Incident Cataracts Following Protracted Low-Dose Occupational Ionizing Radiation Exposures in United States Medical Radiologic Technologists: Statistical Methods for Exploring Heterogeneity of Effects And Improving Causal Inference

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Craig Steven Meyer

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Richard F. MacLehose, Alan R. Lifson

February 2016

Acknowledgements

This research was funded by the intramural program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services. Craig Meyer was supported by the Midwest Center for Occupational Safety and Health: training grant CDC/NIOSH 2T42 OH008434.

Dedication

This is dedicated to my friends and family who have stuck with me through it all.

Abstract

Background: Medical radiologic technologists are routinely exposed to protracted low-dose occupational ionizing radiation. The U.S. Rad Tech (USRT Study) was begun in 1982 by the National Cancer Institute in collaboration with the University of Minnesota School of Public and the American Registry of Radiologic Technologists to investigate potential health risks from occupational ionizing radiation. Ionizing radiation exposures have been associated with cataracts, which if left untreated can lead to visual impairment or blindness. Phenotypes of cataracts are characterized by their location in the eye lens and include posterior subcapsular and cortical cataracts (most commonly associated with ionizing radiation), and nuclear cataracts (associated with age).

Methods that allow investigators to flexibly examine the extent of heterogeneity across many covariate strata are needed to help characterize the extent of any heterogeneity. One such potential method is boosted regression trees, a machine learning ensemble model that is particularly well suited to prediction while incorporating interactions. As prediction is becoming increasingly important for epidemiologic investigations (causal inference methods commonly require the use of prediction), exploration of the utility of machine learning methods in epidemiology is warranted.

Occupational epidemiologic cohort studies are often susceptible to selection bias from the healthy worker survivor effect (HWSE), whereby less healthy individuals leave work and accrue less exposure compared to healthier individuals who stay at work and continue to accrue exposure. As a result, the association between exposure and an outcome may be attenuated, or even reversed in some cases. G-methods are a family of

analytical tools that were developed to address situations that may be affected from time-varying confounding and structural bias as seen in the HWSE. One such method, the parametric g-formula, is a rigorous computational model that has been used to correct effects estimates for potential bias from the HWSE.

Objective: The overall objective of this research is to explore the relationship between protracted low-dose exposures to occupational ionizing radiation and the risk of cataracts in medical radiologic technologists in the United States and its territories, and to propose methodologic techniques to help estimate causal effects in such settings. The overall objective of this research will be accomplished in three separate manuscripts.

Manuscript 1: *Aim:* To estimate the overall association between protracted exposure to low-dose occupational ionizing radiation and incident cataracts in medical radiologic technologists. *Methods:* Cox regression models were used to model time to cataract predicted by ionizing radiation. Technologists were followed from year first worked as a radiologic technologist starting at age 18 or older, until report of cataracts or administrative censoring at the third survey. *Results:* After adjustment for birth year, sex, and race / ethnicity (N=69,798), ionizing radiation was significantly associated with increased hazard of cataracts with a time-varying effect ($p<0.001$) that while initially elevated, decreased over time. Hazard ratios of cataract per 10-mSv increment of radiation were statistically significant at age 20 [HR=1.09; 95% CI = (1.04, 1.14)] and age 30 [HR = 1.04; 95% CI = (1.00, 1.09)], but were not significant after age 30. Sensitivity analyses indicated strong evidence that selection bias from the HWSE were present and may have explained the time-varying effect. Additionally, a literature review

found five population-based studies of cataract subtype prevalence over time, and indicated that there was potential for misclassification of cataracts in the USRT study that may have biased effect estimates.

Manuscript 2: *Aim:* Use boosted regression trees to fully characterize the distribution of the effect of occupational ionizing radiation on cataracts in medical radiologic technologists. *Methods:* A boosted regression tree model was used to build a prediction model of cataracts. The cohort was restricted to those ages 24–44 at baseline (N=43,513). Predictions from the model were used to calculate risk differences of cataracts between high dose (75^{th} percentile of observed badge dose: 61.31 mSv) and low-dose (25^{th} percentile of observed badge dose: 23.90 mSv) occupational ionizing radiation in strata of potential effect modifiers. *Results:* Overall, there was a significant population average effect [RD=0.002; 95% CI = (0.002, 0.015)]. Additionally, subgroups were found with larger risks than the population average including those born earliest, those with diabetes, macular degeneration, glaucoma, or were overweight (BMI > 25) at baseline. Those who were youngest and those without macular degeneration conversely had lower risk differences compared to the average.

Manuscript 3: *Aim:* Use the parametric G-formula to adjust effect estimates for the healthy worker survivor effect in the estimated risk of incident radiogenic cataracts in medical radiologic technologists. *Methods*: The parametric g-formula was used to estimate cataract risks under different hypothetical scenarios limiting badge dose in five-year periods to the 80^{th} percentile (badge dose ≤ 18.38 mSv), 60^{th} percentile (badge dose ≤ 9.06 mSv), 40^{th} percentile (badge dose ≤ 4.47 mSv), and 20^{th} percentile (badge dose ≤

2.08 mSv) of observed dose, and a 5-mSv reduction in dose estimates in each period over follow-up (N=69,798). Cumulative incidence risks and risks conditional on survival of cataracts from these treatment regimes were compared to the status quo (no intervention of dose) with risk differences and 95% confidence intervals. Substantively important differences in both cumulative incidence of cataracts and conditional risks of cataracts between the natural course and treatment regimes were found. There was evidence that decreasing the dose of radiation exposure could reduce the risk of cataracts, even at relatively early ages.

Conclusion: Overall, our results indicate that low-dose occupational ionizing radiation exposures elevate the risks of cataracts in medical radiologic technologists in the USRT Study, as our three manuscripts found significant associations between occupational ionizing radiation and cataract risks. Additionally, methods were proposed to explore heterogeneity of effects and improve the causal interpretation of effect estimates in the association between ionizing radiation and cataracts. Validation of cataracts is warranted and future studies would benefit from information regarding phenotypes of cataracts.

Table of Contents

List of Tables

List of Figures

Organization

The organization of this thesis provides an initial introductory chapter followed by three individual papers and a concluding chapter. Because the three individual papers are in preparation for peer-review, there may be some redundancy in material.

Chapter 1

Introduction

Medical radiologic technologists are healthcare professionals that perform imaging of the human body and anatomy to assist in diagnostic evaluation and treatment of disease. They specialize in radiography, nuclear medicine, radiation therapy, and other imaging specialties. Radiologic technologists are routinely exposed to occupational low-dose ionizing radiation and may be at risk for developing health conditions from their work exposures. Ionizing radiation is known to cause tissue damage in the human body[1] and is associated with health risks such as cancers[2] and cardiovascular disease.[3] Additionally, ionizing radiation has been found to be associated with lenticular opacities, or cataracts, that when left untreated lead to visual impairment and blindness.[4] For this reason, the National Cancer Institute (NCI), the University of Minnesota School of Public Health, and the American Registry of Radiologic Technologists (ARRT) began the U.S. Rad Tech (USRT) Study[5] to investigate the possible health risks from protracted low-dose occupational ionizing radiation exposures. The USRT Study eligible cohort (N=146,022) included all medical radiologic technologists in the United States and its territories certified by the ARRT for more than 2 years prior to 1982. A baseline questionnaire (Q1: 1984–1989) was sent to 132,454 technologists assumed to be still alive and completed by 104,504. The second questionnaire (Q2: 1994–1998) was sent to 126,628 participants assumed to be alive at that time and completed by 90,972. And the

third questionnaire (Q3: 2004-2006) was sent to the 102,357 technologists assumed to be alive and had completed the first and / or second questionnaire, and was completed by 73,625.

Radiation is the transfer of electromagnetic energy and comes in the form of ultraviolet radiation, cosmic radiation, and ionizing radiation. Ionizing radiation has the ability to remove electrons from an atom, thus creating ions. Differentiated by the source of energy, ionizing radiation comes from gamma radiation, where gamma rays are emitted from the nucleus of an atom, or X-rays, which are propagated from electrons located outside of the nucleus. X-rays were first discovered by Roentgen in 1895 and quickly came into use for medical purposes.[6] It was not immediately clear that the X-rays could cause damage to the human body, but it became apparent that tissue damage could occur from radiation.[6] It is now known that ionizing radiation can damage or alter the DNA in a cell. However, as more became known about the controlled use of ionizing radiation, the benefits for both the diagnostics and treatment of disease could be garnered without loss of safety. For example, X-rays are used to locate broken bones or tumor growths, radioactive material may be injected into the body to follow the movement of the radiation through the body, and treatments for cancer may use radiation to promote tumor cell tissue death.

The use of medical radiation is steadily increasing in the United Sates. It was estimated in 1982 that the per capita dose was 0.54 mSv and the population collective dose was 124,000 person-Sv. This had increased in 2006 by almost 600% to 3.0 mSv per capita and over 700% to 900,000 person-Sv for the population collective dose.[7] Radiation

dose is measured in different quantities depending on the radiation source, the specific tissue in the body, and the potential for harm to the body. These quantities are referred to as *absorbed dose*, which is the quantity of energy from ionizing radiation deposited in a specific tissue or organ measured in milli-Grays (mGy), and *effective dose equivalent* (or dose equivalent), which is calculated from the absorbed dose to all organs in the body, the relative harm from the source of radiation, and the different sensitivities of tissues in the body and is measured in milli-Sieverts (mSv).[8] Although occupational protective practices are in place for the use of ionizing radiation in medical settings,[9] it is still important to continue actively researching the effects of occupational ionizing radiation on medical professionals.

The etiology of cataracts is most likely multifactorial. Many risk factors include age,[10] diabetes mellitus,[11] obesity,[12] hypertension and cardiovascular disease risk factors,[13,14] use of corticosteroids,[15] genetics,[16] alcohol and tobacco use,[17] trauma to the eye,[18] ultraviolet radiation,[19] and ionizing radiation.[20] Phenotypes of cataracts are characterized by their location in the lens of the eye and include posterior subcapsular (located in the posterior pole of the lens), cortical (begins at the outer edge of then lens and moves inward), nuclear (located in the center of the lens) and mixed types (any combination of the three individual subtypes).[21] Ionizing radiation is mainly associated with posterior subcapsular and cortical cataracts.[22–24] Acute exposures of ionizing radiation have been associated with cataracts in Hiroshima atomic bomb survivors[23] and infants treated with radiologic agents.[25] Smaller doses of ionizing radiation have been associated with cataracts in those exposed to contaminated-building materials,[26]

astronauts,[27] airline pilots,[28] cardiologists,[29] industrial radiographers,[30] and Chernobyl clean-up workers.[31] The relationship between protracted low-dose occupational ionizing radiation exposures and risks of cataracts is still an active area of research, and the USRT Study with its large cohort and extensive follow-up is uniquely capable of investigating risks of cataracts from low-dose ionizing radiation exposures. The goal of *manuscript 1* is to investigate the relationship between protracted low-dose occupational ionizing radiation exposures in medical radiologic technologists in the USRT Study and their risks of cataracts.

Though rarely used in epidemiology, machine learning and data mining techniques have been developing in the field of computer science concurrently to the development of some of the standard regression models employed in epidemiology such as Cox regression, generalized linear, and generalized additive models. The strengths of machine learning techniques like neural networks, regression trees, and boosting include predictions and identifying patterns from data. Their limited use in epidemiology may be from a lack of training in this area, but also because they are less easily interpretable as compared to results from a standard regression model. However, prediction is relevant for the epidemiologist, and the exploration and use of such techniques in epidemiologic studies are warranted.

Standard analyses in epidemiologic investigations often assume homogeneous effects, i.e. that the effect estimate is constant over strata of all the other covariates in the model.[32] Epidemiologists often implement analyses by assuming homogeneity of effect and heterogeneity is often ignored. As a result, reports of epidemiologic investigations do

not always consider ways to express heterogeneity in their results and typically rely on tests of significance for differences in effects between a limited number of strata.[33] Standard regression models have difficulty with these types of situations and tend to have poor power to detect effects.[34]It is difficult enough to describe an interaction between a main effect and one variable let alone two or more variables. Typically as a result of these limitations, non-significant interactions are excluded and the statistical model is biased toward main effects. Furthermore, model building by the analyst, though systematic, may overlook important interactions in the statistical (and physical) model. However, heterogeneity is not an unrealistic assumption in epidemiologic studies. That an exposure has a different effect for different groups of people based on their covariate patterns is likely the rule rather than exception. Therefore, it is beneficial for the epidemiologist to have tools that allow the investigation of heterogeneity in effect estimates across different covariate patterns, and meaningful ways to communicate these results to a general audience.

Boosting is one of the most powerful machine learning ideas introduced in the last twenty years and is one of the ensemble machine learning methods. Boosting is a general procedure that combines the outputs of many simple models to produce a single model, often called a "powerful committee". Any type of statistical model (e.g. linear or logistic regression) can be used in a boosting algorithm; however, classification and regression tree (CART) models have become the most widely adopted.[35] CART models are intuitive and easily interpretable and have several desirable properties including (1) their ability to handle various types of responses, (2) complex interactions are modeled

simply, (3) they are invariant to monotonic transformations.[36] Boosted regression trees (BRT), combines boosting and CART models while drawing on developments from statistics and computer science. BRT have been shown to have excellent predictive properties while easily incorporating interactions.[37] Once a prediction model is obtained, that model can be used to look at the effect measure of interest for any desired covariate pattern. For example, risk differences of radiogenic cataracts can be obtained from the prediction model for every possible covariate patterns and plotted as a histogram (e.g. the RD for high vs. low radiation exposure among Blacks, among Whites, among those born in the 1930's, 1940's, and so on). The goal of *manuscript 2* is to use a BRT model to build a prediction model of the risk of cataracts from low-dose occupational ionizing radiation exposure and characterize the distribution of effects.

Often, the goal of an occupational epidemiologic study is to estimate a causal effect between a workplace exposure and a health outcome. For example, in the present study we are seeking to estimate the effect of occupational ionizing radiation on the risk of cataracts in medical radiologic technologists. Effect estimates from the study's results may inform current occupational health regulatory practices. It is only meaningful to do this if it is possible to causally interpret the effect estimate in this relationship. However, for reasons of study design or other characteristics of the data, confounding bias can limit causal inference of effect estimates. Incomplete control of confounding is a well-recognized source of bias and must be accounted for when estimating effects of exposures.[38] The healthy worker effect (also referred to as healthy worker survivor bias) has long been known as a common source of confounding in occupational health

epidemiologic studies.[38–41] The healthy worker effect was first recognized in 1885 by Ogle with the observation that individuals change job tasks, occupations, and leave employment based on health status.[42] It is now described as a combination of different processes that lead to bias in effect estimates. One such process involved in the healthy worker effect is apparent when comparing mortality and health status of occupational cohorts to the general population. Individuals with poorer health status and pre-existing health conditions are often not selected for employment for a variety of reasons (e.g. unable to look for work, less likely to be hired by employers). As a result, occupational cohorts are often healthier than the general population due to the selection of individuals into the workforce. Direct comparisons between these two groups are not informative.

Robins described the healthy worker bias as the result of time-varying confounding from work status that occurs when a participant leaves the work place as a result of previous exposure, hereafter called the healthy worker survivor effect (HWSE).[43] Healthier workers continue to accrue exposure while less-healthy individuals drop out of the study and accrue less exposure; this may attenuate (or even reverse) the relationship between exposure and disease.[42] To address this type of confounding, Robins first introduced g-methods and the parametric g-formula to account for the HWSE[43] and adjust effect estimates for time-varying confounding in epidemiologic studies.[44–46] G-methods include g-estimation of structural nested models, marginal structural models, and the g-formula (parametric and non-parametric). Though applications of these methods are increasing, there have been relatively few examples of the g-formula in use until recently due to its rigorous computation.[40,47–50]

The g-formula can be used quite generally for many applications. Because there is a structural problem inherent in the HWSE (e.g. technologists can only be exposed to occupational ionizing radiation while at work) standard methods fail to properly account for it in effect estimates. Briefly (more detail will follow in Chapter 4), adjustment for work-status in the regression model (a common solution to account for the HWSE) would induce a type of bias known as collider-stratification bias.[51] The g-formula is able to account for time-varying exposure (occupation ionizing radiation) and time-varying confounding (work-status) that is affected by previous exposure, and not introduce collider-stratification bias.[52] Therefore, we aim to use the parametric g-formula in the present study to account for the HWSE in the USRT cohort.

In the longitudinal setting, the parametric g-formula is straightforward to conceptualize as an extension of traditional epidemiologic standardization that is applied over many time points. Parametric models are built that predict the outcome (as well as other covariates) conditional on exposure (and other covariates). These parametric models are used to predict the potentially counterfactual outcomes. This leads to a natural comparison of "what would have happened" to the whole population under different hypothetical interventions. This can be thought of as an imputation-based causal inference method.[48] Similar to the use of imputation to complete datasets with missing data, the parametric g-formula uses imputation to fill missing counterfactual observations of participants that have only one observed observation for their exposure state.[53] The goal of *manuscript 3* is to use the parametric g-formula to analyze risks of cataracts from occupational ionizing radiation in medical radiologic technicians in the USRT Study

under various exposure levels, adjusting them for potential bias from the HWSE.

Objective

The overall objective of this research is to explore the relationship between protracted low-dose exposures to occupational ionizing radiation and the risk of cataracts in medical radiologic technologists in the United States and its territories, and to propose methodology that may improve exploring heterogeneity of effects in the exposure-disease relationship and causal inference from estimates in this association.

Specific Aims

The overall objective of this research will be accomplished in three separate manuscripts through the following specific aims:

Manuscript 1

1) To estimate the overall association between protracted exposure to low-dose occupational ionizing radiation and incident cataracts in medical radiologic technologists.

Hypothesis 1: Protracted exposure to low-dose occupational ionizing radiation increases the risk of incident cataracts in medical radiologic technologists.

Manuscript 2

2.) Use boosted regression trees to fully characterize the distribution of the effect of occupational ionizing radiation on cataracts in medical radiologic technologists.

Hypothesis 2: Occupational ionizing radiation exposures will have different effects on cataracts in different subgroups of medical radiologic technologists.

Manuscript 3

3.) Use the parametric G-formula to adjust effect estimates for the healthy worker survivor effect in the estimated risk of incident radiogenic cataracts in medical radiologic technologists.

Hypothesis 3: After adjustment for the healthy worker survivor effect using the parametric g-formula, there will be an elevated risk of cataracts from occupational ionizing radiation in medical radiologic technologists.

