medicine.* 2006;355(15):1572-1582.

Oktay K, Okem O. Fertility preservation medicine: A new field in the care of young cancer survivors. *Pediatric Blood and Cancer*. 2009;53:267-273.

Oktem O, Kim SS, Selek U, Schatmann G, Urman B. Ovarian and Uterine Functions in Female Survivors of Childhood Cancers. *The Oncologist.* 2018;23(214-224).

Overbeek A, et al. Chemotherapy-related late adverse effects on ovarian function in female survivors of childhood and young adult cancer: A systematic review. *Cancer Treatment Reviews*. 2017;53:10-24.

Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*. 2015;68(8):855-859.

Paramanantham J, Talmor AJ, Osianlis T, Weston GC. Cryopreserved Oocytes: Update on Clinical Applications and Success Rates. *Obstetrical and Gynecological Survey*. 2015;70(2): 97-114.

Resetkova N, Hayashi M, Kolp LA, Christianson MS. Fertility Preservation for Prepubertal Girls: Update and Current Challenges. *Current Obstetrics and Gynecology Reports*. 2013; 2(4):218-225.

Ries LAG, Eisner MP, Kosary CL, et al. Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975-1995, National Cancer Institute, SEER Program. *National Institutes of Health Publications.* 1999:No. 99-4649.

Robison LL, Armstrong GT, Boice JD, et al. The Childhood Cancer Survivor Study: A National Cancer Institute-supported resource for outcome and intervention research. *Journal of Clinical Oncology.* 2009;27(14):2308-2318.

Robison LL, Hudson MM. Survivors of childhood and adolescent cancer: life-long risks and responsibilities. *Nature Reviews Cancer.* 2014;14(1):61-70.

Robison LL, Mertens AC, Boice JD, et al. Study design and cohort characteristics of the Childhood Cancer Survivor Study: a multi-institutional collaborative project. *Medical and Pediatric Oncology.* 2002;38:229-239.

Rothman KJ, Greenland S, Lash TL. Chapter 10: Precision and Statistics in Epidemiologic Studies. In: *Modern Epidemiology.* 3rd ed. Lippincott Williams & Wilkins; 2008:148-167.

Rothman KJ, Greenland S, Lash TL. Chapter 9: Validity in Epidemiologic Studies. In: *Modern Epidemiology.* 3rd ed. Lippincott Williams & Wilkins; 2008:128-147.

Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*. 2011;22(3):278–295.

Shuster LT, Rhodes DJ, Gostout BS, Grossardt BR, Rocca WA. Premature menopause or early menopause: Long-term health consequences. *Maturitas.* 2010;65(2):161.

Singer D, Mann E, Hunter MS, Pitkin J, Panay N. The silent grief: psychosocial aspects of premature ovarian failure. *Climacteric.* 2011;14(4):428-237.

Sklar CA, Mertens AC, Mitby P, et al. Premature Menopause in Survivors of Childhood Cancer: A Report From the Childhood Cancer Survivor Study. *Journal of the National Cancer Institute.* 2006;98(13):890-896.

Sklar CA, Mostoufi-Moab S. Discussion regarding analysis of the Childhood Cancer Survivor Study dataset (Personal Communication). March 13th, 2018.

Sklar CA. Maintenance of Ovarian Function and Risk of Premature Menopause Related to Cancer Treatment. *Journal of National Cancer Institute Monographs.* 2005;34:25-27.

Spittal MJ, Carlin JB, Currier D, et al. The Australian Longitudinal Study on Male Health sampling design and survey weighting: implications for analysis and interpretation of clustered data. *BMC Public Health*. 2016;16(Suppl 3):1062.

Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology*. 2010;21(1):128-138.

Sun L, Tan L, Yang F, et al. Meta-analysis suggests that smoking is associated with an increased risk of early natural menopause. *Menopause: The Journal of The North American Menopause Society.* 2012;19(2):126-132.

Therneau T, Grambsch P, eds. *Modeling Survival Data: Extending the Cox Model.* 1st ed. Springer; 2000.

Thomas-Teinturier C, et al. Ovarian reserve after treatment with alkylating agents during childhood. *Human Reproduction.* 2014;30(6):1437-1446.

Torrealday S, Pal L. Premature Menopause. *Endocrinology and Metabolism Clinics of North America.* 2015;44:543-557.

Turcotte LM, Liu Q, Yasui Y, et al. Temporal Trends in Treatment and Subsequent Neoplasm Risk Among 5-Year Survivors of Childhood Cancer, 1970-2015. *The Journal of the American Medical Association.* 2017;317(8):814-824.

United Nations. Chapter 6: Construction and use of sample weights. In: *Designing Household Survey Samples: Practical Guidelines.* New York: United Nations Publication; 2008:109.

Vern-Gross TZ, Bradley JA, Rotondo RL, Indelicato DJ. Fertility in childhood cancer survivors following cranial irradiation for primary central nervous system and skull base tumors. *Radiotherapy and Oncology.* 2015;117:195-205.

Ward E, DeSantis C, Robbins A, Kohler B, Jemal A. Childhood and Adolescent Cancer Statistics, 2014. *CA: A Cancer Journal for Clinicians.* 2014;64:83-103.