Chapter 2

Manuscript 1. Self-reported incident cataracts following protracted low-dose occupational ionizing radiation exposures in U.S. medical radiologic technologists

Introduction

Cataracts are a clouding of the lens that can cause visual impairment and potentially lead to blindness. The etiology of cataracts is not fully understood but is most likely multifactorial, with risk factors including age,[54] genetics,[16] diabetes,[11] hypertension,[13] cardiovascular disease risk factors,[14] obesity,[12] ultraviolet (UV) radiation,[19] use of corticosteroids,[15] tobacco and alcohol use,[17] and trauma to the eye.[18] It has been well documented that the lens is radiosensitive with cataract formation resulting from exposures to radiation.[4,20,55] Studies have concluded that the development of radiation-induced cataracts are a latent effect of acute radiation exposures seen in the Hiroshima atomic bomb survivors[23] and infants treated with radiologic agents.[25,56] Though there has also been evidence of cataracts from lower-dose ionizing radiation exposures in airline pilots,[28] cardiologists,[29,57] astronauts and cosmonauts, [24,27,58] clean-up workers from the Chernobyl nuclear reactor accident,[59] and individuals exposed to radioactive contamination in building materials,[26] the risk of cataracts from protracted low-dose exposures to ionizing radiation, particularly in occupational settings, is still an active area of research.

Of interest to the present study are medical radiologic technologists in the U.S. Rad Tech (USRT) Study. It constitutes the largest occupational cohort of medical radiologic technologists in the world, and with its extensive follow-up, is uniquely capable of investigating risks of cataracts from protracted occupational ionizing radiation exposures. Cataract types are characterized by their location in the lens and include nuclear, cortical, posterior subcapsular, and mixed types.[21] Ionizing radiation exposures are most commonly associated with posterior subcapsular cataracts[22] located in the posterior pole of the lens, but have also been associated with cortical cataracts.[23,24] Previous report of cataracts in the USRT Study were ascertained by self-report for any type of cataract and had a positive association with occupational ionizing radiation, though it was not statistically significant.[60] However, recent methodological enhancements made to the dosimetric system for occupational ionizing radiation exposures in this cohort[61] warrant further investigation of the potential risks of cataracts from occupational exposures to ionizing radiation. Utilizing the new dosimetric exposure information we addressed the following aim in this study: to estimate the overall association between protracted exposures to low-dose occupational ionizing radiation and self-reported incident cataracts in medical radiologic technologists.

Methods

Study Population

The USRT Study began in 1982 as a collaborative effort between the American Registration of Radiologic Technologists (ARRT), the National Cancer Institute, and the University of Minnesota School of Public Health.[5] The study cohort includes medical

radiologic technologists certified in radiography, nuclear medicine, radiation therapy and other imaging specialties. Briefly, a cohort of all radiologic technologists who were certified by the ARRT for more than 2 years prior to 1982 and were residents of the United States and its territories were eligible to participate in the study (N=146,022). A baseline questionnaire (Q1: 1984–1989) was mailed to 132,454 radiologic technologists assumed to be alive and was completed by 104,504 participants. A second questionnaire (Q2: 1994–1998) was sent out to 126,628 technologists and completed by 90,972. The third questionnaire (Q3: 2004–2006) was sent out to the 102,357 study participants that had completed the first and / or second questionnaire assumed to be still alive, and was completed by 73,625 technologists.

Cataracts

Cataracts were ascertained from self–report on the third survey (Q3: 2004–2006). On Q3, participants were asked, "Did a doctor ever tell you that you had any of the following eye conditions?" followed by options for macular degeneration, cataract, and glaucoma. Participants were directed to provide the year they were first diagnosed for the conditions for which they marked YES. If participants marked YES for cataract, they were additionally asked if they had any cataracts removed and the year of first extraction.

Occupational Ionizing Radiation

The historical dose reconstruction for occupational ionizing radiation exposure estimates of the participants in the USRT Study was recently revised and is reported elsewhere in detail.[61] Briefly, the dosimetric system for occupational ionizing radiation

incorporated information from a variety of sources including: individual cohort badge dose measurements where available (badge readers are worn by the technologist to record radiation dose); published badge dose estimates from 11 publications for years preceding 1960; information gathered from the Rad Tech questionnaires (Q1, Q2, and Q3) about personal protective measures (e.g. lead apron use, placement of badges); the number of years and hours per week worked as a radiologic technologist; the types of procedures performed (e.g. X-rays, fluoroscopy); the type of facility (e.g. hospital or clinic), as well as work in military or civilian locations; and information about historical technical specifications and regulatory occupational dose limits. Organ dose conversion coefficients were calculated[62] and used in the individualization of organ dose estimates. For each participant, individual annual probability density functions of badge dose were developed to capture the range of plausible values of the technologists' true annual badge dose based on the information about the population dose distribution, period specific protection practices, and individual work-related characteristics. Multiple values were drawn for each technologist using Monte Carlo simulation techniques and the final measurement was reported as the arithmetic mean from the dose realizations. The final dosimetry model estimates resulted in annual badge dose (mSv) measurements and organ dose (mGy) measurements for 12 organs including the eye lens for 110,000 participants.

Ultraviolet Radiation

From the 73,625 cohort members who completed the third survey, UV exposure information collected included past sun exposure (time spent outdoors) during different age ranges (<13, 13–19, 20–39, 40–64, ≥65) and location of residence (city, state,

country) during those ages. Participants were asked, "When you were (this AGE), on weekdays (between the hours of 9AM and 3PM), about how many hours per day did you usually spend in strong sunlight?", with options of 0, less than 1 hour, 1–2 hours, 3–4 hours, or 5–6 hours. The question was repeated for weekends in summer. Additionally, environmental UV exposures were obtained by linking reported geographic locations during the same ranges with NASA's Total Ozone Mapping Spectrometer database[63] which provides daily information on noon-time erythemal dose estimates (UVA and UVB). Mean values for summer months during 1978–2003 were selected and assigned values of UV ambient exposure to individuals.[64] Ambient exposures were assumed to be the same for years of exposure preceding 1978 and would reflect the relative differences in UV levels between locations. The final measure of environmental exposure was cumulative from birth up to each year of observation.

Health Conditions and Characteristics

Health conditions were coded as 'Yes' if a participant had reported a condition for any of the first three surveys, and 'No' if they had never reported the condition. Selected health conditions included known potential risk factors of diabetes, hypertension, myocardial infarction, stroke, and other cardiovascular diseases (e.g. angina pectoris). Smoking was ascertained on all three surveys and categorized as Ever / Never. Body mass index (kg / m^2) was calculated at baseline from self-reported weight and height at Q1 or Q2, whichever was the first survey completed.

Statistical Analysis

Descriptive analyses (e.g. means, medians, frequencies, proportions, standard deviations, and ranges) were used for participant characteristics for the entire study population, and by cataracts and cataract extraction.

For the present study, participants were followed from age first worked as a medical radiologic technologist with age 18 being the earliest start point, until report of cataracts or date of completion of the third questionnaire. Participants missing information on cataracts (N=7), those reporting a diagnosis of cataract but missing year of diagnosis (N=651), missing occupational radiation exposures (N=5), had unlikely survey responses (N=3), reported never working as a radiologic technologist (N=597), had any occupational exposures before age 18 (N=2,439) or a cataract diagnosis before age 18 (N=51), had a cataract diagnosis before beginning work as a technologist (N=15), had missing data for date of Q3 completion (N=6), had only one year of observation (N=6), or had been asked to be removed from the study (N=1) were dropped from the analyses leaving a total sample of 69,844. Beginning at age of first employment, participants were followed until report of cataract diagnosis or administrative censoring at completion of Q3. Using chronologic age as the time scale,[65] Cox regression models were used to estimate the relative hazard of cataracts associated with cumulative occupational ionizing radiation exposures up to the given age using badge dose (mSv) estimates. Additionally, to test the linearity of dose-response a quadratic term for dose was included in the model. Self-reported cataract diagnosis was used as the outcome and individual occupational ionizing radiation exposures were summed up to each year of observation. The proportional hazards assumption was tested for all variables in the model. Because initial results for this aim produced somewhat counter-intuitive results in

which the association with radiation exposure varied by age at follow-up, we further conducted two sensitivity analyses as possible sources of this result: 1) we evaluated whether lack of specificity regarding cataract type in this cohort could bias associations; and 2) we quantitatively evaluated whether the healthy worker survivor effect (HWSE) could bias the association between occupational ionizing radiation and cataracts in medical radiologic technologists.

Sensitivity Analysis 1: *Bias from lack of cataract specificity*

To evaluate whether lack of specificity regarding cataract type in this cohort could bias associations, a literature review was conducted to estimate age-specific prevalence of cataracts by cataract subtype: nuclear, cortical, and posterior subcapsular cataracts. Five population-based studies were found from 1997-2006 from different regions in the world including Barbados,[66] Tanzania,[67] Taiwan,[68] Los Angeles, CA[69] and Australia.[70] Age ranges varied for each study, but the overall range included ages 40-90. Data were pooled to calculate age-specific prevalences of each subtype (nuclear, cortical, and subcapsular) separately in 10-year increments for the total pooled sample (N=20,988). Additionally, the relative proportion of each subtype was estimated by age group for three comparisons: 1) the relative proportion of each subtype to one another; 2) the relative proportion of posterior subcapsular cataracts only to cortical and nuclear cataracts grouped together; and 3) the relative proportion of posterior subcapsular and cortical cataracts grouped together to nuclear cataracts only. These comparisons were chosen to compare the proportion of subtypes more commonly associated with radiation exposures

(posterior subcapsular and cortical cataracts) to those associated with age (cortical cataracts) over time.

Sensitivity Analysis 2: *Healthy Worker Survivor Effect*

Occupational cohorts are often susceptible to bias from the HWSE.[42,71] In studies with cumulative exposures, healthy individuals are likely to accrue higher exposures as compared to less healthy individuals who may leave work due to their poor health, thus attenuating, or even reversing, the true relationship between the exposure and the health outcome.[72] Because of the potential for the healthy worker survivor effect bias, a full assessment of the component associations was warranted. Recent work by Naimi et al.[39] provided a useful framework to identify the component associations of healthy worker survivor bias.

Figure 1 shows a directed acyclic graph for the HWSE and is comprised of three key components that all need to be present to indicate strong evidence of the HWSE. Component 1 is the relationship between previous exposure and subsequent work status; component 2 is the relationship between previous work status and subsequent exposure; component 3 is the relationship between work status and the outcome of interest.

To estimate components associations, data was transformed into a person-year format, so that there was one observation for each year of follow-up for each person. To evaluate the association in component 1, occupational ionizing radiation was lagged both one year and ten years, and a pooled logistic regression model was used to predict work-status (leaving work) for the subsequent year conditional on the prior year and 10-year prior radiation dose. The association in component 2 already existed since occupational

exposure to ionizing radiation was only possible if an individual worked in a given year, thus inducing a relationship between work status and subsequent exposure. Component 3 was evaluated using a Cox regression model of time-to-cataracts predicted by work status (leaving work) for 1) the total sample; 2) those that leave work and don't return; and 3) those with intermittent work status.

Results

The 69,844 medical radiologic technologists in the study sample had 2,470,569 years of person-time (Mean=35.4 person-years) ranging in observation from 18 to 97 years of age. Participants were predominantly female (78%), White (95%), married or in a married-like relationship (68%), with a college education or more (78%), and had mean age 57.0 years (SD=8.7) at Q3. Cataracts were reported by 13% (N=9,060) of the sample with a mean age of cataract diagnosis at 62.1 years (SD=10.3). Cataract extractions were reported by 5% (N=3,391) with a mean age of cataract diagnosis of 66.1 years (SD=11.3). See Table 1 for summary statistics of the full sample and by report of cataracts and cataract extractions, separately. The highest percentage cataracts were reported between the ages of 65–69 (19%), with a large majority between the ages of 55–65 (see Table 2 for details).

After adjustment for sex, race/ethnicity, and birth year, badge dose was associated with an increased rate of reported cataracts (N=69,798 without missing information on baseline covariates). This association had a statistically significant time varying effect ($p<0.001$). Additionally, the quadratic term for badge was dose was significantly time varying ($p<0.001$). However, as the hazard ratios using the quadratic terms for dose were

nearly identical to the hazard ratios from the model with only the linear term for dose, the final model included only the linear term. Table 3 shows the hazard ratios and 95% confidence intervals of cataracts for badge dose per 10-mSv increment of dose estimated at ages 20 [HR=1.09; 95% CI= (1.04, 1.14)], 30 [HR=1.04; 95% CI = (1.00, 1.09)], 40 [HR=1.03; 95% CI = (0.99, 1.07)], 50 [HR=1.02; 95% CI = (0.98, 1.07)], 60 [HR=1.01; 95% CI = (0.97, 1.06)], and 70 [HR=1.01; 95% CI = (0.97, 1.05)]. As shown, hazard ratios were elevated at younger ages and decreased monotonically with greater age. Similar results were seen with models of eye lens dose (mGy), and with adjustment for ultraviolet radiation (environmental exposures and time spent outdoors), age first worked as a technologist, and health characteristics and conditions (body mass index, smoking, report of hypertension, diabetes, stroke, myocardial infarction) though these are not reported here.

Figure 2 shows the results of the pooled analyses of cataract type by age from the five population-based published studies (N=20,988) for the first sensitivity analysis. In the top left corner, the prevalence of each subtype of cataract increases with age, though posterior subcapsular cataracts appear to have lower rates of increase over time. In the top right corner, the relative prevalence of each subtype is compared to one another; at younger ages cortical cataracts constitute a larger proportion of overall cataracts. By age 70, the proportion of cortical cataracts is approximately equal to nuclear cataracts. Posterior subcapsular cataracts appear to be relatively constant over time with a little over 10% of the total cataracts compared to the other two types. In the lower left corner is the comparison of the relative proportions of cortical and nuclear cataracts combined to posterior subcapsular cataracts only. The relative prevalences appear to be constant over

time with posterior subcapsular again with around 10% of the total cataracts over time. The lower right corner is the comparison of posterior subcapsular and cortical cataracts combined to nuclear cataracts only. The posterior subcapsular and cortical cataracts combined comprise 80% of the total cataracts at age 40 and decreases to below 60% by age 70 and stabilizes until age 90.

Table 4 shows the results of the second sensitivity analyses of the component associations of the healthy worker survivor effect. For component 1, the relationship between previous exposure and leaving work, a 1-year lag of a 100-mSv increment of badge dose was significantly associated ($p<0.001$) with a 56% decrease in the odds of not working (OR=0.44, 95%CI=[0.43, 0.44]). Similarly, a 10-year lag of a 100-mSv increment of badge dose was significantly associated ($p<0.001$) with a 43% decrease in the odds of not working (OR=0.57, 95%CI=[0.57, 0.58]). The Cox regression models for component 3 showed a significant ($p=0.001$) increase in the hazard ratio of cataracts when not at work for the entire sample (HR=1.14, 95%CI=[1.05, 1.23]). The association was similar for those who had intermittent work status (HR=1.17, 95%CI=[1.05, 1.31]), and even stronger for the sub-sample of individuals that left work as a radiologic technologist and did not return (HR=1.74, 95%CI=[1.38, 2.20]).

Discussion

The present study found a significant association between cataracts and cumulative occupational ionizing radiation for both badge dose (mSv) and eyelens dose (mGy). This relationship varied significantly over time; The HR was most elevated at younger ages and decreased over time. Additionally, strong evidence was found to

indicate the presence of the healthy worker survivor effect, with significant results found for all three of the HWSE component associations. Cumulative badge dose exposures lagged both 1 and 10 years were associated with staying at work for component 1. Component 2 existed a priori, while leaving work was significantly associated with report of cataracts for the entire sample and the subgroup analyses for intermittent work status and those who leave work and never return for component 3.

Our main result of an elevated risk of cataract is consistent with findings from recent research on the risks of cataracts from protracted low-dose occupational ionizing radiation in a cohort of Chinese industrial radiographers[30] where exposed workers were more likely than unexposed workers to develop cortical cataracts [HR=2.58; 95%CI=(1.36, 3.82)] and posterior subcapsular cataracts [HR=3.57; 95%=(1.27, 4.79)], but not nuclear cataracts [HR=0.93; 95%CI=(0.78, 1.11)]. A previous report of cataracts in the USRT cohort found technologists in the highest dose category (eye lens dose: mean, 60 mGy) compared to the lowest dose category (eye lens dose: mean, 5 mGy) had an adjusted relative hazard of 1.18 [95%CI=(0.99, 1.40)], though this analysis restricted the sample of technologists to ages 24–44 at baseline (N=35,705).[60] It should be noted that the present study included older members of the cohort that would have been historically exposed to higher doses of occupational ionizing radiation. Other studies have similarly shown a relationship between low-dose ionizing radiation exposures and cataracts.[26,28,29,31,57,58,73]

There are three main types of cataract defined by their clinical appearance and location on the lens: nuclear, cortical, and posterior subcapsular.[21] Ionizing radiation is generally, though not exclusively, associated with cortical and posterior subcapsular

ocular opacities.[20] Historically, posterior subcapsular opacification was thought to be a characteristic of the radiation damage to the lens equator, although more recent data suggest that radiation induced opacities can be found in the lens cortex as well.[24,74] Various studies have found that nuclear cataracts are not associated with radiation exposure.[20,55] Age-related cataracts are most commonly found in the nuclear region and cortical cataracts are also commonly found in diabetic patients.[11,75,76] The association in the present study significantly decreased with increasing age and is difficult to explain biologically. One possible explanation for this result includes potential differential distributions of cataract subtypes by age. Since the subtypes of cataracts most commonly associated with ionizing radiation are relatively rare, they could have been overwhelmed at later ages by other types of cataracts whose prevalence increase with age. This would have led to increasing amounts of non-differential misclassification with increasing age and could have led to the observed decrease in the association with age. Since information on cataract subtypes were lacking in the USRT study, we were unable to investigate the relationship between ionizing radiation and specific cataract subtypes. Based on the time-varying result found in the Cox regression analysis, we hypothesized that the effect could in truth be constant, but the misclassification of cataracts would increase with age as age-related cataracts dominate the overall prevalence of cataracts, thus inducing a time-varying effect between ionizing radiation and cataracts.

The literature review on age-specific cataract rates found five population studies and none in cohorts of medical radiologic technologists. First, it is clear that posterior subcapsular cataracts alone were not predominant at younger ages. It appeared that cortical cataracts were predominant from age 40 until about age 70, and the same was

true when grouping cortical and posterior subcapsular cataracts together. The populations in these studies are different from medical radiologic technologists in that they were not subject to protracted low dose exposures to ionizing radiation. However, as there is evidence of the relationship between ionizing radiation and posterior subcapsular cataracts (most commonly) and cortical cataracts, it may be that in medical radiologic technologists, the predominance of these two types of cataracts at younger ages may be more pronounced even than seen in the populations from the literature review. Given the significant relationship found between low-dose ionizing radiation exposures and cataracts in technologists in the present study, it would be important for future investigations of the USRT cohort to validate cataracts and cataract subtypes.