Willems S, Schat A, van Noorden M, Fiocco M. Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. *Statistical Methods in Medical Research*. 2018;27(2):323-335.

Wolbers M, Koller MT, Witteman JCM, Steyerberg EW. Prognostic Models with Competing Risks: Methods and Application to Coronary Risk Prediction. *Epidemiology.*2009;20(4):555-561.

Yu J, Huang J, Rosenwaks Z. Assisted Reproductive Techniques. In: Rosenwaks Z, Wassarman PM, eds. *Human Fertility: Methods and Protocols.* 1st ed. Humana Press; 2014: 171-231.

Yuan Y, Zhou QM, Li B, Cai H, Chow EJ, Armstrong GT. A threshold-free summary index of prediction accuracy for censored time to event data. *Statistics in Medicine.* 2018;37(10): 1671-1681.

Zhou QM, Zheng Y, Chibnik LB, Karlson EW, Cai T. Assessing incremental value of biomarkers with multi-phase nested case-control studies. *Biometrics.* 2015;71:1139-1149.

Zhou X, Obuchowski NA, McClish DK. Chapter 2: Measures of Diagnostic Accuracy. In: *Statistical Methods in Diagnostic Medicine.* 2nd ed. John Wiley & Sons, Inc.; 2011:13-56.

Zhou Y, McArdle J. Rationale and Applications of Survival Tree and Survival Ensemble Methods. *Psychometrika.* 2015;80(3):811-833.

Appendices

Appendix A Menstrual history survey questions

Ovarian status classifications for female original cohort participants were derived using variables from follow-up 1 (items 19-19d), follow-up 4 (items F13-F16, J33-J34) and the follow-up 5 questionnaire (items G13-G16, J35-J36). Ovarian status classifications for female expansion cohort participants were derived using variables from the expansion baseline questionnaire (items E13 –E16, I33-I34) and the follow-up 5 questionnaire. Copies of the surveys administered to the CCSS participants were obtained from the Childhood Cancer Survivor Study website (https://ccss.stjude.org/tools-and-documents/questionnaires/baseline-and-follow-up-questionnaires.html) and the specific questions pertaining to ovarian status classification are included below.

Menstrual History – for females 18 years or older

The following questions pertain to your menstrual history. Previously we asked a few questions about your menstrual periods. Now we wish to obtain more detailed information. This will help us understand how past treatments affect a woman's pattern of menstruation and the timing of her menopause.

19. Have you ever had a menstrual period naturally; that is, without needing hormones or medication?

- ○ Yes → Go to Question 19a
- ○ No → Skip to Question 20
- ○ Not sure → Go to Question 19a

19a. At what age did you have your first menstrual period?

_____ **years old**

19b. At what age did you last have a menstrual period naturally, without needing medication or hormones to bring it on?

_____ **years old**

19c. Which of the following statements best describes you? (Select only ***one***)

- a. ○ I am having regular periods and I **am not** taking birth control pills or female hormones *(example: Premarin, estrogen).*
- b. ○ I am having regular periods but **I am** using birth control pills to prevent a pregnancy.
- c. ○ My menstrual periods are irregular and I **am** taking birth control pills or female hormones to regulate my periods.
- d. ○ I am currently pregnant.
- e. ○ I am *not* having menstrual periods naturally but I **am** taking birth control pills or female hormones.
- f. ○ I am *not* having menstrual periods naturally and I **am not** taking birth control pills or female hormones.
- g. ○ Other, please specify:

If you selected a, b, c, or d, please go to Question 20.
If you selected e, f, or g, please go to Question 19d.

19d. What caused your menstrual periods to stop? (Select only ***one***)

- ○ Normal or early menopause
- ○ Surgery *(example: a hysterectomy)*
- ○ Pregnancy
- ○ Other, please specify:

Figure A1 Follow-up 1 survey (2000)

F13. **FEMALES** - Have you had a menstrual period naturally, that is, without needing hormones or medication?

☐ No ☐ Yes **If yes, age at first occurrence:** ☐☐

If no, ⟶ Go to Question F15.

F14. **FEMALES** - At what age did you last have a menstrual period naturally, without needing hormones or medication?

☐☐ **years and** ☐☐ **months old**

F15. **FEMALES** - Which one of the following statements best describes you? **(Select only one)**

- ☐ a. I am having regular periods and I am not taking birth control pills or female hormones (example: Premarin, estrogen)
- ☐ b. I am having regular periods but I am using birth control pills to prevent a pregnancy
- ☐ c. My menstrual periods are irregular and I am taking birth control pills or female hormones to regulate my periods
- ☐ d. I am currently pregnant
- ☐ e. I am not having menstrual periods naturally but I am taking birth control pills or female hormones
- ☐ f. I am not having menstrual periods naturally and I am not taking birth control pills or female hormones
- ☐ g. Other

If Other, please describe.

If you selected a, b, c, or d ⟶ Go to Question G1.
If you selected e, f, or g ⟶ Go to Question F16.

F16. **FEMALES** - What caused your menstrual periods to stop? **(Select only one)**

- ☐ Normal or early menopause
- ☐ Surgery (example: a hysterectomy)
- ☐ Pregnancy
- ☐ Don't know
- ☐ Other

If Other, please describe.