Another possible explanation for these results is the potential for selection bias in the cohort and the healthy worker survivor effect in the effect estimates. The healthy worker survivor effect can occur in occupational cohorts when healthy individuals remain at work and accrue higher exposures while those who are not healthy leave work and accrue lower cumulative exposure. This resulting bias can attenuate the relationship between occupational exposure and health outcomes, or in some cases a reversal in the relationship. [71,77–79] The results of our exploration of the components of the healthy worker survivor bias (which provided strong evidence all three component association were present in the analyses) suggest this selection bias could explain the decreased hazard ratio of cataracts from ionizing radiation with increasing age as a result of the HWSE. Standard regression methods fail to address the structural problem of time-varying confounding of exposure and work status.[43,77,80] Future planned analyses aim to address the HWSE and correct the biased effect estimates using g-methods. The strength

of the results in the present analyses warrant using more sophisticated methods that seek to improve the causal interpretations of the effect estimates.

Diagnoses of cataract are recorded through self-report asking only if the participant has been diagnosed with a cataract and the date of first diagnosis. Report of cataract diagnoses was not confirmed from participants' medical records. Moreover, radiogenic cataracts are generally associated with certain phenotypes such as posterior sub-capsular or cortical cataracts, and this information is not available in the USRT study. There is little that can be done to address this problem in the present study. However, the results of the analyses in this study suggest that further investigation of phenotypic subtypes could be of value. Future studies in this cohort may need to obtain medical records to validate cataracts and cataract phenotypes to alleviate concerns about misclassification of cataracts and allow the investigator to analyze cataract phenotypes that are known to be radiogenic.

In conclusion, a significant association was found between occupational ionizing radiation and cataracts in a cohort of U.S. medical radiologic technologists that decreased over time. The uncertainty around this relationship may necessitate further study. In particular, g-methods that have been shown to account for time-varying confounding in occupational cohorts should be employed to address any potential bias from the healthy worker survivor effect. Additionally, future studies in the USRT Study should include medical chart abstraction to validate cataracts and cataract phenotypes.

Table 1. Summary statistics of personal characteristics of United States medical radiologic technologists for the total sample (n=69,944) and by report of cataracts and cataract extraction, U.S. Rad Tech Study, third survey 2004–2006[a]

	Total (N=69,844)	Cataract- Yes (N=9,060)	Cataract - No (N=60,784)	Extraction- Yes (N=3,391)	Extraction - No (N=65,344)
Age at Third Survey (Mean, STD)	57.0 (8.7)	67.3 (9.7)	55.5 (7.2)	70.8 (10.1)	56.2 (7.8)
Cataracts	13% (9,060)	–	–	–	–
Cataract Extraction	5% (3,391)	–	–	–	–
Age of Cataract Diagnosis	–	62.1 (10.3)	–	66.1 (11.3)	–
Badge Dose (mSv)	53.2 (44.4)	105.8 (80.1)	47.4 (34.2)	133.0 (99.5)	93.7 (65.7)
Eye Lens Dose (mGy)	54.5 (47.2)	108.4 (89.3)	46.2 (34.7)	139.4 (113.2)	94.4 (71.0)
Sex					
Female	78% (54,700)	77% (6,935)	79% (47,835)	73% (2,462)	79% (51,448)
Male	22% (15,074)	23% (2,125)	21% (12,949)	27% (929)	21% (13,896)
Race					
White	95% (66,335)	95% (8,617)	95% (57,718)	95% (3,235)	95% (62,069)
Other	5% (3,463)	5% (435)	5% (3,028)	5% (152)	5% (3,235)
Education					
High School or Less	1% (537)	3% (245)	1% (292)	4% (126)	1% (405)
Vocational	7% (4,852)	7% (677)	7% (4,175)	9% (302)	7% (4,4491)
Radiation Technology Program	11% (7,442)	9% (857)	11% (6,585)	6% (216)	10% (6,847)
College Graduate or More	78% (54,407)	78% (7,087)	78% (47,320)	81% (2,730)	78% (51,185)
Marital Status					
Never	6% (4,110)	8% (760)	6% (3,350)	9% (304)	6% (3,713)
Married / Married-like	69% (48,297)	55% (4,940)	71% (43,357)	47% (1,582)	70% (49,954)
Widowed	6% (3,977)	7% (636)	6% (3,341)	7% (243)	6% (3,656)
Divorced / Separated	19% (13,404)	30% (2,715)	17% (10,689)	37% (1,260)	18% (11,972)

a Values presented as mean(standard deviation) or %(n). Numbers may not add to total due to missingness.

Table 2. Frequencies and percentages of self-reported cataracts by age of diagnosis (N=9,060), U.S. Rad Tech Study, 2004-2006

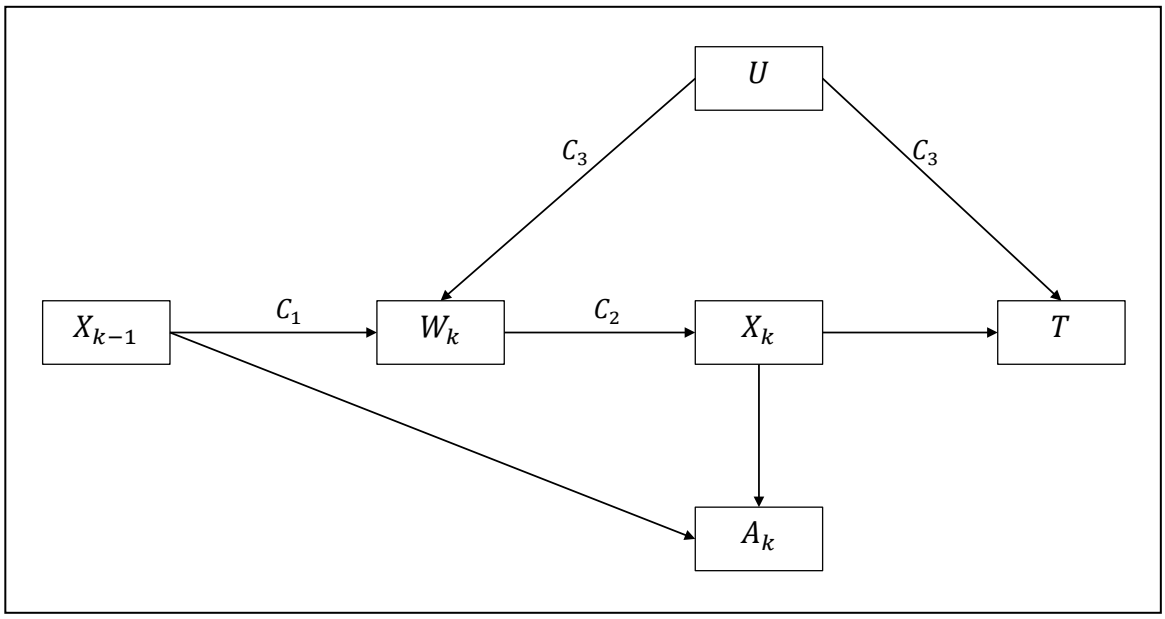
Age	N (%)	Age	N (%)
19-24	21 (0.2)	55-59	1,464 (16.2)
25-29	28 (0.3)	60-64	1,680 (18.5)
30-34	51 (0.6)	65-69	1,736 (19.2)
35-39	105 (1.2)	70-74	1,250 (13.8)
40-44	233 (2.6)	75-79	637 (7.0)
45-49	558 (6.2)	80-84	234 (2.6)
50-54	1,000 (11.0)	85-95	63 (0.7)

Table 3. Age-specific hazard ratios and 95% confidence intervals of cataracts for occupational ionizing radiation badge dose per 10mSv in U.S. medical radiologic technologists (N=69,798), U.S. Rad Tech Study[a]

Age	Hazard Ratio	95% CI
20	1.09	[1.04, 1.14]
30	1.04	[1.00, 1.09]
40	1.03	[0.99, 1.07]
50	1.02	[0.98, 1.07]
60	1.01	[0.97, 1.06]
70	1.01	[0.97, 1.05]

a Model adjusted for sex (female / male), race/ethnicity (White / Other), and birth year.

Figure 1. Conceptual model of healthy worker survivor effect[a]



a Abbreviations: for year k, X=point exposure, W=work status, A=cumulative exposure, U=unmeasured confounder, T=survival time; C_1= component association 1: previous exposure relationship to subsequent work status; C_2=component association 2: work status relationship to subsequent exposure; C_3=component association 3: work status relationship to survival time.

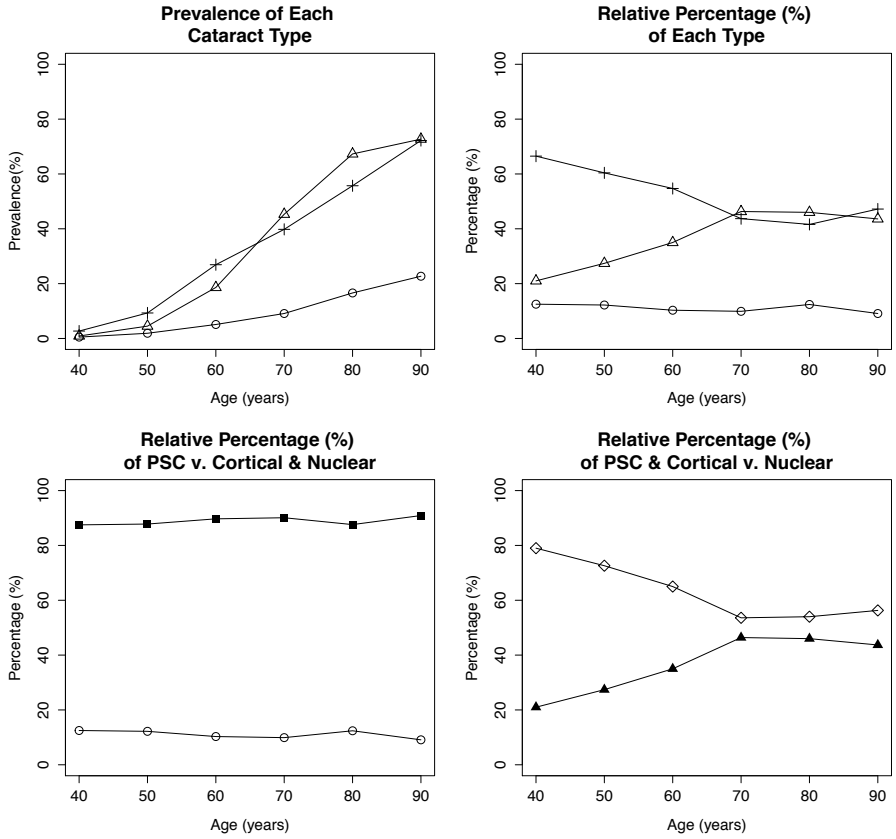


Figure 2. Age-specific rates of posterior subcapsular (PSC), nuclear and cortical cataracts calculated from pooled data (N=20,988) of five published studies including the Los Angeles Latino Eye Study,[69] the Barbados Eye Study,[66] the Blue Mountains Eye Study,[70] and cataract prevalence studies in Tanzania[67] and Taiwan.[68] Top left: Age-specific prevalence of PSC, nuclear and cortical cataracts. △=nuclear cataracts; ＋=cortical cataracts; ○=posterior subcapsular cataracts. Top right: Relative percentage of each type of cataract. △=nuclear cataracts; ＋=cortical cataracts; ○=posterior subcapsular cataracts. Bottom left: Relative percentage of posterior subcapsular cataracts compared to nuclear and cortical cataracts. ■=nuclear and cortical cataracts; ○=posterior subcapsular cataracts. Bottom right: Relative percentage of posterior subcapsular and cortical cataracts compared to nuclear cataracts. ◇=posterior subcapsular and cortical cataracts; ▲=nuclear cataracts.

Table 4. Odds ratios of not working for 1-year and 10-year lag of badge dose / 100mSv and 95% confidence intervals for component 1 of the healthy worker survivor effect, hazard ratios of cataracts and 95% confidence intervals for component 3 of the healthy worker survivor effect, U.S. Rad Tech Study[a]

Component 1: Previous Exposure → Work-Status	**Odds Ratio [95% CI][b]**
badge dose / 100mSv: 1-year lag	0.44 [0.43, 0.44]
badge dose / 100mSv: 10-year lag	0.57 [0.57, 0.58]
Component 3: Work Status → Survival Time	**Hazard Ratio [95% CI][c]**
Not-working: Total Sample (n=69,799)	1.14 [1.05, 1.23]
Not-working: Leave Work - Never Return (n=37,477)	1.74 [1.38, 2.20]
Not-working: Intermittent Work Status (n=32,302)	1.17 [1.05, 1.31]

a Component 1 is the association between previous exposure to subsequent work status (leaving work); Component 3 is the association between work status (leaving work) and survival time.
b Logistic regression models predicting not-working adjusted for sex (female / male), race / ethnicity (White / Other), follow-up age and birth year.
c Cox regression models of cataracts adjusted for sex (female / male), race / ethnicity (White / Other) and birth year.

Chapter 3

Manuscript 2. Using boosted regression trees to explore effect measure modification: an example predicting cataracts in U.S. medical radiologic technologist

Background and Rationale

Examining effect measure modification is fraught with difficulties. On the one hand, standard tests of heterogeneity have low power and epidemiologists are often reluctant to include interaction terms in models. This may lead to a failure to characterize the entire range of effects that a putative exposure might have. While heterogeneity of effects may be explored by including interaction terms in models, this is typically limited in scope to a small number of covariates, if occurring at all. On the other hand, heterogeneity of effects may be investigated by statistical 'fishing expeditions', which attempt to find some extreme strata of the data in which the effect estimate is statistically significant. These fishing expeditions, by only reporting the most extreme effect estimates, are misleading by failing to characterize the entire range of effects that a putative exposure might have. Methods that allow investigators to flexibly examine the extent of heterogeneity across many covariate strata are needed to help characterize the extent of any heterogeneity.

Historically, model building in epidemiology has been oriented toward parsimony, biasing models to main effects only.[81] It is challenging to explain two-way interactions, let alone higher-order interactions. Standard generalized linear models have

difficulty with interpretability when interactions are considered with several variables at the same time and tend to have poor power to detect effects.[82] The direct implications are the exclusion of interaction terms in the statistical model that do not reach statistical significance. However, in order to capture the overall risk pattern in the population, the statistical model requires higher-order terms for valid statistical inference.[81]

Identifying sub-groups of the population with different levels of risks are key to developing public health interventions. However, the perils of looking for these subgroups through statistical fishing expeditions involve interpreting subgroups without the context of the overall distribution of heterogeneity in the entire population.[34] Data dredging is circumspect because results may be due to chance and not generalizable outside of that specific dataset.[83] By reporting only one significant result from an interaction between covariates in the model and the target outcome, there is a chance of doing more harm than good.

The goal of the present work is to demonstrate the use of boosted regression trees, a statistical method developed in the machine learning literature, for estimating the distribution of the effect of an exposure over relevant covariates that are potential effect modifiers. Boosted regression trees have been shown to have excellent properties while easily and flexibly incorporating interaction terms.[84,85] Using data from the U.S. Rad Tech (USRT) study,[5] the largest occupational cohort of medical radiologic technologists in the world, we use boosted regression trees as a more sophisticated and flexible model to estimate risk differences of cataracts between high and low-dose occupational ionizing radiation within pre-defined population strata of the other model covariates. These risk differences can be combined to estimate the distribution of stratum specific effects from

ionizing radiation on cataracts in the entire study population. The following sections provide a description of classification and regression trees (CART),[36] the base model in the boosted model, and boosted regression trees,[35] followed by the example predicting cataracts in medical radiologic technologists.

Methods

Study Population and Measures

The USRT Study began in 1982 as a collaborative effort between the American Registration of Radiologic Technologists (ARRT), the National Cancer Institute, and the University of Minnesota School of Public Health.[5] The study cohort includes medical radiologic technologists certified in radiography, nuclear medicine, radiation therapy and other imaging specialties. Briefly, a cohort of all radiologic technologists who were certified by the ARRT for more than 2 years prior to 1982 and were residents of the United States and its territories were eligible to participate in the study. A baseline questionnaire (Q1: 1984–1989) was mailed to 132,454 radiologic technologists assumed to be alive. Two follow-up surveys have occurred (Q2: 1994–1998, Q3: 2004–2006). The third questionnaire was sent to living members of the original cohort who had completed at least one of the first two questionnaires. Following a previous analysis of these data reported by Chodick et al.[60] participants were selected aged 24–44 at the first questionnaire and with a report of cataract after the first questionnaire, for a total of 43,513 in the present study. Posterior subcapsular cataracts, a subtype of cataracts based on the location of the opacity in the eye lens, are most commonly associated with ionizing radiation exposures and occur at younger ages.[22,86] The age restriction at Q1 was

chosen since information on subtypes of cataracts was lacking in the Rad Tech Study. Participants were followed from age 18 until report of cataracts or administrative censoring at Q3. Those reporting the year of diagnosis for cataracts before the baseline survey were not included in the analysis.

It has been established that the eye lens is one of the most susceptible body parts to the effects of ionizing radiation leading to cataracts.[4,87] Several other risk factors for cataracts have been identified including ultraviolet (UV) radiation, diabetes mellitus, hypertension, hypercholesterolemia, arthritis, age, race/ethnicity, gender, smoking tobacco, alcohol use, genetics, use of corticosteroids, aspirin use, previous history of eye trauma or infections, and nutritional deficiencies.[11,15,17,20,54,87–89] Not all of these measures were ascertained in the USRT Study questionnaires. Further description of the measures included in the boosted regression tree prediction model is found below.

Cataracts and year of diagnosis were ascertained by self-report on the third survey. Participants were asked, "Did a doctor ever tell you that you had any of the following eye conditions?" followed by options for macular degeneration, cataract, and glaucoma. Participants were directed to provide the year they were first diagnosed for those conditions they reported being diagnosed.

Estimates of occupational ionizing radiation exposure in the USRT cohort were derived through a historical dose reconstruction which is described in detail by Simon et al.[61] Briefly, the estimates of occupational ionizing radiation exposure incorporated information from a variety of sources including: historical technical specifications and regulatory occupational dose limits; individual cohort badge dose measurements where available; published badge dose estimates from 11 publications for years preceding 1960;

information gathered from the USRT questionnaires about the number of years and hours per week worked as a radiologic technologist, whether the facility was civilian (hospital or clinic) or military; and personal protective measures such as lead apron use. Individual annual probability density functions of badge dose were developed for each individual to capture the plausible ranges of the technologists' true dose. Monte Carlo simulation techniques were used to draw several estimates of badge dose for each year a technologist reported working, with the arithmetic mean of all realizations of badge dose as the final estimate. The present analysis uses a cumulative measure of the annual badge doses. The annual badge dose measurements were added from the age of 18 or year first worked after the age of 18 until the end of follow-up.

Annual UV radiation exposure dose estimates were constructed from residential history information collected on the third survey and satellite data. Residential history was reported as the location they lived the longest in 5 age intervals: <13, 13–19, 20–39, 40–64, and 65 and older. The ambient UV exposures were estimated by linking participants reported residence (cite, state, country) to NASA's Total Ozone Mapping Spectrometer (TOMS) database.[63] Using daily information of erythemal dose estimates including UVA and UVB from the TOMS database during 1978–2003, mean values for the summer months were assigned to each participant. Ambient exposures were assumed to be the same for years of exposure preceding 1978. The final UV dose for the present study was a cumulative measure of summer UV exposures from childhood to the end of follow-up.

Other variables that are known potential risk factors for cataracts and were collected in the USRT Study included in the boosted regression tree model were birth

year, body mass index at baseline (Underweight / Normal / Overweight / Obese), race (White / Other), sex (female / male), hair color (black / brown / medium brown / dark brown / light brown / red / reddish-brown / blonde), iris color (blue / dark brown / grey, blue-green / light brown), education (high school or less / rad tech / vocational / college), income (<$25,000 / $25,000–49,9999 / $50,000–74,999 / $75,000– 99,999 / $100,000+), and marital status (married / divorced / never married). Health conditions were included as 'Yes' if a participant had reported a condition for any of the first three surveys, and 'No' if they had never reported the condition. The following health conditions included arthritis, rheumatoid arthritis, diabetes, glaucoma, macular degeneration, hyperthyroidism, hypothyroidism, other thyroid conditions (e.g. Hashimoto's disease), osteoporosis, myocardial infarction, stroke, hypertension, and other cardiovascular diseases (e.g. angina pectoris). Health behaviors included aspirin use, alcohol intake, smoking status (Former / Current / Never), vitamin C supplementation, and vitamin E supplementation.

Classification and Regression Trees

CART models are statistical tools that can model non-linear relationships and higher order interactions. These models are relatively straightforward, involving a series of binary 'splits' of the predictor variables (called recursive partitioning) to create a series of decision rules to create subgroups of the population and predict the outcome of interest in those subgroups. CART models are intuitive, easily interpretable, and have several desirable properties. First, CART models can handle various types of responses including continuous, binary, and categorical. Second, they have an easy representation of complex

interactions, as shown by the decision rules of the tree creating subgroups of the population. Third, they are invariant to monotonic transformations of variables, that is, transformations that maintain the order of the values of a variable, thus there is no need to transform variables before fitting the model. And finally, CART models provide inherent variable selection. That is, while there may be several variables specified as potential predictors of the outcome, not all of them will be included in the final CART model. This last point is most beneficial from the standpoint of interactions. While we can specify a CART model that, in theory, could include all possible interactions, only some of them will be included in the final model, as described below.

Each CART model starts with the entire sample in a parent node, and the algorithm will choose the first predictor variable by evaluating which variable best separates the sample into two groups by testing the homogeneity of the groups using what is called an impurity function, commonly the Gini function.[90] The two new groups, or nodes, are called the children. The relevant outcome parameter is estimated in each child node (e.g. the probability, or mean, of the outcome in each group). Next, each child node will become the subsequent parent node, and may similarly be split on a predictor variable again by testing homogeneity of the groups into more child nodes. Various stopping rules are implemented to end the fitting process (for example, each node may contain no less than five observations). Each terminal node is called the 'leaf' of the tree as the final model looks like an inverted tree. The recursive partitioning of the predictor space makes modeling interactions intuitive by splitting the sample into subgroups. Additionally, the tree may split on a variable more than once capturing non-linear associations. The end result is a model that estimates the probability of the outcome (for

dichotomous variables) or mean (for continuous variables) for each terminal node based on the categorizations determined by the CART model.

Figure 3 shows an example of a single CART model. The full sample starts in the parent node at the top of the figure. The CART algorithm choses sex as the first split variable (as this variable out of all candidate variables best minimized the impurity function), and splits the sample by sex into female and male, with female now becoming a terminal node as there are no other splits for that group (e.g. by meeting the stopping criterion). These child nodes can then become parent nodes, as is seen when the male group is further split on age with one group being males less than age 44 and the other group males greater or equal to age 44. The older males are now in a terminal node, while the younger males are split on smoking status. In the final model we see the two-way interaction between sex and age, where amongst males there is a split in an older (age $\geq$ 44) and younger (age < 44) group. We can also see the three-way interaction between sex, age, and smoking, where amongst the younger (age < 44) males, they have been split into smokers and non-smokers. In general, for a single CART model, cross validation is used to "prune" the tree, that is remove some of the splits, so that the optimal number of splits are chosen for the final model.

Although useful, CART models are limited in that they can result in poor predictive capabilities and are not as accurate as other regression methods.[85,91] Further, they tend to be unstable, meaning that models fit on different samples of the data may change the structure of the tree substantially.[92] Taken together these limitations of the individual CART model reduce their utility for accurate prediction. However, paradoxically, combining several simple CART models into one ensemble model (via

boosting) can attenuate the limitations of the individual CART models.

Boosted Regression Trees

Boosting is a statistical procedure that combines a large number of relatively simple models in order to obtain an ensemble estimate that is essentially a weighted average of the simple models. Introduced in 1995 by Freund and Schapire[93–95], boosting was originally designed for classification of categorical variables, but has been extended to regression of continuous outcomes and survival analysis.[96,97] Any type of model can be used in a boosting algorithm; however, CART models are the most widely used. Simulation studies have demonstrated that boosting algorithms can have excellent predictive capacity, often better than standard epidemiologic regression models.[98]

Boosted regression trees differ from more traditional CART models in that they do not estimate a single best CART model. Rather, they combine a large number of CART models adaptively, to optimize statistical performance (e.g. minimizing mean squared error in each node of the tree). Most importantly, boosted regression tree models are remarkably powerful in terms of their predictive capabilities, while incorporating the appealing features of CART models such as incorporating interactions and variable selection throughout the modeling procedure.[99] The adaptive, iterative process (known as functional gradient descent) of boosting seeks to improve upon the performance of the previous succession of models with an emphasis on prediction of the outcome. For example, the boosted regression tree model will start by fitting an initial CART model and then determine the classification errors of the predictions (i.e. residuals) of that fitted model. Weights are generated from the first CART model using the residuals with

observations that were not classified correctly by the model receiving a higher weight. These weights are used to fit a second weighted CART model. Predictions and classification errors are computed from the second model and used to generate weights for a third CART model. The contribution of each individual weighted CART model to the ensemble is set in advance and is called the learning rate (the shrinkage parameter, a value < 1). In other words, the weighted predictions may only move a certain amount based on the shrinkage parameter. The optimal number of CART models in the final ensemble and the learning rate are inversely related to one another (i.e. a smaller learning rate requires more CART models, while a larger learning rate requires less CART models). The final number of CART models for the ensemble (the stopping point) is determined through cross-validation.[100] By specifying the maximum number of splits in a given CART model, we can determine the maximum possible interaction level. Throughout this manuscript, we choose CART models with 3 maximum splits to allow for, at most, three-way interactions in a given CART model. The general algorithm for fitting the boosted model is shown in Figure 4. The boosting algorithm used in the present study is known as "stochastic" gradient boosting, which is implemented by taking a random sample of observations for each of the iterations (the "bag fraction"). Taking a smaller random sample at each iteration improves computational speed and produces a more accurate model.[101] To summarize, to fit the boosted regression tree the meta-parameters set in advance include the interaction depth (the number of splits for each individual CART model), the learning rate (the shrinkage parameter, between 0 and 1), and the bag fraction (the proportion of the data to be sampled at each iteration), and cross validation is used to determine the final number of CART models in the ensemble and is

dependent on values set for the meta-parameters.[100]

Once the final boosted model (the weighted average of the individual CART models) has been fitted, we are able to use the predictions in a variety of ways. In the present study, we use the final fitted model to make predictions to calculate stratum-specific risk differences of cataract from high to low-dose occupational ionizing radiation, as described below for the radiation exposure example. In this manner, we investigate whether there are two-way interactions with badge dose (readings of ionizing radiation dose from badge readers worn by the technologist, called personal dose equivalent) and each of the covariates in the relationship to the risk of cataracts in medical radiologic technologists. We have limited each individual CART model to three-way interactions, and this method could be extended to three-way (and higher-order depending on the interaction depth) interactions using the boosted model, although we have not done so here. Further description is provided below.

Analysis

Boosted classification regression trees were used to model the risk for cataracts from occupational ionizing radiation exposure and other risk factors. Cataract occurrence was modeled as the outcome in the base CART model. All variables including cumulative occupational ionizing radiation badge dose, cumulative summer UV environmental exposure, personal characteristics, and health conditions and behaviors were included as possible features to split the regression tree during model fitting. Boosted models were fitted using the "gbm"[102] and "dismo"[103] packages in R.[104] The final estimated boosted regression tree model consisted of 2,350 trees, an interaction

depth of 3, a learning rate of 0.01, and a 0.75 bag fraction.

After fitting the final model, predictions from the model were used to calculate the risk differences of cataract comparing high badge dose (75th percentile: 61.31 mSv) to low badge dose (25th percentile: 23.90 mSv) occupational ionizing radiation in the strata for the individual model covariates. To do this, marginal predicted probabilities[105,106] were calculated within each strata of possible effect modifiers (predictions were made for percentiles of continuous effect modifiers). For example, when computing the effect of high badge dose vs. low badge dose radiation among those with a high school education, we estimated predicted probabilities of cataract if a participant had a high school education and received the 75th percentile of the distribution of the badge dose, averaged over the distribution of other covariates in the dataset. Similarly, we estimated the predicted probabilities of cataract for everyone setting their occupational badge dose to the 25th percentile and with a high school education. Then, we would take the average of each set of predicted probabilities, and subtract the probability for low dose from the probability for high dose to obtain the risk difference to estimate RD = E[Pr(Y |Set[X = 1], Set[Z = 1])] − E[Pr(Y |Set[X = 0], Set[Z = 1])]. Here, RD is the risk difference, Y is cataracts the outcome of interest, X is badge dose, and Z is the covariate stratum of interest. This was repeated for each stratum of the possible effect modifier. If the possible effect modifier was a continuous measure (e.g. alcohol intake, UV Radiation), predictions were made in percentiles (1^{st}, 10^{th}, 25^{th}, 30^{th}, 40^{th}, 50^{th}, 60^{th}, 70^{th}, 80^{th}, 90^{th}, 99^{th}) of the distribution for that variable. Standard errors around the risk differences were obtained by bootstrapping. After estimating the risk differences and their standard errors, the risk differences were plotted in a histogram and

compared to the population averaged risk difference of cataract comparing the 75th percentile to the 25th percentile of badge dose. The risk differences with values that were either noticeably higher or lower than the population average were visualized in a forest plot with 95% confidence intervals.

Results

The present study contained 43,513 medical radiologic technologists from the United States who were ages 24–44 at baseline and reported cataract after the baseline survey. Participants were predominantly female (82.2%), Caucasian (96.3%), and college educated (82%). A total of 2,641 (6.1%) cataracts were reported during follow-up. We estimated 94 stratum specific risk differences and their standard errors in the analysis. Figure 5 shows the distribution of these stratum specific risk differences. Overall, there was a statistically significant population average effect (RD= 0.009; 95% CI = [0.002, 0.015]) from occupational ionizing radiation on the risk of cataracts, averaging over any possible effect modification. However, there were subgroups with much larger effect estimates, compared to the population average. This is clearly illustrated in Figure 6, the forest plot of selected risk differences. Those individuals born in the 1st percentile of birth year in 1939 (0.032; 95% CI = [0.004, 0.060]) or born in the 10th percentile of birth year in 1944 (RD = 0.024; 95% CI = [0.007, 0.041]), and those who were overweight (RD = 0.019; 95% CI = [0.009, 0.029]) had the highest stratum-specific effect of cataract from the effect of occupational ionizing radiation. Additionally, subgroups with a diagnosis of diabetes (RD = 0.016; 95% CI = [0.004, 0.027]), glaucoma (RD = 0.015; 95% CI = [-0.001, 0.031]), and macular degeneration (RD = 0.014; 95% CI = [-0.005, 0.033]) had

higher risk differences of cataract from ionizing radiation compared to the population average. Those born in the 99th percentile of birth year in 1961 (RD = 0.004; 95% CI = [0.001, 0.007]) and without macular degeneration (RD = 0.005; 95% CI = [-0.003, 0.013]) conversely had a lower risk difference compared to the population average.

Discussion

In this paper, we have demonstrated the use of a more flexible model to explore the distribution of a potentially heterogeneous set of effects over a variety of strata. Similar methods for exploring heterogeneity could be applied using a more traditional logistic regression model. However, the flexibility of the boosted regression tree model does not require defining which specific interaction terms must be in the model, and instead relies on the fitting criterion and interaction depth to determine the extent of the interactions present. Heterogeneity of effects may be present because there exist certain strata in which the exposure of interest has a pronounced effect due to a biological interaction. Identifying these subgroups may be important for developing public health interventions. However, exploration of data in hopes of finding strata in which the exposure has a significant effect is a substantial concern.[83] Having tools that allow the investigator to estimate and visualize the overall distribution of effects may jointly address these points, by helping to identify high-risk or low-risk subgroups while also portraying stratum specific significant results in the context of all other stratum specific results.

The boosted regression model we estimated indicated an increased risk of cataract occurrence associated with an increased dose of occupational ionizing radiation.

Additionally, subgroups were identified that had higher risks of cataract from occupational ionizing radiation as compared to the population averaged risk of cataracts. In particular, higher risks for cataracts from occupational ionizing radiation were seen in those born in 1939 and 1944, those with diabetes, glaucoma, and macular degeneration, and being overweight (BMI>25) at baseline. Diabetes and age are known risk factors for cataracts, but it is interesting to note the higher risks of cataracts from radiation in these groups. There may be a synergistic effect between ionizing radiation and diabetes in the risk of cataracts in medical radiologic technologists. Additionally, previous research has found ionizing radiation is associated with other eye conditions like glaucoma[107] though not macular degeneration. This result may indicate that those with other eye conditions may have higher risks to the effects of radiation from their genetic susceptibilities. Alternatively, it is possible from an epidemiologic perspective of screening and detection bias, that individuals that are diagnosed with cataracts are also more likely to report other eye conditions. This is unable to be determined, but should be considered when interpreting the results of this analysis.

One of the main limitations to this study is the ascertainment of cataract. Diagnoses of cataract are recorded by self-report asking only if the participant has been diagnosed with a cataract and the year of first diagnosis. Cataract diagnoses were not confirmed from participants' medical records. Additionally, cataracts and the exact year of diagnosis were not ascertained until the third survey (Q3). Therefore, cataracts first diagnosed before ascertainment at Q3 may be subject to recall bias. As is common with many occupational cohorts, selection biases may be present due to loss of follow-up from less healthy individuals continuing in the study. This could bias the relationship between

the occupational ionizing radiation and cataract. Moreover, radiogenic cataracts are generally associated with certain phenotypes such as posterior subcapsular cataracts, and this information is not available in the U.S. Rad Tech Study. Results from this study may warrant validation of cataract subtypes in future studies.

Building accurate prediction models is increasingly important in epidemiology. For example, current causal models like marginal structural models, propensity score models, and the parametric g-formula require the prediction of a probability (such as the probability of treatment assignment or survey response).[45,108] Causal models will only perform well if the prediction is accurate. Indeed, Greenland argues that causal inference is a prediction problem.[109] For the sole purpose of prediction, aspects of biological mechanism and interpretation do not necessarily restrain the rules for model building and evaluation of interaction.[110] From this viewpoint, it would be useful to build a large prediction model and estimate the effect of exposure over all possible covariate patterns on the additive scale, regardless of which scale the model used. It will be impossible to achieve good prediction without modeling interactions in a flexible manner.[81] The flexibility is important because there may not be a particular biological rationale for where interactions might occur. Therefore, it becomes increasingly important to include interactions in the prediction model to improve accurate prediction.

In conclusion, an overall increase in the risk of cataracts was found for higher doses compared to lower doses of occupational ionizing radiation in a cohort of U.S. medical radiologic technologists. Additionally, subgroups were found that indicated they had much higher risks compared to other subgroups. Boosted regression trees were successfully used to find heterogeneity of effects in an epidemiologic study. This is one

of the first studies to our knowledge that has demonstrated the utility of boosted regression trees for epidemiology. Packages and software used in the analysis are freely available using the R statistical computing environment.[102–104] Other common statistical software packages such as Stata[111,112] have also incorporated software tools to fit boosted regression models. Predictions are becoming increasingly important for epidemiologic studies and further study of potential applications of boosted regression trees is warranted.

Figure 3. Classification and Regression Tree of Cataracts in the U.S. Rad Tech Study

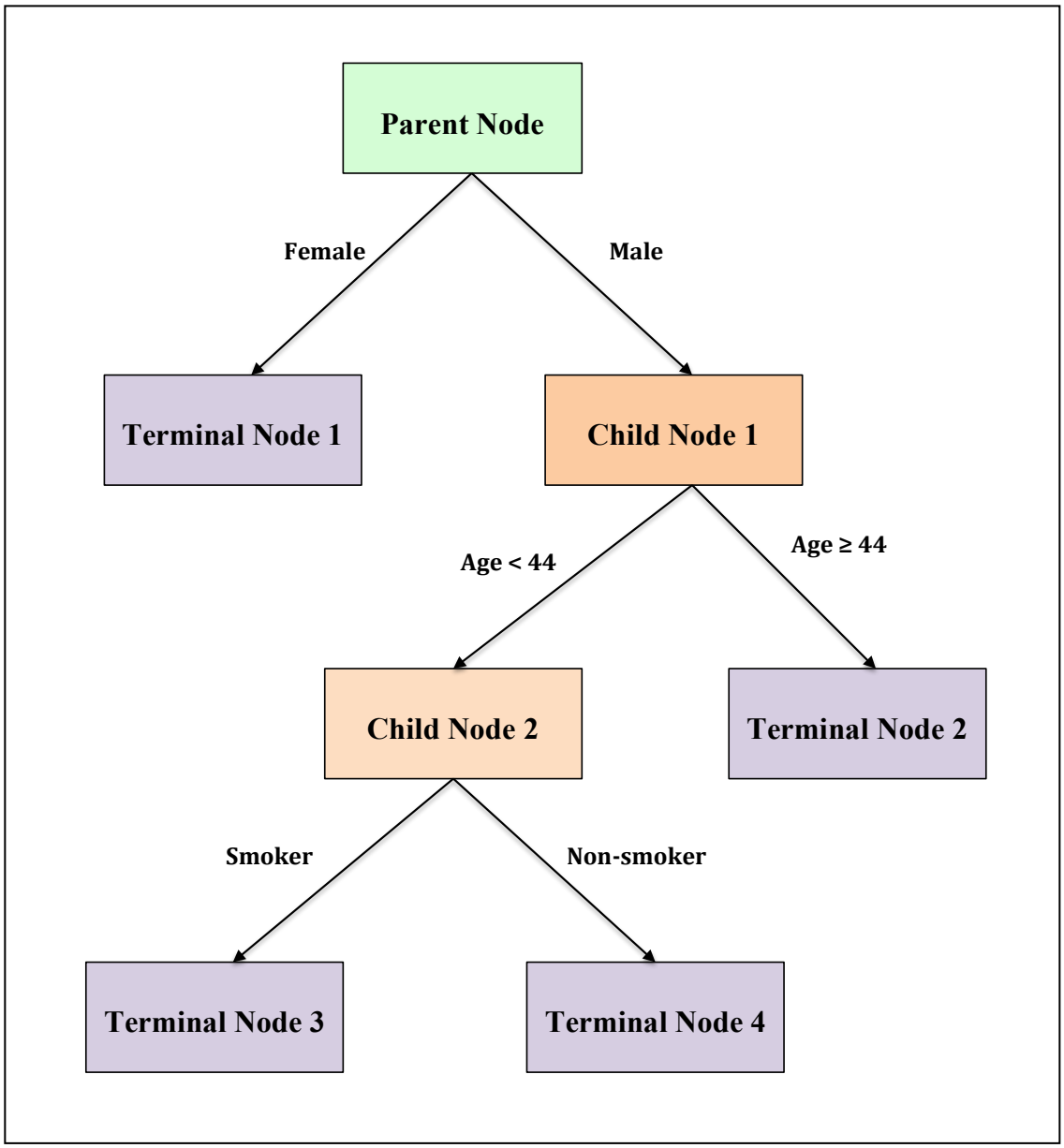


Figure 4. Boosting Algorithm: i = individual, $\hat{p}$ = estimated probability, m = iteration.