Please indicate if you have ever had any of the following surgical procedures done.

	No	Yes	Not sure	If yes, age at first occurrence (years)
J23. Any lung surgery?	☐	☐	☐	☐☐
If yes, specify.				
J24. Periodontal (gum) surgery?	☐	☐	☐	☐☐
J25. Heart transplant?	☐	☐	☐	☐☐
J26. Lung transplant?	☐	☐	☐	☐☐
J27. Kidney transplant?	☐	☐	☐	☐☐
J28. Liver transplant?	☐	☐	☐	☐☐
J29. Bone marrow transplant?	☐	☐	☐	☐☐
J30. Other organ transplant?	☐	☐	☐	☐☐
If yes, specify transplant.				
J31. Cataract surgery?	☐	☐	☐	☐☐
Males ⟶ Go to Question J35.				
J32. Removal of one ovary?	☐	☐	☐	☐☐
J33. Removal of both ovaries?	☐	☐	☐	☐☐
J34. Removal of uterus?	☐	☐	☐	☐☐
Females ⟶ Go to Question J37.				

Figure A2 Follow-up 4 survey (2007)

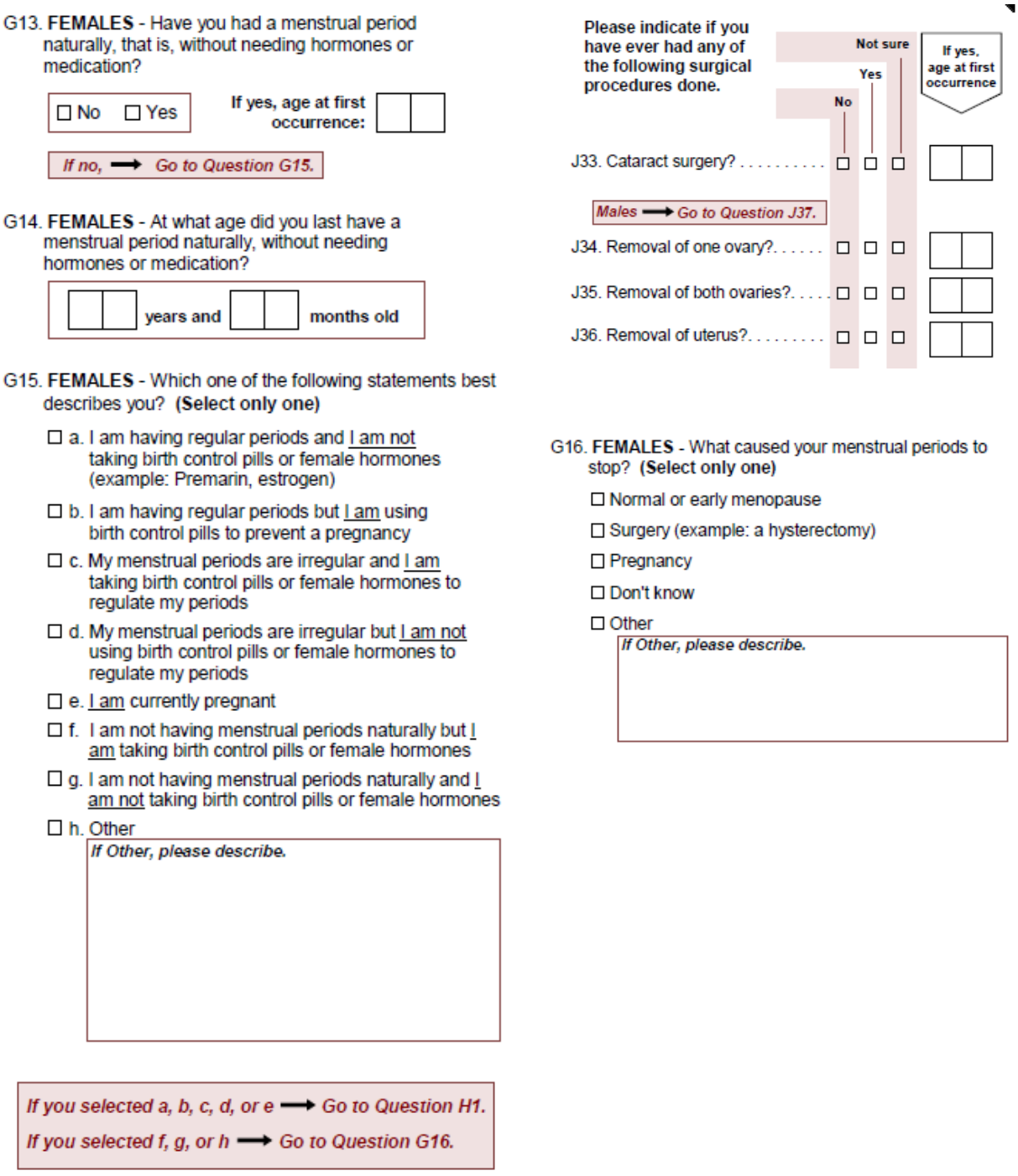


Figure A3 Follow-up 5 survey (2014)