BRT Algorithm

1. Initialization:
   1a. Assign $\hat{p}_i^1 = \hat{p}$
2. For $m = 1, \ldots,$:
   2a. Compute residuals $\hat{r}_i = Y_i - \hat{p}_i^m$
   2b. Construct a tree, T^m based on the sample, with observations weighted by the magnitude of the residuals.
   2c. Obtain new predictions $\hat{p}_i^m$ as a weighted sum of $\hat{p}_i^{m-1}$ and predictions from the current tree T^m.

Figure 5. Stratum-specific risk differences of cataracts between high-dose (75th percentile of cumulative badge dose; 61.3 mSv) and low-dose (25th percentile of cumulative badge dose; 23.9 mSv) occupational ionizing radiation

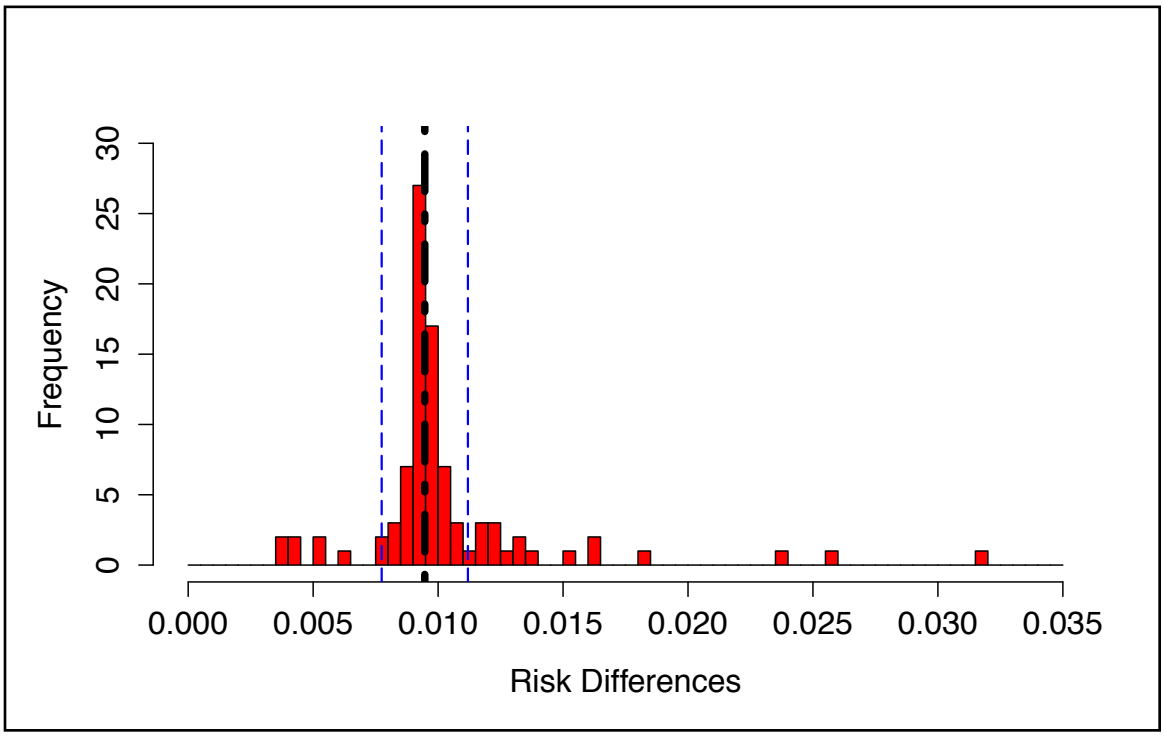
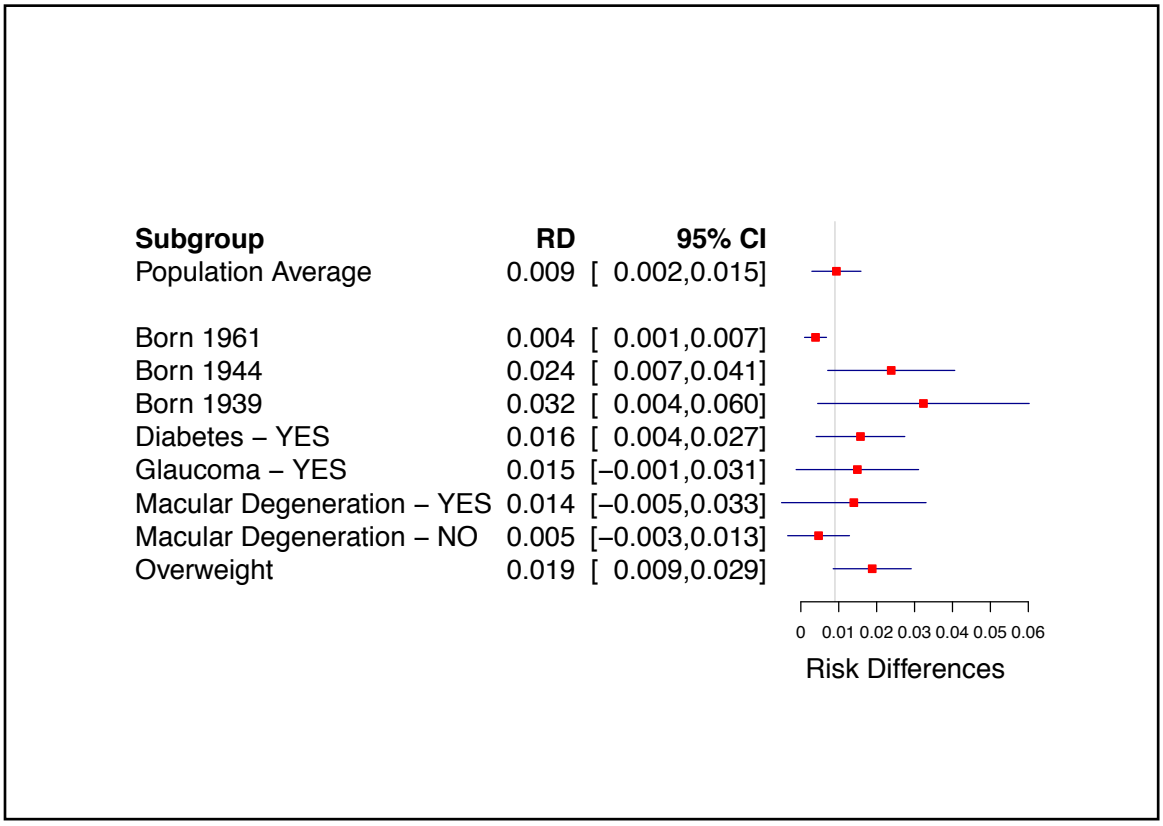


Figure 6. Forest plot of select stratum-specific risk differences for cataracts between high-dose (75th percentile of cumulative badge dose; 61.3 mSv) and low-dose (25th percentile of cumulative badge dose; 23.9 mSv) occupational ionizing radiation



Chapter 4

Manuscript 3. Exposures to occupational ionizing radiation in U.S. medical radiologic technologists and risks of cataracts: an application of the parametric g-formula

Introduction

Exposure to ionizing radiation is known to be associated with risks of cataracts.[4] Studies have shown cataract risk is associated with high-dose ionizing radiation in Hiroshima bomb survivors[23] and Chernobyl clean-up workers.[31] Evidence has also shown low-dose ionizing radiation is associated with cataracts among pilots,[28] cardiologists,[29] and industrial radiographers.[30] Medical radiologic technologists are exposed to protracted low-dose occupational ionizing radiation and may be at risk of developing work-related cataracts. The U.S. Rad Tech (USRT) Study constitutes the largest occupational cohort of medical radiologic technologists in the world.[5] With its extensive follow-up and recent improvements to the dosimetry system estimating annual ionizing radiation exposures,[61] the USRT cohort provided the unique opportunity to investigate risks of cataracts from low-dose occupational ionizing radiation exposures. In a previous manuscript, we estimated that ionizing radiation significantly increased the rate of cataract. However, the hazard ratio we estimated varied by follow-up duration (age) with larger effects at younger ages (earlier follow-up) that waned with greater age (longer follow-up). Exploratory analyses found component associations of the healthy worker survivor bias

were present and suggested this bias may have lead to the time-varying effect.

The healthy worker survivor effect (HWSE) has long been known as a common source of confounding in occupational health studies.[38–41] The HWSE was first recognized in 1885 by Ogle with the observation that individuals change job tasks, occupations, and leave employment based on health status.[42] Robins described the healthy worker effect as the result of time-varying confounding from work status that occurs when a participant leaves the work place as a result of previous exposure such that work status is both a confounder and an intermediate on the causal pathway.[43] Individuals may leave work as a result of their health status, return to work for periods of time, then leave again and so on, resulting in time-varying confounding. In studies with cumulative exposures, this manifests as a bias known as the HWSE, whereby healthy individuals are likely to accrue higher exposures as compared to less healthy individuals who may leave work due to their poor health. Thus, the HWSE can attenuate or even reverse the true relationship between the exposure and the health outcome.[72]

A directed acyclic graph (DAG) illustrating the HWSE is shown in Figure 7. Of note, W_1 (work status at time 1) is a confounder of the effect of X_1 (radiation exposure at time 1) on Y_1 (cataracts at time 1) in this DAG and, as such, we would like to control for it. However W_1 is also a collider because it is directly caused by other variables, namely U (health status) and X_0 (radiation exposure at time 0), and controlling for W_1 will induce collider-stratification bias.[51] If we were to condition on work-status, this bias (the HWSE) may induce a relationship between exposure and outcome by opening a back door path through U, an unmeasured confounder (e.g. health status). Because conventional methods (such as conditioning on work status in a regression model) are unable to address the

structural problem inherent in the HWSE, a family of analytical methods called g-methods were developed to adjust for this type of bias.[43] Recent studies have applied the parametric g-formula in occupational cohorts[40,113–115] to correct effect estimates for healthy worker survivor bias.

The parametric g-formula can be conceptualized as an extension of traditional epidemiologic standardization to treatments applied over many time points.[116] Though it is straightforward to conceptualize in this way, it can be challenging to implement the g-formula in practice. Parametric regression models are built that predict the outcome and relevant covariates. These models are used to estimate the predicted covariate value an individual would have under different hypothetical treatment regimes. This allows natural comparisons of "what would have happened" to the study population under different hypothetical interventions; for instance, a comparison of what would have happened if individuals had never received more than the annual regulatory threshold of ionizing radiation throughout their career relative to their actual experience. In the present study we use the parametric g-formula to adjust for possible bias from the HWSE when estimating the effect of occupational ionizing radiation treatment regimes among medical radiologic technologists on cataracts.

Methods

Study Population

The USRT cohort consists of 146,022 radiologic technologists certified by the American Registry of Radiologic Technologists for more than two years prior to 1982. A baseline survey was sent out to all living technologists in the years 1984–1989. A second

follow-up survey was sent in 1994–1998, with a third survey sent in 2004–2006 to participants that responded to the first and / or second survey and completed by 73,625 technologists. In the present study, participants were followed from age first worked as a technologist starting at age 18 or older until report of cataracts or administrative censoring at the completion of the third survey. Participants were dropped in the analyses if they had exposures before the age of 18 (N=2,439), report of cataracts age 18 or younger (N=51) or before year first worked as a technologist (N=15), had only one year of observation (N=6), reported never working as a medical radiologic technologist (N=597), had unlikely survey responses (N=3), asked to be removed from the study (N=1), or had missing information on cataracts (N=7) and year of diagnosis (N=651), occupational ionizing radiation exposures (N=5), date of completion of the third survey (N=6), or baseline covariates age, race, or birth year (N=46), for a final sample of 69,798 participants.

Cataracts

Cataracts were ascertained by self-report on the third survey. Participants were asked to report if they had been diagnosed with cataracts (YES / NO) and the year of first diagnosis. Information regarding the type of cataract (i.e. nuclear, cortical, posterior subcapsular, or mixed) was not collected on the survey. Age of diagnosis was calculated using reported birth year subtracted from year of cataract diagnosis.

Occupational Ionizing Radiation and Work Status

Recent enhancements were made to the dosimetry system in the USRT study and have been reported in detail elsewhere.[61,62] Briefly, the goal of the dosimetry system is to capture the plausible range of badge dose (effective dose equivalent, calculated from badge readers used to capture ionizing radiation exposure) and organ dose to twelve organs (absorbed doses, including the eye lens). Annual probability density functions of dose were developed for individuals using information from the three surveys regarding years and hours per week worked, personal protective characteristics (e.g. lead apron use, badge reader placement), types of facilities worked (e.g. hospital or clinic, military or civilian), individual badge dose readings where available, literature based badge dose for years preceding 1960, and historical technical specifications and regulatory dose limits. Using Monte Carlo simulation techniques, multiple dose realizations were estimated for each year of reported work as a technologist, with the arithmetic mean of all realizations reported as the final estimate of annual dose. The final estimates resulted in annual badge dose (mSv) and eyelens dose (mGy) for 110,000 technologists in the USRT cohort for each year reported working as a technologist. The present study uses badge dose estimates only for comparison to current regulatory standards of annual effective dose equivalent (the weighted sum of dose, source of radiation, and relative level of harm to tissue from radiation) as measured by badge dose. Participants' work status was determined by assigning a value of one to each year they reported working as a technologist, and zero otherwise.

Statistical Analysis

Data were collapsed from annual periods to five-year intervals, such that exposure and work history and cataracts were summarized for ages 18–20, 21–25, 26–30, 31–35, 36–40, 41–45, 46–50, 51–55, 56–60, 61–65, 66–70, 71–75, 76–80, 81–85 and 86–90. For example, if a participant reported a diagnosis of cataract at age 33, the outcome indicator would be assigned a value of one for cataracts in the age interval 31–35. For intervals without a report of cataracts or administrative censoring, the outcome indicator was assigned a value of zero. Similarly, if a participant reported working at any age during an interval they were assigned a value of one, and zero if they did not work at any age during that interval.

A complete description and theory of the parametric g-formula have been described elsewhere.[47,117] Briefly, the steps we implemented for the parametric g-formula were as follows. First, we fit a pooled linear regression model for 1) the level of occupational ionizing radiation exposure conditional on work status in that period, cumulative previous badge dose exposures, and baseline covariates (age, sex, birth year), a pooled logistic regression model for 2) the probability of working in a given period conditional on previous work status and cumulative ionizing radiation exposure up to the previous time and baseline covariates, and a pooled logistic regression model for 3) the probability of cataracts in a given period conditional on work status in that period, cumulative previous ionizing radiation exposure up to that period, and baseline covariates (see Appendix for details). Next, we created a pseudo-population of the same size as the original cohort with the same distribution of baseline covariates. For this pseudo-population, we create the follow-up experience of work status, badge dose exposure, and cataracts using parameter estimates from the fitted models (above). Finally, from the

outcome data in the pseudo population, we estimate the age-specific cumulative incidence of cataracts and age-specific risks of cataracts conditional on survival. Finally, we repeated this procedure for different treatment regimes (discussed below) and computed risk differences to compare the different regimes. The general steps in the algorithm are outlined below.

Steps in the parametric g-formula (for each exposure scenario):

1. Fit parametric models using either pooled linear or logistic regression for each component (see Appendix), using the observed data in the USRT cohort (*N* =69,798).
2. Create pseudo-population using the total sample size (*N*=69,798) such that the baseline covariate distributions of sex, race / ethnicity, and birth year are the same as the original sample.
3. Construct follow-up in order of work status, badge dose exposure, then cataracts on *N* pseudo-participants using the regression models fit in Step 1.
4. Compute the age-specific cumulative incidence risks and age-specific risks of cataracts conditional on survival.
5. Repeat steps 1–4 500 times, using bootstrapped resampling to refit the models in step 1 and allow estimation of 95% confidence intervals around effects in step (4).

First, as a check of the pooled linear and logistic regression models, we estimated the overall risk of cataracts using the observed entry and censoring times, and observed

exposure history data. Then we estimated the risk of cataracts by simulating the exposure history but still using the observed entry and censoring times. Under these conditions, we found the component models estimated similar risks of cataracts and radiation dose as observed in the original USRT cohort.

To implement the g-formula we considered several possible treatment regimes that corresponded to hypothetical interventions in the study population's radiation exposure. As a reference treatment regime, we estimated the cumulative and conditional risks of cataracts under the observed exposure to ionizing radiation; that is, we set no limits on the exposure level that was determined by the parametric models, and we allowed left truncation at the observed age of study entry. This is referred to as the "natural course" and used as the reference level for all other treatment regimes we consider. One possible intervention would be to restrict every worker to receive less than the regulatory limit. The United States Nuclear Regulatory Commission (USNRC) has determined the current annual regulatory limit of effective dose equivalent to 50 mSv per year or 250 mSv every five years.[118] Occupational ionizing radiation doses are significantly lower than the annual limit set by the USNRC for almost all exposures of the USRT cohort except for the oldest technologists working in the earliest years of study follow-up, with only 0.05% of the annual point exposures from almost 1.5 million measurements of badge being greater than 50 mSv. Hence, we did not consider that treatment regime. Instead, we considered treatment regimes of radiation dose thresholds using percentiles of the observed badge dose five-year point exposures for 1) 80^{th} percentile: 18.38 mSv, 2) 60^{th} percentile: 9.06 mSv, 3) 40^{th} percentile: 4.47 mSv, and 4) 20^{th} percentile: 2.08 mSv. For example, for the regime using the 80^{th} percentile of badge

dose, we set values of estimated badge dose that were above the threshold of 18.38 mSv to 18.38 mSv for that five-year period, and those that were below the threshold remained the same. This would correspond to a real-world policy intervention, whereby a technologist would not be allowed to accrue cumulative ionizing radiation exposure greater than 18.38 mSv over a five-year period. We then estimated the risks of cataracts with the adjusted dose estimates for that treatment regime. Similarly, we would repeat the analysis setting the threshold to the 60^{th} percentile of badge dose, such that any estimates of dose that were above the 60^{th} percentile, were set to 9.06 mSv while those below would remain the same, and so on for each treatment regime. Additionally, for comparison to the previous study in which we estimated the risk of cataracts for a 10-mSv increment of badge dose, we computed (as an approximation) the risk of cataracts under a scenario that reduced each badge dose estimated in each period of follow-up in the pseudo-population by 5 mSv (a 10-mSv increment of ionizing radiation dose over 10 years) with negative values set to a value of zero. Thus, we estimated risks of cataracts under the natural course and five different interventions of dose. For all scenarios, censoring due to dropout was not allowed and administrative censoring occurred at age 90. After estimating the age-specific cumulative risks and conditional risks of cataracts, we calculated risk differences of cataracts comparing each intervention separately to the natural course as reference. Finally, we repeated the analysis with 500 bootstrapped samples to construct 95% confidence intervals for each risk difference.

Results

Table 5 shows the cumulative incidence of cataracts for each scenario, and the risk differences and 95% confidence intervals of cataracts comparing each intervention of dose to the natural course for select intervals ages 20–90. In general, we see that differences in risks are initially modest but statistically significant at ages 35–40 and then increasing in magnitude with increasing age. Trends are similar for the treatment regimes limiting point exposures to the 80th percentile (badge dose $\leq$ 18.38 mSv), 60th percentile (badge dose $\leq$ 9.06 mSv), 40th percentile (badge dose $\leq$ 4.47 mSv), and 20th percentile (badge dose $\leq$ 2.08 mSv) of observed dose. Overall, the treatment regime reducing 5 mSv in dose at each age-period had smaller risk reductions as compared to the other treatment regimes with only a few significant differences. All cumulative incidence risk differences are illustrated in Figure 8. In Table 6 we present conditional risks of cataracts for the natural course and all treatment regimes and risk differences and 95% confidence intervals. Similar to the cumulative incidence risks, trends indicate modest differences in risks at earlier ages that increase in magnitude until the end of follow-up at age 90. Significant differences in cataract risks between treatment regimes and natural course begin at ages 35–50 until ages 75–80 for the treatment regimes limiting dose to the 60th percentile and 80th percentile. Significant differences in conditional risks of cataracts start at ages 35–50 for the 20th percentile and 40th percentile and continue until the end of follow-up at age 90. The treatment regime of a 5 mSv reduction in dose did not have any significant differences in the conditional risks compared to the natural course throughout all of follow-up. Figure 9 illustrates the conditional risk differences of cataracts between the natural course and each intervention of dose.

Discussion

The present study found substantively important differences in both cumulative incidence of cataracts and conditional risks of cataracts between the natural course and treatment regimes. There was evidence that decreasing the dose of radiation exposure could reduce the risk of cataracts, even at relatively early ages. This contrasts with the results in the previous analysis, which found significant hazards of cataracts at earlier ages that attenuated toward the null at later ages. Additionally, the differences in risk increased in magnitude and became quite substantial with increasing age. Confidence intervals of the effect estimates were rather precise, indicating little uncertainty about the sizes of these effects in cataract risk reduction. The present analysis appears to indicate that the risk of cataracts among medical radiologic technologists in the United States could be reduced if the threshold for a maximum allowable dose were changed.

The parametric g-formula has several assumptions to causally interpret parameter estimates some of which are 1) *exchangeability* – no unmeasured confounding, 2) *correct model specification* – models reflect the underlying biological model, 3) *positivity* – that there are exposed and unexposed at every combination of the confounders, and 4) *consistency* – that there are well-defined interventions of the exposure. These assumptions are not exclusive to the g-formula and in fact are true for the Cox regression models used in our previous analysis and indeed generally for all statistical modeling. We cannot rule out the possibility that there is unmeasured confounding in our study, which would affect the exchangeability assumption. Additionally, though the parametric regression models used in the g-formula may not be correctly specified, we have used the DAG of the proposed HWSE mechanism to guide

their specification. There is a lack of positivity due to the fact that there is zero probability of exposure when a technologist is not at work, precluding the use of marginal structural models. However, in a time-varying setting the g-formula with dynamic treatment regimes is more flexible regarding these types of violations to positivity.[43,119] The consistency assumption means that a person's counterfactual outcome is consistent with their observed outcome. To maintain consistency, the interventions must be well-defined such that an exposure level could be hypothetically assigned to another individual who has a different exposure level.[120] The consistency assumption is reasonable in the present study because we are able to manipulate the level of ionizing radiation in a manner that is supported (consistent) in the observed data. Another potential limitation of the g-formula is known as the g-null paradox, which states that under the null hypothesis, one is not able to correctly specify the correct parametric regression models used in the g-formula.[52] The result of the g-null paradox is that in large samples, the g-formula may reject the null hypothesis of no association between exposure and outcome (in our case, ionizing radiation and cataracts) even when no association exists. As it is already established there is an association between ionizing radiation and the development of cataracts,[20,22] we are less concerned with the g-null paradox in the present study.

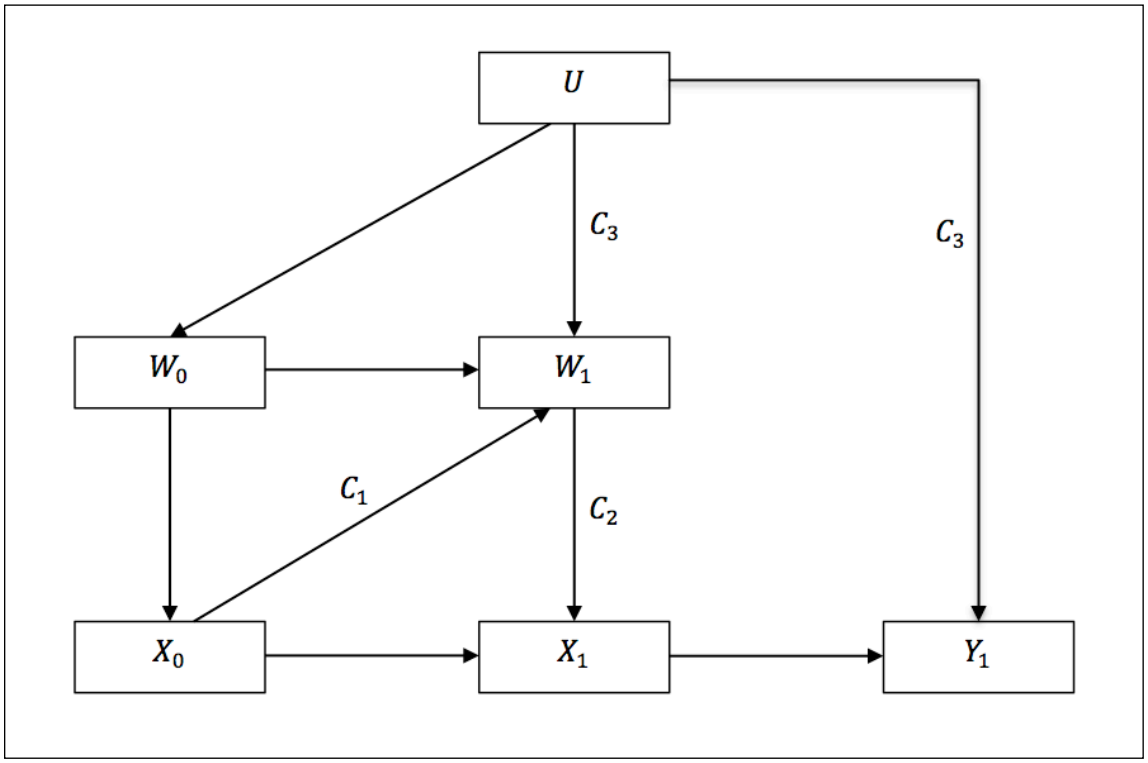
Subtypes of cataracts are characterized by their location in the eye lens and include posterior subcapsular, cortical, nuclear, and mixed phenotypes. Of these, ionizing radiation is most commonly associated with posterior subcapsular or cortical subtypes.[22–24] Ascertainment of cataracts in the USRT Study does not specify the subtype. In a previous manuscript, we reported the results of a literature review investigating age-

specific rates of cataracts by subtypes. The review found five population based studies.[66–70] Populations in these studies differed from medical radiologic technologists in that they were not exposed to protracted low–dose occupational ionizing radiation. We found that relative rates of the subtypes differed by age, with radiogenic subtypes (posterior subcapsular and cortical) predominant at earlier ages and age-related nuclear cataracts predominating with increasing age. We cannot rule out possible non-differential misclassification of cataracts in the present study. A large proportion of cataracts (40%) were diagnosed in the USRT cohort between ages 65–80, the age ranges where we found the largest significant differences in the rates of cataracts in the present study. It may warrant validation of cataracts and subtypes in the USRT cohort, to address the bias from misclassification in the effect estimates between ionizing radiation and cataracts.

Often, the goal of an occupational epidemiologic study is to estimate a causal effect between a workplace exposure and a health outcome. Effect estimates from the study's results may inform current occupational health standards. It is only meaningful to do this if it is possible to causally interpret the effect estimate in this relationship. However, for reasons of study design or other characteristics of the data, confounding bias can limit causal inference of effect estimates. Incomplete control of confounding is a well-recognized source of bias and must be accounted for when estimating effects of exposures.[38] In the present study we are seeking to estimate the effect of occupational ionizing radiation on the risk of cataracts in medical radiologic technologists. Previous analyses strongly suggested that a well-known source of confounding from selection bias (the HWSE) in occupational health epidemiologic investigations was present. As a result, we have implemented the parametric g-formula to adjust our effect estimates between

ionizing radiation and cataracts for bias stemming from the structural problem of no exposure when not at work and confounding of unmeasured health–status of the HWSE. Our results indicate that interventions (using percentiles of observed dose) reducing exposures of five-year occupational ionizing radiation attenuate the risks of cataracts starting at relatively earlier ages compared to no intervention (the status quo) of occupational exposure in the USRT Study cohort. As such, the medical radiologic technologists in the USRT cohort appear to have significantly elevated risk in the incidence of cataracts from protracted low-dose occupational ionizing radiation exposures.

Figure 7. Conceptual model of healthy worker survivor effect[a]

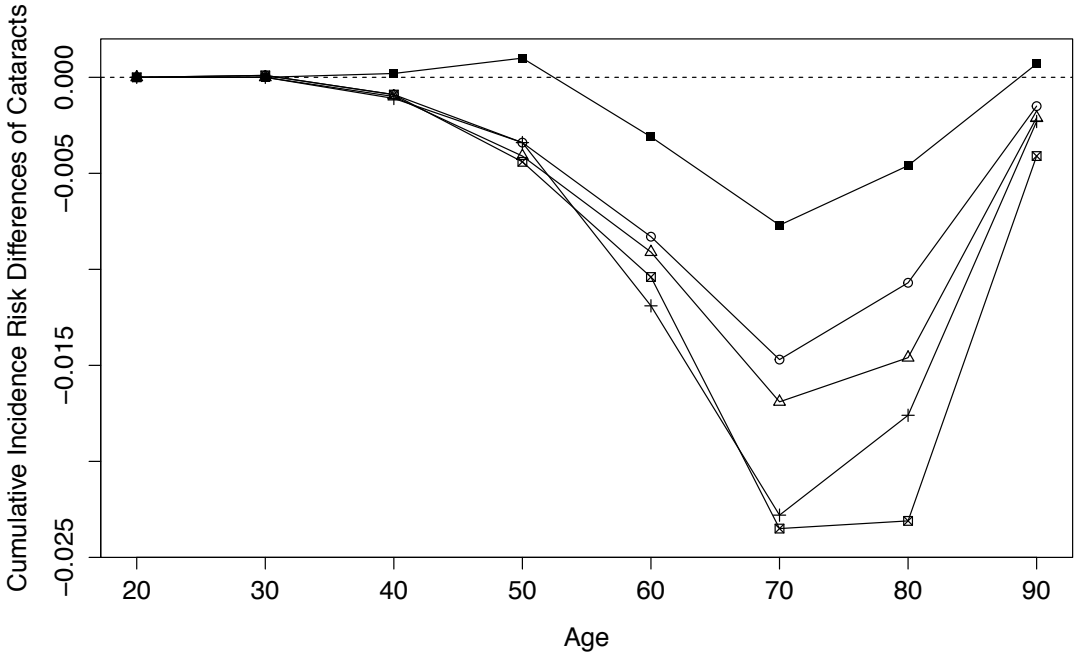


a Abbreviations: W_0=work status at time 0; W_1=work status at time 1; X_0=cumulative exposure up to time 0; X_1=cumulative exposure up to time 1; U=Unmeasured health status; Y=cataracts at time 1; C_1= component association 1: previous exposure relationship to subsequent work status; C_2=component association 2: work status relationship to subsequent exposure; C_3=component association 3: work status relationship to survival time. In this DAG, work status at time 1 is a collider, such that conventional methods conditioning on work status open a back door path between exposure and outcome through unmeasured health status and result in a biased estimate between occupational ionizing radiation and cataracts.

Table 5. Cumulative incidence of cataracts for simulated natural course of ionizing radiation badge dose (mSv), and simulated interventions of badge dose limited to ≤ 18.38 mSv, ≤ 9.06 mSv, ≤ 4.47 mSv, ≤ 2.08 mSv, and 5 mSv reduction in dose, and risk differences of cataracts and 95% confidence intervals comparing interventions to natural course, Ages 20–90, U.S. Rad Tech Study

	Risks: 18-20	RD [95% CI]	Risks: 55-60	RD [95% CI]
Natural Course	0.0000	ref	0.0903	ref
≤ 18.38 mSv	0.0000	0.0000 [-0.3894, 0.3888]	0.0820	-0.0083 [-0.0115, -0.0042]
≤ 9.06 mSv	0.0000	0.0000 [-0.3893, 0.3887]	0.0813	-0.0091 [-0.0131, -0.0055]
≤ 4.47 mSv	0.0000	0.0000 [-0.3897, 0.3891]	0.0784	-0.0119 [-0.0139, -0.0063]
≤ 2.08 mSv	0.0000	0.0000 [-0.3893, 0.3890]	0.0800	-0.0104 [-0.0146, -0.0066]
5 mSv reduction in Dose	0.0000	0.0000 [-0.3895, 0.3893]	0.0872	0.0031 [-0.0059, 0.0012]
	Risks: 25-30		**Risks: 65-70**	**RD [95% CI]**
Natural Course	0.0004	ref	0.3238	ref
≤ 18.38 mSv	0.0005	0.0001 [-0.0003, 0.0002]	0.3091	-0.0147 [-0.0200, -0.0072]
≤ 9.06 mSv	0.0004	0.0000 [-0.0003, 0.0003]	0.3069	-0.0169 [-0.0233, -0.0099]
≤ 4.47 mSv	0.0004	0.0000 [-0.0003, 0.0003]	0.3011	-0.0228 [-0.0265, -0.0125]
≤ 2.08 mSv	0.0005	0.0001 [-0.0003, 0.0002]	0.3004	-0.0235 [-0.0287, -0.0150]
5 mSv reduction in Dose	0.0005	0.0000 [-0.0003, 0.0003]	0.3161	-0.0077 [-0.0124, -0.0009]
	Risks: 35-40		**Risks: 75-80**	**RD [95% CI]**
Natural Course	0.0038	ref	0.7446	ref
≤ 18.38 mSv	0.0029	-0.0009 [-0.0015, -0.0001]	0.7339	-0.0107 [-0.0174, -0.0060]
≤ 9.06 mSv	0.0028	-0.0010 [-0.0016, -0.0001]	0.7300	-0.0146 [-0.0226, -0.0097]
≤ 4.47 mSv	0.0027	-0.0011 [-0.0017, -0.0002]	0.7270	-0.0176 [-0.0263, -0.0129]
≤ 2.08 mSv	0.0029	-0.0009 [-0.0016, -0.0001]	0.7215	-0.0231 [-0.0304, -0.0158]
5 mSv reduction in Dose	0.0040	0.0002 [-0.0008, 0.0007]	0.7400	-0.0046 [-0.0134, -0.0013]
	Risks: 45-50		**Risks: 85-90**	**RD [95% CI]**
Natural Course	0.0203	ref	0.9746	ref
≤ 18.38 mSv	0.0169	-0.0034 [-0.0056, -0.0021]	0.9731	-0.0015 [-0.0046, -0.0003]
≤ 9.06 mSv	0.0162	-0.0041 [-0.0059, -0.0022]	0.9725	-0.0162 [-0.0058, -0.0015]
≤ 4.47 mSv	0.0168	-0.0034 [-0.0059, -0.0024]	0.9723	-0.0034 [-0.0068, -0.0019]
≤ 2.08 mSv	0.0159	-0.0044 [-0.0063, -0.0026]	0.9705	-0.0044 [-0.0077, -0.0029]
5 mSv reduction in Dose	0.0212	0.0010 [-0.0025, 0.0011]	0.9753	0.0007 [-0.0036, 0.0003]

Figure 8. Age-specific cumulative incidence risk differences of cataracts comparing simulated interventions of occupational ionizing radiation exposures to natural course[a]

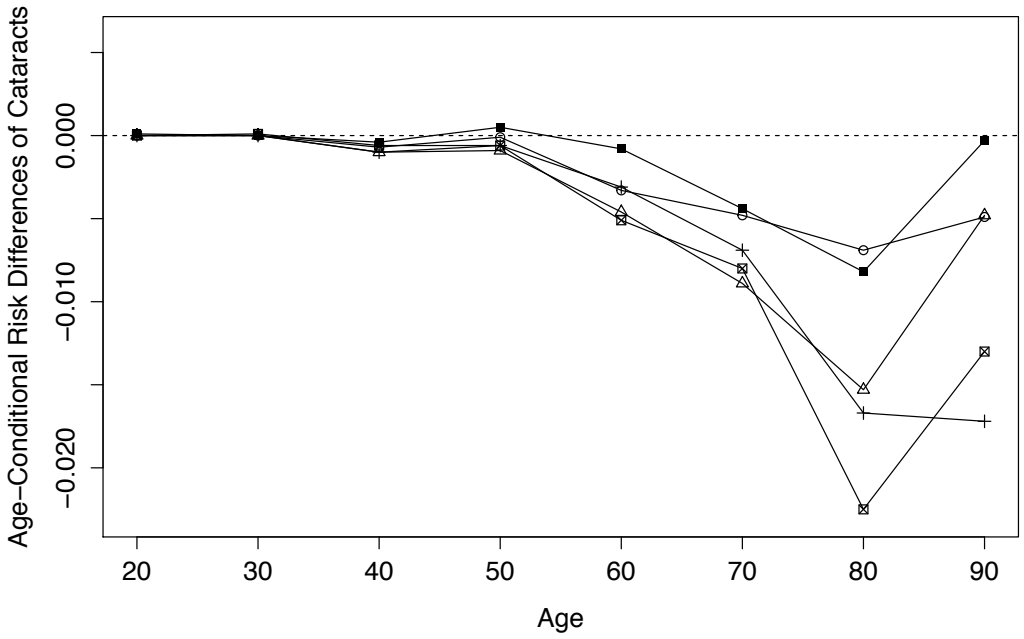


a ○=Risk differences of cataracts for ionizing radiation doses no larger than the 80th percentile of badge dose (18.38 mSv) compared to natural course; △=Risk differences of cataracts for ionizing radiation doses no larger than the 60th percentile of badge dose (9.06 mSv) compared to natural course; +=Risk differences of cataracts for ionizing radiation doses no larger than the 40th percentile of badge dose (4.47 mSv) compared to natural course; ⊠=Risk differences of cataracts for ionizing radiation doses no larger than the 20th percentile of badge dose (2.08 mSv) compared to natural course; ■=Risk differences of cataracts for 5 mSv reduction in badge dose for each five-year period of follow-up compared to natural course.