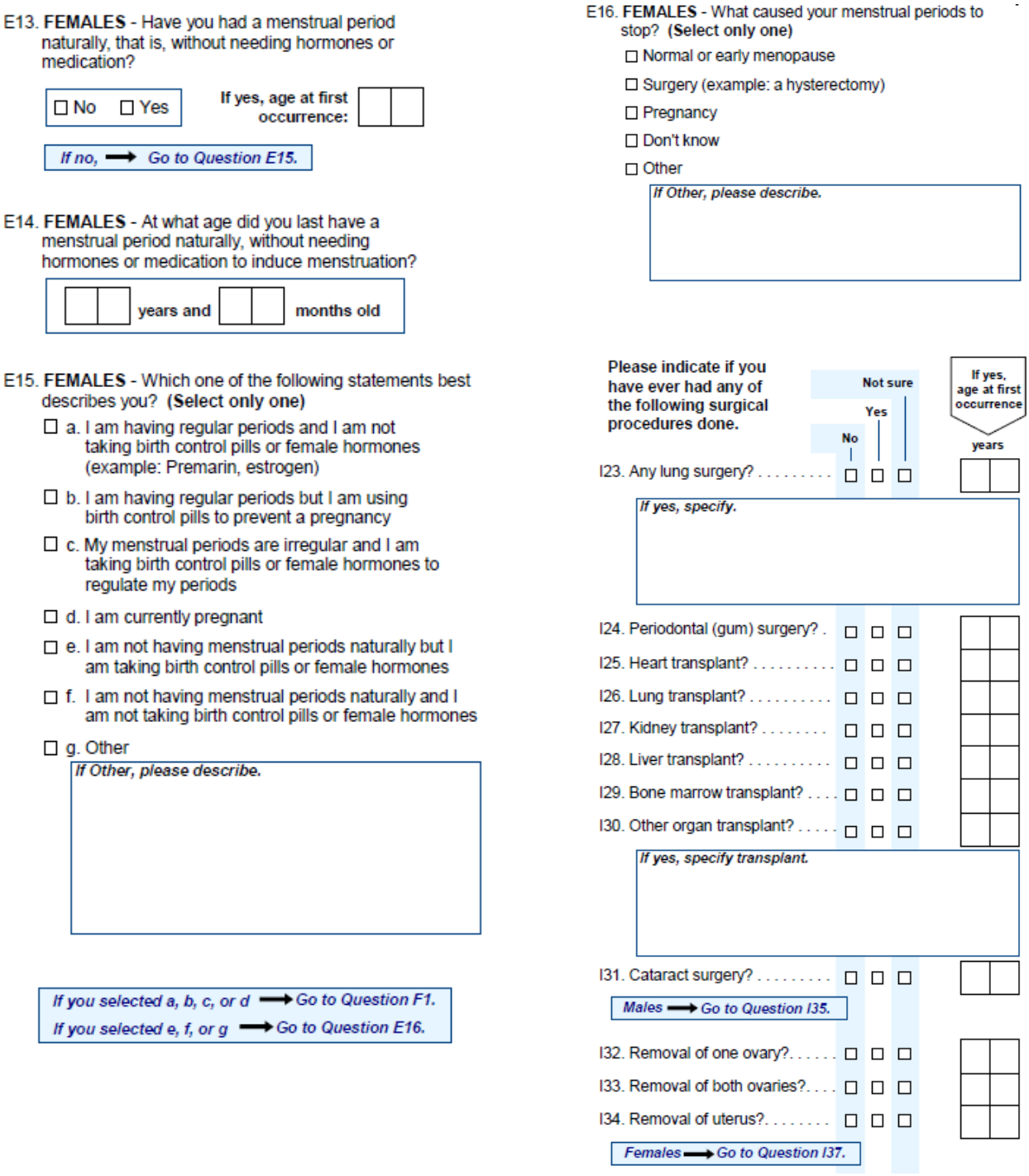


Figure A4 Expansion cohort baseline survey (2008 – present)

Appendix B Simulation study coefficient estimates

There were no significant differences in the estimates of the coefficient from the Cox PH model for weighted and unweighted samples as shown in Table B1, and all estimates were not significantly different from the true coefficient value of 1.115. In each study, the weighted sample produced larger standard deviations for the coefficient estimate than the unweighted sample, illustrating the bias-variance trade-off. In settings i) and ii), where the censoring distribution was non-informative, coefficient estimates had slightly non-significant differences between the weighted and unweighted samples. In settings iii) and iv), with an independent relationship between the risk score and censoring distributions, coefficient estimates were identical between the weighted and unweighted samples.