Table 6. Risks of cataracts conditional on survival for simulated natural course of ionizing radiation badge dose (mSv), and simulated interventions of badge dose limited to ≤ 18.38 mSv, ≤ 9.06 mSv, ≤ 4.47 mSv, ≤ 2.08 mSv, and 5 mSv reduction in dose, and risk differences of cataracts and 95% confidence intervals comparing interventions to natural course, Ages 20–90, U.S. Rad Tech Study

	Risks: 18-20	RD [95% CI]	Risks: 55-60	RD [95% CI]
Natural Course	0.0000	ref	0.0479	ref
≤ 18.38 mSv	0.0000	0.0000 [-0.0002, 0.0001]	0.0445	-0.0033 [-0.0048, -0.0003]
≤ 9.06 mSv	0.0000	0.0000 [-0.0001, 0.3876]	0.0433	-0.0046 [-0.0056, -0.0010]
≤ 4.47 mSv	0.0000	0.0000 [-0.0001, 0.3876]	0.0447	-0.0031 [-0.0059, -0.0013]
≤ 2.08 mSv	0.0000	0.0000 [-0.0001, 0.3876]	0.0428	-0.0051 [-0.0065, -0.0015]
5 mSv reduction in Dose	0.0001	0.0001 [-0.0001, 0.0001]	0.0471	-0.0008 [-0.0034, 0.0010]
	Risks: 25-30		**Risks: 65-70**	**RD [95% CI]**
Natural Course	0.0003	ref	0.1773	ref
≤ 18.38 mSv	0.0003	0.0000 [-0.0002, 0.0002]	0.1725	-0.0048 [-0.0100, -0.0011]
≤ 9.06 mSv	0.0003	0.0000 [-0.0002, 0.0002]	0.1684	-0.0089 [-0.0119, -0.0031]
≤ 4.47 mSv	0.0003	0.0000 [-0.0002, 0.0001]	0.1704	-0.0069 [-0.0143, -0.0043]
≤ 2.08 mSv	0.0003	0.0000 [-0.0002, 0.0001]	0.1693	-0.0080 [-0.0156, -0.0056]
5 mSv reduction in Dose	0.0003	0.0000 [-0.0002, 0.0002]	0.1729	-0.0044 [-0.0080, 0.0005]
	Risks: 35-40		**Risks: 75-80**	**RD [95% CI]**
Natural Course	0.0026	ref	0.4592	ref
≤ 18.38 mSv	0.0019	-0.0007 [-0.0011, -0.0001]	0.4523	-0.0069 [-0.0166, -0.0011]
≤ 9.06 mSv	0.0016	-0.0010 [-0.0012, -0.0001]	0.4439	-0.0153 [-0.0201, -0.0047]
≤ 4.47 mSv	0.0016	-0.0010 [-0.0012, -0.0003]	0.4425	-0.0167 [-0.0235, -0.0075]
≤ 2.08 mSv	0.0020	-0.0006 [-0.0012, -0.0002]	0.4367	-0.0225 [-0.0259, -0.0099]
5 mSv reduction in Dose	0.0023	-0.0004 [-0.0006, 0.0004]	0.4510	-0.0082 [-0.0135, 0.0013]
	Risks: 45-50		**Risks: 85-90**	**RD [95% CI]**
Natural Course	0.0103	ref	0.7464	ref
≤ 18.38 mSv	0.0102	-0.0001 [-0.0030, -0.0008]	0.7415	-0.0049 [-0.0234, 0.0062]
≤ 9.06 mSv	0.0094	-0.0009 [-0.0031, -0.0008]	0.7416	-0.0048 [-0.0267, 0.0023]
≤ 4.47 mSv	0.0097	-0.0006 [-0.0030, -0.0007]	0.7292	-0.0172 [-0.0300, -0.0002]
≤ 2.08 mSv	0.0097	-0.0006 [-0.0033, -0.0010]	0.7334	-0.0130 [-0.0301, -0.0017]
5 mSv reduction in Dose	0.0118	0.0005 [-0.0016, 0.0007]	0.7507	-0.0003 [-0.0190, 0.0110]

Figure 9. Risk differences of cataracts conditional on survival comparing simulated interventions of occupational ionizing radiation exposures to natural course[a]



a ○=Risk differences of cataracts for ionizing radiation doses no larger than the 80th percentile of badge dose (18.38 mSv) compared to natural course; △=Risk differences of cataracts for ionizing radiation doses no larger than the 60th percentile of badge dose (9.06 mSv) compared to natural course; +=Risk differences of cataracts for ionizing radiation doses no larger than the 40th percentile of badge dose (4.47 mSv) compared to natural course; ⊠=Risk differences of cataracts for ionizing radiation doses no larger than the 20th percentile of badge dose (2.08 mSv) compared to natural course; ■=Risk differences of cataracts for 5 mSv reduction in badge dose for each five-year period of follow-up compared to natural course.

Chapter 5

Discussion

In the dissertation, we found in *manuscript 1* a significant association between occupational ionizing radiation and the hazard of cataracts for both dose equivalent (badge dose, mSv) and absorbed dose (eye lens dose, mGy). This relationship varied significantly by age with the hazard ratio elevated at earlier ages and decreasing over time. Additionally, the relationship was statistically significant at earlier ages while the 95% confidence intervals of the hazard ratios contained the null value starting at ages > 30. A sensitivity analysis assessing the components of the healthy worker survivor effect (HWSE) indicated strong evidence that this type of bias was present in the U.S. Rad Tech (USRT) Study cohort. Specifically, component 1 (the association between previous exposure to subsequent work-status) showed a significant association between an increase in both 1-year and 10-year lagged cumulative badge dose and staying at work. Component 2, the association between work-status and subsequent exposure, existed a priori (technologists were only exposed to occupational ionizing radiation if they indeed worked in a given year). And component 3 (the association between work-status to survival time) indicated a significant association between leaving work and cataracts for the entire sample, the sub-sample of those with intermittent work status, and most strongly for the sub-sample of those who left work and did not return. A subsequent literature review investigating age-specific rates of the subtypes of cataracts (posterior subcapsular, cortical, and nuclear) found five population-based studies to identify if there

was a possibility for misclassification of radiogenic cataracts.[66–70] Although the populations in these studies differed from the USRT cohort in that they were not exposed to occupational ionizing radiation, they did reveal the potential for non-differential misclassification of cataracts in the present study. The analysis pooling data from the population studies indicated that cortical and posterior subcapsular cataracts (those most commonly associated with exposures to ionizing radiation) were predominant at earlier ages as compared to nuclear cataracts (the phenotype associated with increasing age). It could be the true effect was constant over time, and the increasing amounts of non-differential misclassification of cataracts at later ages could explain the decreasing association between ionizing radiation and cataracts. Although we can only speculate whether this distribution is different for radiologic technologists, it does introduce uncertainty into the classification of cataracts for the USRT cohort.

In *manuscript 2,* we presented methods for building a prediction model of cataracts in the USRT cohort using boosted regression trees, a type of ensemble machine learning analytical method. In this analysis, we restricted the cohort to ages 24–44 at baseline similar to a previous analysis by Chodick et. al.[60] Subtypes of cataracts commonly associated with ionizing radiation occur at younger ages,[10] and the restriction was chosen as information regarding subtypes was not available in the present study. We calculated risk differences of cataracts from high-dose (75th percentile badge dose, 61.31 mSv) to low-dose (25th percentile badge dose, 23.90 mSv) occupational ionizing radiation in subgroups of the USRT cohort. Elevated risks compared to the population average were found in those with diabetes, macular degeneration, glaucoma, BMI > 25 at

baseline, and those born in the earliest years of this restricted cohort. Those who were youngest at baseline and those without macular degeneration had lower risks compared to the population average. This method illustrated both the benefit of identifying potential subgroups where ionizing radiation may interact biologically to increase the risk of cataracts (useful for designing public health interventions) while also placing these risks in the context of other stratum-specific risks in the entire population.

In *manuscript 3,* we estimated age-specific risks of cataracts (both cumulative incidence risks and risks conditional on survival) for the natural course of ionizing radiation and several interventions limiting the level of dose to different percentiles of the observed absorbed dose (badge dose, mSv) or a 5-mSv reduction for each dose estimate in the USRT cohort. Overall, there was evidence that decreasing the dose of radiation exposure could reduce the risk of cataracts, even at relatively early ages. Trends were similar for the treatment regimes limiting dose to the 80^{th} percentile (badge dose $\leq$ 18.38 mSv), 60^{th} percentile (badge dose $\leq$ 9.06 mSv), 40^{th} percentile (badge dose $\leq$ 4.47 mSv), and 20^{th} percentile (badge dose $\leq$ 2.08 mSv) of observed dose, indicating significant risk reduction at ages 35–40 for cumulative incidence and conditional risks throughout the end of follow-up to age 90. This contrasts with the results found in *manuscript 1*, which found significant elevated risks of cataracts at younger ages compared to non-significant risks at older ages.

Overall, there appears to be elevated risk of cataracts in the USRT cohort from occupational ionizing radiation. This is supported by similar evidence of cataract risks from protracted low-dose occupational ionizing radiation exposures in industrial

radiographers[30] and cardiologists.[29,57] The characterization of this relationship is dependent on the method of analysis and study design. As we have highlighted, the effect estimates in *manuscript 1* may be affected by bias from the HWSE. To limit the effects of time-varying confounding and structural bias of work-status (and unmeasured confounding from health status), we have adjusted our estimates of cataract risks for the HWSE in *manuscript 3*. Although these estimates are not directly comparable (the contrasts in *manuscript 1* compare an increment change in dose while the contrasts in *manuscript 3* compare cataract occurrence under designated hypothetical treatment regimes), they do indicate that there are significant elevated risks of cataracts. The identification of certain subgroups with elevated risks of cataracts compared to the population average in *manuscript 2* illustrated the potential for synergistic effects (biologic interaction) of ionizing radiation and other health conditions (e.g. diabetes, macular degeneration, glaucoma). Though we cannot exclude the potential for reporting bias (i.e. those with cataracts may also be more likely to report other eye conditions such as glaucoma), these results provide context for characterization of the distribution of effects from ionizing radiation and the risks of cataracts.

Strengths and Limitations

One of the main strengths to our study is the size of the occupational cohort of radiologic technologists in the USRT Study and its extensive follow-up. Being the largest cohort of its kind, it is uniquely capable of investigating the risks of cataracts from protracted low-dose occupational ionizing radiation exposures. Additionally, recent

enhancements to the historical dose reconstruction[61] have provided a comprehensive dosimetry system that aims to estimate the true distribution of ionizing radiation exposure in the USRT cohort. The dosimetry system has provided estimates for both dose equivalent (badge dose, mSv) and absorbed dose (eye lens dose, mGy), allowing us to compare risks of cataracts from ionizing radiation using different measures of dose. The similar results found for both measurements add strength to the findings.

As we have discussed throughout the three manuscripts, cataracts were ascertained by self-report and the lack of specificity for the types of cataracts in the USRT cohort are a type of information bias and a limitation in the present study. As such, we are unable to determine with the present data how this lack of specificity impacts our current results. It is highly likely there is non-differential misclassification of cataracts (i.e. each participant has the same probability of misclassification of the outcome), and this has been proposed as a possible explanation for the time-varying effect of ionizing radiation and higher risks of cataracts at younger ages compared to older ages in *manuscript 1*. Although we were unable to account for this lack of specificity in the first manuscript, we chose to restrict our population in *manuscript 2* to potentially remove cataracts that would most likely be related to age and not ionizing radiation. Although this may have introduced another type of selection bias into the study, these analyses juxtaposed with those in the first manuscript represent a comprehensive evaluation of cataract risks with the data that is available.

We also noted through sensitivity analyses assessing the component associations of the HWSE, the results in *manuscript 1* are likely strongly affected by selection bias

from the HWSE. Although this is a limitation that is common to occupational epidemiologic investigations and is a type of selection bias, the application of the parametric g-formula in *manuscript 3* addresses this limitation. The quantitative assessment of the components of the HWSE justified the use of g-methods to correct effect estimates. The assessment is very useful as g-methods are computationally intensive, and should only be implemented when appropriate. Because the parametric g-formula is able to circumvent non-positivity (i.e. that there isn't a probability of exposure > 0 in all levels of confounders), we needed to use it as opposed to another g-method like marginal structural models that are unable to account for non-positivity. In the USRT cohort, technologists are only exposed when at work, and therefore have a zero probability of exposure when not at work, resulting in non-positivity. Marginal structural models use the inverse probability of treatment weight estimator, and when the probability is inverted the weight goes to infinity and is not estimable for a non-zero probability of exposure, thus precluding their use in the present study. It is important to note that although we have adjusted our estimates for the risks of cataracts from ionizing radiation in *manuscript 3* for any potential bias from the HWSE, these results may still be affected by the lack of specificity of cataract phenotype. A limitation of the parametric g-formula is the g-null paradox, where the null hypothesis of no association can be rejected in large samples even when no association exists. For this reason, it is not recommended to use the parametric g-formula for exploratory and hypothesis generating investigations, and instead is only recommended when there is a known association.[52] As there is a known association between ionizing radiation and cataracts[4] we are less concerned about

the g-null paradox. Regardless, the strength of this analysis is the use of rigorous computational methods grounded in counterfactual and causal inference theory.

Conclusion

In conclusion, the present study has provided a comprehensive evaluation of risks of cataracts from protracted low-dose occupational ionizing radiation exposures in medical radiologic technologists in the United States and its territories. Overall, our results indicate that low-dose ionizing radiation exposures elevate the risks of cataracts, as our three manuscripts found significant associations between occupational ionizing radiation and cataract risks. Additionally, methods were presented to explore heterogeneity of effects and improve the causal interpretation of effect estimates in the association between ionizing radiation and cataracts. Validation of cataracts is warranted and future studies would benefit from information regarding phenotypes of cataracts.

Bibliography

1. Stewart F a, Akleyev a V, Hauer-Jensen M, et al. ICRP publication 118: ICRP statement on tissue reactions and early and late effects of radiation in normal tissues and organs--threshold doses for tissue reactions in a radiation protection context. *Ann ICRP*. 2012;41(1-2):1–322. doi:10.1016/j.icrp.2012.02.001.
2. Thompson DE, Mabuchi K, Ron E, et al. Cancer incidence in atomic bomb survivors. Part II: Solid tumors, 1958-1987. *Radiat Res*. 1994;137(2 Suppl):S17–S67.
3. Adams MJ, Hardenbergh PH, Constine LS, Lipshultz SE. Radiation-associated cardiovascular disease. *Crit Rev Oncol Hematol*. 2003;45:55–75. doi:10.1016/S1040-8428(01)00227-X.
4. Kleiman NJ. Radiation cataract. *Ann ICRP*. 2011;41(3-4):80–97. doi:10.1016/j.icrp.2012.06.018.
5. Boice JD, Mandel JS, Doody MM, Yoder RC, McGowan R. A health survey of radiologic technologists. *Cancer*. 1992;69(2):586–598. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1728391.
6. Bushong SC. *Radiologic science for technologists: physics, biology, and protection*. Saint Louis, MO: Elsevier Mosby; 2013.
7. Mettler F a, Thomadsen BR, Bhargavan M, et al. Medical radiation exposure in the U.S. in 2006: preliminary results. *Health Phys*. 2008;95(5):502–507. doi:10.1097/01.HP.0000326333.42287.a2.
8. Turner JE. *Atoms, radiation, and radiation protection*. 3rd Editio. Wiley-VCH Verlag GmbH & Co.; 2007.
9. Unites States Nuclear Regulatory Commission. Part 20 - Standards for Protection Against Radiation. Available at: http://www.nrc.gov/reading-rm/doc-collections/cfr/part020/full-text.html#part020-1001.
10. Asbell PA, Dualan I, Mindel J, Brocks D, Ahmad M, Epstein S. Age-related cataract. *Lancet*. 2005;365:599–609.
11. Klein BEK, Klein R, Lee KE. Diabetes, Cardiovascular Disease, Selected Cardiovascular Disease Risk Factors, and the 5-year Incidence of Age-related Cataract and Progression of Lens Opacities: The Beaver Dam Eye Study. *Am J Ophthalmol*. 1998;126(6):782–790.
12. Leske MC, Wu SY, Hennis a, Connell a M, Hyman L, Schachat a. Diabetes, hypertension, and central obesity as cataract risk factors in a black population. The Barbados Eye Study. *Ophthalmology*. 1999;106(1):35–41.
13. Klein BEK, Klein R, Jensen SC, Linton K. Hypertension and Lens Opactities from the Beaver Dam Eye Study. *Am J Opthalmology*. 1995;119(March):640–646.
14. Klein BEK, Klein R, Lee KE. Cardiovascular Disease, Selected Cardiovascular Disease Risk Factors, and Age-related Cataracts: The Beaver Dam Eye Study. *Am J Ophthalmol*. 1997;123(March):338–346.
15. Cumming RG, Mitchell P, Leeder SR. Use of inhaled corticosteroids and the risk of cataracts. *N Engl J Med*. 2006;337(1):8–14.
16. Litt M, Kramer P, LaMorticella DM, Murphey W, Lovrien EW, Weleber RG.