Table B1 Mean coefficient estimates

	Truth	**Setting i)**	**Setting ii)**	**Setting iii)**	**Setting iv)**
	β	Mean $\hat{\beta}$ (SD)	Mean $\hat{\beta}$ (SD)	Mean $\hat{\beta}$ (SD)	Mean $\hat{\beta}$ (SD)
Unweighted	1.115	1.111 (0.028)	1.119 (0.034)	1.121 (0.037)	1.118 (0.043)
Weighted		1.113 (0.039)	1.125 (0.042)	1.121 (0.050)	1.118 (0.060)

Appendix C Inverse probability-of-censoring weight calculations

Notation

Let:

t_0 = time point of interest

X_i = event time for the ith individual

δ_i = censoring indicator (0 is censored, 1 is the event, 2 is the competing risk event)

$G(\cdot)$ = the estimated survivor function of the censoring distribution (computed treating all competing events the same as if they were events of interest)

$\hat{c}_{t_0,i}$= the estimated weight at time t_0 for the ith individual

The inverse probability-of-censoring weights (IPCW) are calculated as follows:

$$\hat{c}_{t_0,i} = \frac{I(X_i < t_0)I(\delta_i = 1)}{\hat{G}(X_i)} + \frac{I(X_i \geq t_0)}{\hat{G}(t_0)}$$

Table C1 IPCW for survival models with competing risks

Situation	Time at event	δ_i	Status at t_0	Weight at t_0
1	$X_1 < t_0$	0	unknown	0
2	$X_2 < t_0$	1 or 2	1 or 2	$\frac{1}{\hat{G}(X_2)}$
3	$X_3 \geq t_0$	0, 1 or 2	0	$\frac{1}{\hat{G}(t_0)}$

X_i is the observed survival time for the ith individual δ_i, is the event indicator, t_0 is the time point of interest, and $\hat{G}(\cdot)$ is an estimate of the censoring distribution.

Table C2 IPCW for time-specific logistic regression with competing risks

Situation	Time at event	δ_i	Status at t_0	Weight at t_0
1	$X_1 < t_0$	0 or 2	unknown	0
2	$X_2 < t_0$	1	1	$\frac{1}{\hat{G}(X_2)}$
3	$X_3 \geq t_0$	0, 1 or 2	0	$\frac{1}{\hat{G}(t_0)}$

X_i is the observed survival time for the ith individual δ_i, is the event indicator, t_0 is the time point of interest, and $\hat{G}(\cdot)$ is an estimate of the censoring distribution.

Appendix D Computing calibration curves for competing risk prediction models

Step 1: Generate predicted probabilities at the time point of interest $t = t_0$ using a competing risk model

Step 2: Rank the observations based on their predicted probability values and divide into n groups

Step 3: Within each group, calculate the average predicted risk at t_0

Step 4: Within each group, calculate the cumulative incidence function (which accounts for competing risks) at t_0

For each group, obtain the observed and predicted probabilities:

Group	**Observed**	**Predicted**
Group 1	Cumulative Incidence at t_0 (from individuals in Group 1)	Mean predicted probabilities (from individuals in Group 1)
Group 2	Cumulative Incidence at t_0 (from individuals in Group 2)	Mean predicted probabilities (from individuals in Group 2)
. . .	. . .	. . .
Group n	Cumulative Incidence at t_0 (from individuals in Group n)	Mean predicted probabilities (from individuals in Group n)

Step 5: Compute a lowess curve of observed versus predicted probabilities

Appendix E Checking independence of competing risk events

Suppose T_1 and T_2 are two statistically independent processes, both subject to a non-informative censoring process, C.

Let $F_1(t)$ denote $P(T_1 < t)$, and $F_2(t)$ denote $P(T_2 < t)$.

$$X = \min(T_1, T_2), \quad \delta = \begin{cases} 0 & XX = C \\ 1 & XX = T_1 \\ 2 & XX = T_2 \end{cases}$$

$$XX = \min(X, C)$$

Then:

$$F_X(t) = P(X < t) = P(\min(T_1, T_2) < t) \tag{1}$$

$$\stackrel{\text{def}}{=} \text{CIF}_1(t) + \text{CIF}_2(t) \tag{2}$$

$$= P(T_1 < t, T_2 \geq t) + P(T_1 \leq T_2 < t) + P(T_2 < t, T_1 \geq t) + P(T_2 \leq T_1 < t) \tag{3}$$

$$= P(T_1 < t, T_2 \geq t) + P(T_2 < t, T_1 \geq t) + P(T_1 < t, T_2 < t) \tag{4}$$

Using the independence assumption

$$= F_1(t)S_2(t) + F_2(t)S_1(t) + F_1(t)F_2(t) \tag{5}$$

$$= F_1(t) + F_2(t) - F_1(t)F_2(t) \tag{6}$$

This provides a graphical check for the independence of T_1 and T_2.

Appendix F Model analysis

1. Time-specific Logistic Regression with Competing Risks (at 15 years post diagnosis)

Assuming independent censoring and sampling, the observation weights for logistic regression are given by:

$$IPCW_{Cox}^{1} \times p_j$$

The censoring distribution, $\hat{G}(t)$, is estimated using Cox proportional hazards regression with treatment period and age at diagnosis as covariates.

Table F1 TLR-CR univariate analysis

Variable	Coefficient	Odds Ratio	p-value
Age at Cancer Diagnosis	0.050	1.051	0.008
BMT Exposure	1.996	7.360	<0.001
CED Value	0.015	1.015	0.326
Hispanic Origin	0.737	2.090	0.026
Minimum Ovarian RT Dose	0.101	1.106	<0.001
Race			
Black	-0.509	0.601	0.487
Asian or Pacific Islander	0.620	1.859	0.405
American Indian or Alaskan Native	1.453	4.276	0.176
Smoked at least 100 cigarettes	-0.013	0.987	0.973
Treatment Period			
1975 – 1979	0.101	1.106	0.887
1980 – 1984	0.046	1.047	0.946
1985 – 1989	0.460	1.584	0.513
1990 – 1994	0.991	2.694	0.119
1995 – 1999	1.418	4.129	0.027

BMT exposure = No is the reference category for the BMT exposure variable, Hispanic origin = No is the reference category for the Hispanic origin variable, White is the reference category for the race variable, smoked at least 100 cigarettes = No is the reference category for the "smoked at least 100 cigarettes" variable, and 1970-1974 is the reference category for the treatment period variable.

Race Overall Significance:

Wald Test:

Wald $chi^2(3) = 3.06$

p-value = 0.3821

Treatment Period Overall Significance:

Wald Test:

Wald $chi^2(5) = 15.44$

p-value = 0.0086

TLR-CR Univariate Analysis Summary:

Age at cancer diagnosis, BMT exposure, Hispanic origin, minimum ovarian RT dose, and treatment period have p-values less than 0.20 and are retained for multivariate regression. CED value will be included in multivariate regression as it is a biologically important variable, and clinicians have recommended its inclusion. The race and "smoked at least 100 cigarettes" variables were not significant in the analysis (p-values > 0.20) and therefore are not retained for multivariate analysis.

Table F2 TLR-CR intermediate multivariate analysis

Variable	Coefficient	Odds Ratio	p-value
Age at Cancer Diagnosis	0.054	1.055	0.007
BMT Exposure	1.395	4.035	0.002
CED Value	0.001	1.001	0.963
Hispanic Origin	0.516	1.675	0.148
Minimum Ovarian RT Dose	0.093	1.097	0.001
Treatment Period			
1975 – 1979	0.207	1.230	0.776
1980 – 1984	0.182	1.200	0.793
1985 – 1989	0.478	1.613	0.491
1990 – 1994	1.044	2.841	0.119
1995 – 1999	1.316	3.728	0.057

BMT exposure = No is the reference category for the BMT exposure variable, Hispanic origin = No is the reference category for the Hispanic origin variable, and 1970-1974 is the reference category for the treatment period variable.

Treatment Period Overall Significance:

Wald Test:

Wald $chi^2(5) = 9.44$

p-value = 0.0928

TLR-CR Intermediate Multivariate Analysis Summary:

After accounting for the other variables in the model, age at cancer diagnosis, BMT exposure, and minimum ovarian RT dose are significant with p-values less than 0.05. CED value will continue to be included as a biologically important variable, even though the associated p-value is larger than 0.05 (p-value = 0.972). Both the treatment period variable and the Hispanic origin variable are non-significant. As model prediction and performance was improved when the treatment period variable was included in the model, it is retained in the multivariate regression model.

Table F3 TLR-CR multivariate analysis

Variable	Coefficient	Odds Ratio	p-value
Age at Cancer Diagnosis	0.054	1.055	0.007
BMT Exposure	1.438	4.212	<0.0001
CED Value	-0.001	0.999	0.972
Minimum Ovarian RT Dose	0.094	1.099	<0.0001
Treatment Period			
1975 – 1979	0.247	1.280	0.730
1980 – 1984	0.216	1.241	0.754
1985 – 1989	0.557	1.745	0.403
1990 – 1994	1.140	3.127	0.073
1995 – 1999	1.434	4.195	0.022

BMT exposure = No is the reference category for the BMT exposure variable, and 1970-1974 is the reference category for the treatment period variable.

Treatment Period Overall Significance:

Wald Test:

Wald $chi^2(5) = 11.32$

p-value = 0.0453

TLR-CR Multivariate Analysis Summary:

After accounting for the other variables in the model, age at cancer diagnosis, BMT exposure, minimum ovarian RT dose, and treatment period are significant with p-values less than 0.05. CED value will continue to be included as a biologically important variable, even though the associated p-value is larger than 0.05 (p-value = 0.972).

2. Fine-Gray Regression

Assuming independent sampling, sampling weights are included as observation weights during model development (p_j). Censoring is inherently accounted for during model development with Fine-Gray regression.

Table F4 FGR univariate analysis

Variable	Coefficient	Subdistribution Hazard Ratio	p-value
Age at Cancer Diagnosis	0.043	1.044	0.004
BMT Exposure	1.738	5.686	<0.001
CED Value	0.022	1.022	0.006
Hispanic Origin	0.569	1.766	0.031
Minimum Ovarian RT Dose	0.087	1.091	<0.001
Race			
Black	0.118	1.125	0.750
Asian or Pacific Islander	0.783	2.188	0.085
American Indian or Alaskan Native	0.638	1.893	0.533
Smoked at least 100 cigarettes	0.070	1.073	0.754
Treatment Period			
1975 – 1979	-0.414	0.661	0.172
1980 – 1984	-0.528	0.590	0.073
1985 – 1989	0.082	1.085	0.785
1990 – 1994	0.248	1.281	0.397
1995 – 1999	0.359	1.432	0.290

BMT exposure = No is the reference category for the BMT exposure variable, Hispanic origin = No is the reference category for the Hispanic origin variable, White is the reference category for the race variable, smoked at least 100 cigarettes = No is the reference category for the "smoked at least 100 cigarettes" variable, and 1970-1974 is the reference category for the treatment period variable.

Treatment Period Overall Significance:

Wald Test:

Wald $chi^2(5) = 12.92$

p-value = 0.0242

Race Overall Significance:

Wald Test:

Wald $chi^2(3) = 3.32$

p-value = 0.3444

FGR Univariate Summary:

All variables except for the "smoked at least 100 cigarettes" variable and the race variable have p-values less than 0.20, and are therefore retained for multivariate regression. As the p-values for the "smoked at least 100 cigarettes" variable and the race variable are non-significant (p-values = 0.754 and 0.3444 respectively), they will not be retained in the multivariate analysis.

Table F5 FGR intermediate multivariate analysis

Variable	Coefficient	Subdistribution Hazard Ratio	p-value
Age at Cancer Diagnosis	0.046	1.047	0.002
BMT Exposure	1.296	3.655	<0.001
CED Value	0.014	1.014	0.143
Hispanic Origin	0.472	1.603	0.088
Minimum Ovarian RT Dose	0.077	1.080	<0.001
Treatment Period			
1975 – 1979	-0.293	0.746	0.347
1980 – 1984	-0.457	0.633	0.137
1985 – 1989	-0.003	0.997	0.991
1990 – 1994	0.298	1.347	0.342
1995 – 1999	0.255	1.290	0.499

BMT exposure = No is the reference category for the BMT exposure variable, Hispanic origin = No is the reference category for the Hispanic origin variable, and 1970-1974 is the reference category for the treatment period variable.

Treatment Period Overall Significance:

Wald Test:

Wald $chi^2(5) = 8.28$

p-value = 0.1412

FGR Intermediate Multivariate Analysis Summary:

After accounting for the other variables in the model, age at cancer diagnosis, BMT exposure, and minimum ovarian radiation dose are significant, and will be retained. Both the Hispanic origin variable and the treatment period variable are non-significant. CED value will continue to be included as a biologically important variable, even though the associated p-value is larger than 0.05 (p-value = 0.143). As model prediction and performance was improved when the treatment period variable was included in the model, it is retained in the multivariate regression model.

Table F6 FGR multivariate analysis

Variable	Coefficient	Subdistribution Hazard Ratio	p-value
Age at Cancer Diagnosis	0.045	1.046	0.003
BMT Exposure	1.272	3.568	<0.001
CED Value	0.016	1.016	0.074
Minimum Ovarian RT Dose	0.075	1.078	<0.001
Treatment Period			
1975 – 1979	-0.315	0.730	0.305
1980 – 1984	-0.400	0.670	0.178
1985 – 1989	0.140	1.150	0.643
1990 – 1994	0.333	1.395	0.275
1995 – 1999	0.318	1.374	0.378

BMT exposure = No is the reference category for the BMT exposure variable, and 1970-1974 is the reference category for the treatment period variable.

Treatment Period Overall Significance:

Wald Test:

Wald $chi^2(5) = 9.86$

p-value = 0.0793

FGR Multivariate Analysis Summary:

After accounting for the other variables in the model, minimum ovarian RT dose, BMT exposure, and age at cancer diagnosis are significant with p-values less than 0.05. CED value will continue to be included as a biologically important variable, even though the p-value is slightly larger than 0.05 (p-value = 0.074).

Using the Wald test, the treatment period variable was only marginally significant. However, when the variable was excluded, there was slight confounding with BMT exposure observed. Additionally, model prediction and performance was improved when the treatment period variable was included in the model, and therefore it is retained in the multivariate regression model.

3. Random Survival Forest with Competing Risks

Table F7 RSF-CR variable importance

Variables	CV Set 1		CV Set 2		CV Set 3		CV Set 4		CV Set 5		Entire Training Set	
	Rank	Value	Rank	Value	Rank	Value	Rank	Value	Rank	Value	Rank	Value
Procarbazine Dose	1	0.0411	1	0.0368	2	0.0245	2	0.0341	1	0.0377	1	0.0381
Age at Cancer Diagnosis	2	0.0372	5	0.0156	6	0.0061	5	0.0103	5	0.0133	7	0.0123
BMT Exposure	3	0.0229	2	0.0343	3	0.0226	3	0.0282	2	0.0199	2	0.0322
Age at Menarche	4	0.0156	6	0.0109	5	0.0091	6	0.0069	7	0.0043	5	0.0149
Year of Cancer Diagnosis	5	0.0098	3	0.0301	4	0.0146	1	0.0368	6	0.0110	3	0.0276
Maximum Abdomen RT Dose	6	0.0034	10	-0.0022	9	-0.0032	7	0.0046	9	-0.0028	9	0.0037
CED Value	7	0.0025	8	0.0049	10	-0.0042	8	0.0026	10	-0.0097	8	0.0088
Minimum Ovarian RT Dose	8	0.0022	4	0.0183	7	0.0050	4	0.0193	3	0.0186	4	0.0169
Cancer Diagnosis Category	9	0.000	7	0.0090	8	0.0005	9	0.0023	8	0.0017	10	-0.0001
Maximum Pelvic RT Dose	10	-0.0039	9	-0.0021	1	0.0267	10	0.0022	4	0.0134	6	0.0126

CV is cross validation and 'Value' represents the variable importance (VIMP), calculated by comparing the prediction performance for a variable which has been permuted to the original prediction performance.

Appendix G Test set calibration curves for 12 and 18 years post cancer diagnosis

Figure G1 Calibration curves for 12 years post cancer diagnosis

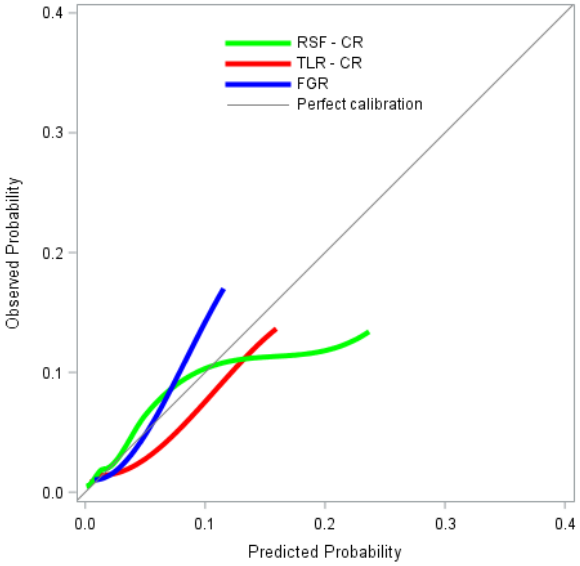
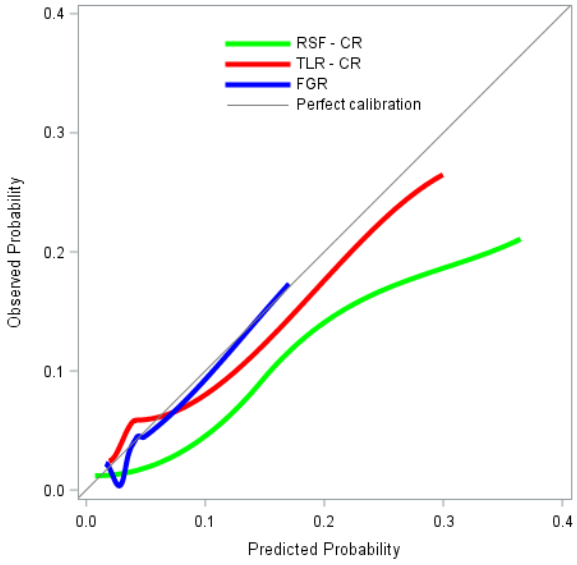


Figure G2 Calibration curves for 18 years post cancer diagnosis



Appendix H Examining cancer diagnoses with improved survival

Cancer diagnoses with large increases in survivorship (defined as Group 2) were characterized by an increase in 5-year survival of greater than 23%, based on information from "Ward E, et al., Childhood and Adolescent Cancer Statistics, 2014. CA Cancer J Clin 2014; 64:83-103". Cancer diagnoses included in Group 1 had increases in 5-year survival of 16% or less.

Table H1 Cancer diagnoses with and without large increases in survivorship

Group 1	Group 2
Astrocytomas (16%)	Acute lymphoblastic leukemia (33%)
Hodgkin lymphoma (10%)	Acute myeloid leukemia (43%)
Kidney tumors (15%)	Ewings sarcoma (30%)
Other CNS tumors (16%)	Medulloblastoma (23%)
Soft tissue sarcoma (15%)	Neuroblastoma (25%)
	Non-Hodgkin lymphoma (38%)
	Osteosarcoma (26%)
	Other bone tumors (24%)
	Other leukemia (36%)
Number of Observations = 1,707	Number of Observations = 2,347

CNS is central nervous system; percent in brackets represents the absolute increase in 5-year survival percent from 1975-1979 to 2003-2009

Table H2 Stratified time-specific logistic regression with competing risks output

Variable	*Group 1* **OR (p-value)**	*Group 2* **OR (p-value)**
Minimum Ovarian RT Dose	1.127 (<0.001)	1.047 (0.408)
BMT Exposure	1 (omitted)	6.601 (0.001)
CED Value	1.003 (0.878)	0.998 (0.924)
Age at Cancer Diagnosis	1.077 (0.028)	1.038 (0.120)
Treatment Period		
1975 – 1979	1.480 (0.678)	1.041 (0.973)
1980 – 1984	1.861 (0.492)	0.706 (0.773)
1985 – 1989	2.974 (0.229)	1.212 (0.870)
1990 – 1994	2.261 (0.364)	4.235 (0.170)
1995 – 1999	2.584 (0.295)	6.524 (0.078)
Significance of Treatment Period Variable (Wald Test)	Non-significant (p = 0.783)	Significant (p = 0.037)

Evaluated at 15 years post cancer diagnosis. 1970-1974 is the reference category for the treatment period variable, and BMT exposure = No is the reference category for the BMT exposure variable; RT is radiation therapy, BMT is bone marrow transplant, CED is the cyclophosphamide equivalent dose, and OR is the odds ratio