Autosomal dominant congenital cataract associated with a missense mutation in the human alpha crystallin gene CRYAA. *Hum Mol Genet*. 1998;7(3):471–474. doi:10.1093/hmg/7.3.471.
17. Robman L, Taylor H. External factors in the development of cataract. *Eye (Lond)*. 2005;19(10):1074–82. doi:10.1038/sj.eye.6701964.
18. Eagling EM. Ocular damage after blunt trauma to the eye. Its relationship to the nature of the injury. *Br J Ophthalmol*. 1974;58(2):126–140.
19. West SK, Duncan DD, Muñoz B, et al. Sunlight exposure and risk of lens opacities in a population-based study: the Salisbury Eye Evaluation project. *JAMA*. 1998;280(8):714–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9728643.
20. Ainsbury E a, Bouffler SD, Dörr W, et al. Radiation cataractogenesis: a review of recent studies. *Radiat Res*. 2009;172(1):1–9. doi:10.1667/RR1688.1.
21. Luntz MH. Clinical Types of Cataract. In: Tasman W, Jaeger EA, eds. *Duane's Clinical Opthamology*. Volume 1. Philadelphia, PA: Lippincott & Company; 1992.
22. Otake M, Schull WJ. Radiation Cataract. *J Radiat Res*. 1991;32(Supplement):283–293.
23. Minamoto a, Taniguchi H, Yoshitani N, et al. Cataract in atomic bomb survivors. *Int J Radiat Biol*. 2004;80(5):339–45. doi:10.1080/09553000410001680332.
24. Chylack LT, Peterson LE, Feiveson AH, et al. NASA study of cataract in astronauts (NASCA). Report 1: Cross-sectional study of the relationship of exposure to space radiation and risk of lens opacity. *Radiat Res*. 2009;172(1):10–20. doi:10.1667/RR1580.1.
25. Hall P, Lundell M, Olsson K, Holm L-E. Lenticular opacities in individuals exposed to ionizing radiation in infancy. *Radiat Res*. 1999;152(2):190–195.
26. Hsieh WA, Lin I-F, Chang WP, Chen W-L, Hsu YH, Chen M-S. Lens opacities in young individuals long after exposure to protracted low-dose-rate gamma radiation in 60Co-contaminated buildings in Taiwan. *Radiat Res*. 2010;173(2):197–204. doi:10.1667/RR1850.1.
27. Cucinotta FA, Manuel EK, Jones J, et al. Space Radiation and Cataracts in Astronauts. *Radiat Res*. 2001;156(5):460–466.
28. Rafnsson V, Olafsdottir E, Hrafnkelsson J, Sasaki H, Arnarsson A, Jonasson F. Cosmic Radiation Increases the Risk of Nuclear Cataract in Airline Pilots: A Population-Based Case Control Study. *Arch Opthalmology*. 2013;123(August):1102–05.
29. Vano E, Kleiman NJ, Duran A, Rehani MM, Echeverri D, Cabrera M. Radiation cataract risk in interventional cardiology personnel. *Radiat Res*. 2010;174(4):490–5. doi:10.1667/RR2207.1.
30. Lian Y, Xiao J, Ji X, et al. Protracted low-dose radiation exposure and cataract in a cohort of Chinese industry radiographers. *Occup Environ Med*. 2015;72(9):oemed–2014–102772. doi:10.1136/oemed-2014-102772.
31. Worgul B V., Kundiyev YI, Sergiyenko NM, et al. Cataracts among Chernobyl Clean-up Workers: Implications Regarding Permissible Eye Exposures. *Radiat Res*. 2007;167(2):233–243. doi:10.1667/RR0298.1.
32. Rothwell PM. Subgroup analysis in randomised controlled trials: importance,

indications, and interpretation. *Lancet*. 2005;365(9454):176–186. doi:10.1016/S0140-6736(05)17709-5.
33. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Stat Med*. 2002;21(19):2917–2930. doi:10.1002/sim.1296.
34. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran P a., Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; *J Clin Epidemiol*. 2004;57(3):229–236. doi:10.1016/j.jclinepi.2003.08.009.
35. Hastie T, Tibshirani R, Friedman JH. Chapter 10: Boosting and Additive Trees. In: *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Second Edi. Springer Science+Business Media, LLC; 2009:337–387.
36. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks; 1984.
37. Maclin R, Opitz D. An Empirical Evaluation of Bagging and Boosting. *Fourteenth Natl Conf Artifical Intell*. 1997.
38. Hernán M a., Hernández-Díaz S, Robins JM. A Structural Approach to Selection Bias. *Epidemiology*. 2004;15(5):615–625. doi:10.1097/01.ede.0000135174.63482.43.
39. Naimi AI, Cole SR, Hudgens MG, Brookhart MA, Richardson DB. Assessing the component associations of the healthy worker survivor bias: occupational asbestos exposure and lung cancer mortality. *Ann Epidemiol*. 2013;23(6):334–41. doi:10.1016/j.annepidem.2013.03.013.
40. Cole SR, Richardson DB, Chu H, Naimi AI. Analysis of Occupational Asbestos Exposure and Lung Cancer Mortality Using the G Formula. *Am J Epidemiol*. 2013;177(9):989–996. doi:10.1093/aje/kws343.
41. Rao RS, Sigurdson AJ, Doody MM, Graubard BI. An application of a weighting method to adjust for nonresponse in standardized incidence ratio analysis of cohort studies. *Ann Epidemiol*. 2005;15(2):129–36. doi:10.1016/j.annepidem.2004.05.007.
42. Arrighi HM, Hertz-Piccioto I. The Evolving Concept of the Healthy Worker Survivor Effect Author (s): H . Michael Arrighi and Irva Hertz-Picciotto Published by : Lippincott Williams & Wilkins Stable URL : http://www.jstor.org/stable/3702361 . content in a trusted digital archive . We. *Epidemiology*. 1994;5(2):189–196.
43. Robins J. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chronic Dis*. 1987;40:139S–161S. doi:10.1016/S0021-9681(87)80018-8.
44. Hernán M a, Cole SR, Margolick J, Cohen M, Robins JM. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiol Drug Saf*. 2005;14(7):477–91. doi:10.1002/pds.1064.
45. Hernán M a, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11(5):561–70. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10955409.
46. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne J a C. Methods for

dealing with time-dependent confounding. *Stat Med*. 2013;32(9):1584–618. doi:10.1002/sim.5686.
47. Westreich D, Cole SR, Young JG, et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Stat Med*. 2012;31(18):2000–9. doi:10.1002/sim.5316.
48. Ahern J, Hubbard A, Galea S. Estimating the effects of potential public health interventions on population disease burden: a step-by-step illustration of causal inference methods. *Am J Epidemiol*. 2009;169(9):1140–7. doi:10.1093/aje/kwp015.
49. Taubman SL, Robins JM, Mittleman M a, Hernán M a. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *Int J Epidemiol*. 2009;38(6):1599–611. doi:10.1093/ije/dyp192.
50. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol*. 2011;173(7):731–8. doi:10.1093/aje/kwq472.
51. Greenland S. Quantifying Biases in Causal Models: Classical Confounding vs. Collider-Stratification Bias. *Epidemiology*. 2003;14(3):300–306.
52. Robins JM, Hern MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G, eds. *Longitudinal Data Analysis*. CRC Press; 2009:553–599.
53. Westreich D, Edwards JK, Cole SR, Platt RW, Mumford SL, Schisterman EF. Imputation approaches for potential outcomes in causal inference. *Int J Epidemiol*. 2015:dyv135. doi:10.1093/ije/dyv135.
54. Taylor HR. Epidemiology of age-related cataract. *Eye (Lond)*. 1999;13 (Pt 3b:445–8. doi:10.1038/eye.1999.119.
55. Blakely E a, Kleiman NJ, Neriishi K, et al. Radiation cataractogenesis: epidemiology and biology. *Radiat Res*. 2010;173(5):709–17. doi:10.1667/RRXX19.1.
56. Wilde G, Sjöstrand J. A clinical study of radiation cataract formation in adult life following irradiation of the lens in early childhood. *Br J Opthalmology*. 1997;81:261–266.
57. Ciraj-Bjelac O, Rehani MM, Sim KH, Liew HB, Vano E, Kleiman NJ. Risk for radiation-induced cataract for staff in interventional cardiology: is there reason for concern? *Catheter Cardiovasc Interv*. 2010;76(6):826–34. doi:10.1002/ccd.22670.
58. Rastegar N, Eckart P, Mertz M. Radiation-induced cataract in astronauts and cosmonauts. *Graefes Arch Clin Exp Ophthalmol*. 2002;240(7):543–7. doi:10.1007/s00417-002-0489-4.
59. Chumak V V., Worgul B V., Kundiyev YI, et al. Dosimetry for a Study of Low-Dose Radiation Cataracts among Chernobyl Clean-up Workers Dosimetry for a Study of Low-Dose Radiation Cataracts. *Radiat Res*. 2007;167(2):606–614. doi:10.1667/RR0298.1.
60. Chodick G, Bekiroglu N, Hauptmann M, et al. Risk of cataract after exposure to low doses of ionizing radiation: a 20-year prospective cohort study among US radiologic technologists. *Am J Epidemiol*. 2008;168(6):620–31.

doi:10.1093/aje/kwn171.
61. Simon SL, Preston DL, Linet MS, et al. Radiation Organ Doses Received in a Nationwide Cohort of U.S. Radiologic Technologists: Methods and Findings. *Radiat Res*. 2014;182(5):507–528. doi:10.1667/RR13542.1.
62. Simon SL. Organ-specific external dose coefficients and protective apron transmission factors for historical dose reconstruction for medical personnel. *Health Phys*. 2011;101(1):13–27. doi:10.1097/HP.0b013e318204a60a.
63. NASA. Total Ozone Mapping Spectrometer (TOMS) Data Overview. Available at: http://disc.sci.gsfc.nasa.gov/acdisc/TOMS. Accessed October 4, 2015.
64. Freedman DM, Kimlin MG, Hoffbeck RW, Alexander BH, Linet MS. Multiple indicators of ambient and personal ultraviolet radiation exposure and risk of non-Hodgkin lymphoma (United States). *J Photochem Photobiol B*. 2010;101(3):321–5. doi:10.1016/j.jphotobiol.2010.08.001.
65. Korn EL, Graubard BI, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *Am J Epidemiol*. 1997;146(6):528–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9290515.
66. Leske MC, Connell A, Wu S-Y, Hyman L, Schachat A. Prevalence of Lens Opacities in the Barbados Eye Study. *Arch Opthalmology*. 1997;115(1):105–111.
67. Congdon N, West SK, Buhrmann RR, Kouzis a., Muñoz B, Mkocha H. Prevalence of the different types of age-related cataract in an African population. *Investig Ophthalmol Vis Sci*. 2001;42(11):2478–2482.
68. Tsai SY, Hsu WM, Cheng CY, Liu JH, Chou P. Epidemiologic study of age-related cataracts among an elderly Chinese population in Shih-Pai, Taiwan. *Ophthalmology*. 2003;110(6):1089–1095. doi:10.1016/S0161-6420(03)00243-4.
69. Varma R, Torres M. Prevalence of lens opacities in Latinos: The Los Angeles Latino Eye Study. *Ophthalmology*. 2004;111(8):1449–1456. doi:10.1016/j.ophtha.2004.01.024.
70. Tan AG, Wang JJ, Rochtchina E, Mitchell P. Comparison of age-specific cataract prevalence in two population-based surveys 6 years apart. *BMC Ophthalmol*. 2006;6(17). doi:10.1186/1471-2415-6-17.
71. Li C, Sung E. A review of the healthy worker effect in occupational epidemiology. 1999;49(4):225–229.
72. Picciotto S, Brown DM, Chevrier J, Eisen E a. Healthy worker survivor bias: implications of truncating follow-up at employment termination. *Occup Environ Med*. 2013;70:736–42. doi:10.1136/oemed-2012-101332.
73. Chen W-L, Hwang J-S, Hu T-H, Chen M-S, Chang WP. Lenticular Opacities in Populations Exposed to Chronic Low-Dose-Rate Gamma Radiaton from Radiocontaminated Building in Taiwan. *Radiat Res*. 2001;156(1):71–77.
74. Nakashima E, Neriishi K, Minamoto A. A reanalysis of atomic-bomb data, 2000–2002 : a threshold analysis. *Health Phys*. 2004;90(2):154–160.
75. Leske MC, Chylack LT, He Q, Wu S, Schoenfeld E, Friend J. Risk Factors for Nuclear Opalescence in a Longitudinal Study. *Am J Epidemiol*. 1998;147(1):36–41.
76. Leske MC, Wu S, Nemesure B, Hennis A. Risk Factors for Incident Nuclear

Opacities. *Opthalmology*. 2002;109(7):1303–1308.
77. Buckley JP, Keil AP, McGrath LJ, Edwards JK. Evolving Methods for Inference in the Presence of Healthy Worker Survivor Bias. *Epidemiology*. 2015;26(2):204–212. doi:10.1097/EDE.0000000000000217.
78. Stayner L, Steenland K, Dosemeci M, Hertz-Picciotto I. Attenuation of exposure-response curves in occupational cohort studies at high exposure levels. *Scand J Work Environ Health*. 2003;29(4):317–324. doi:10.5271/sjweh.737.
79. Kirkeleit J, Riise T, Bjørge T, Christiani DC. The healthy worker effect in cancer incidence studies. *Am J Epidemiol*. 2013;177(11):1218–24. doi:10.1093/aje/kws373.
80. Chevrier J, Picciotto S, Eisen E a. A comparison of standard methods with g-estimation of accelerated failure-time models to address the healthy-worker survivor effect: application in a cohort of autoworkers exposed to metalworking fluids. *Epidemiology*. 2012;23(2):212–9. doi:10.1097/EDE.0b013e318245fc06.
81. Greenland S. Interactions in epidemiology: relevance, identification, and estimation. *Epidemiology*. 2009;20(1):14–7. doi:10.1097/EDE.0b013e318193e7b5.
82. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Stat Med*. 1983;2(2):243–251.
83. Feinstein A. The problem of cogent subgroups: A clinicostatistical tragedy. *J Clin Epidemiol*. 1998;51(4):297–299. doi:10.1016/S0895-4356(98)00004-3.
84. Friedman JH, Meulman JJ. Multiple additive regression trees with application in epidemiology. *Stat Med*. 2003;22(9):1365–81. doi:10.1002/sim.1501.
85. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. *Proc 23rd Int Conf Mach Learn - ICML '06*. 2006:161–168. doi:10.1145/1143844.1143865.
86. Leske MC, Wu S, Nemesure B, Li X, Hennis A, Connell A. Incidence and progression of lens opacities in the Barbados Eye Studies. *Ophthalmology*. 2000;107(7):1267–1273. doi:10.1016/S0161-6420(00)00155-X.
87. Bouffler S, Ainsbury E, Gilvin P, Harrison J. Radiation-induced cataracts: the Health Protection Agency's response to the ICRP statement on tissue reactions and recommendation on the dose limit for the eye lens. *J Radiol Prot*. 2012;32(4):479–88. doi:10.1088/0952-4746/32/4/479.
88. Oliva MS, Taylor H. Ultraviolet radiation and the eye. *Int Ophthalmol Clin*. 2005;45(1):1–17. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15632523.
89. Colitz CMH, Bomser JA, Kusewitt DF. The Endogenous and Exogenous Mechanisms for Protection From Ultraviolet Irradiation in the Lens. *Int Ophthamology Clin*. 2005;45(1):141–155.
90. Breiman L. Technical note: Some properties of splitting criteria. *Mach Learn*. 1996;24:41–47. doi:10.1007/BF00117831.
91. Colombet I, Ruelland a, Chatellier G, Gueyffier F, Degoulet P, Jaulent MC. Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. *Proc AMIA Symp*. 2000:156–160.
92. Breiman L. Heuristics of instability and stabilization in model selection. *Ann Stat*.

1996;24(6):2350–2383. doi:10.1214/aos/1032181158.
93. Schapire R. The strength of weak learnability. *Mach Learn*. 1990;5(2):197–227.
94. Freund Y, Schapire RE, Avenue P. A Short Introduction to Boosting. *J Japanese Soc Artif Intell*. 1999;14(5):771–780.
95. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an applicaion to boosting. *Comput Learn Theory*. 1995;904:23–37.
96. Friedman BJ, Hastie T, Tibshirani R. Additive Logistic Regression: A Statistical View of Boosting. *Ann Stat*. 2000;28(2):337–407.
97. Schmid M, Hothorn T. Flexible boosting of accelerated failure time models. *BMC Bioinformatics*. 2008;9(269):1–13. doi:10.1186/1471-2105-9-269.
98. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. *Proc 23rd Int Conf Mach Learn - ICML '06*. 2006:161–168. doi:10.1145/1143844.1143865.
99. Buhlmann, P. TH. Boosting Algorithms: Regularization, Prediction and Model Fitting. *Stat Sci*. 2007;22(4):477–505.
100. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *J Anim Ecol*. 2008;77(4):802–13. doi:10.1111/j.1365-2656.2008.01390.x.
101. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;38(4):367–378. doi:10.1016/S0167-9473(01)00065-2.
102. Ridgeway G. Generalized Boosted Models : A guide to the gbm package. 2009;(4):1–12. Available at: http://cran.r-project.org/package=gbm.
103. Hijmans RJ, Phillips S, Leathwick J, Elith J. dismo: Species distribution modeling. 2012.
104. Team RC. R: A Language Environment for Statistical Computing. 2015.
105. Localio a. R, Margolis DJ, Berlin J a. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *J Clin Epidemiol*. 2007;60(9):874–882. doi:10.1016/j.jclinepi.2006.12.001.
106. Muller CJ, Maclehose RF. Estimating predicted probabilities from logistic regression: Different methods correspond to different target populations. *Int J Epidemiol*. 2014;43(3):962–970. doi:10.1093/ije/dyu029.
107. Kiuchi Y, Yokoyama T, Takamatsu M, et al. Glaucoma in atomic bomb survivors. *Radiat Res*. 2013;180(4):422–30. doi:10.1667/RR3273.2.
108. Hernan MA, Brumback B, Robins JM. Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments. *J Am Stat Assoc*. 2001;96(454):440–448.
109. Greenland S. Chapter 5: Causal inference as a prediction problem: Assumptions, identification and evidence synthesis. In: *Causality: Statistical Perspectives and Applications*.; 2012:43–58.
110. Rothman K, Greenland S, Walker A. Concepts of Interaction. *Am J Epidemiol*. 1980;112(4):467–470.
111. Schonlau M. Boosted Regression (boosting): An introductory tutorial and a Stata plugin. *Stata J*. 2005;5(3):330–354. doi:The Stata Journal.
112. StataCorp. Stata Statistical Software. 2013.
113. Keil AP, Edwards JK, Richardson DB, Naimi AI, Cole SR. The Parametric g-

Formula for Time-to-event Data: Intuition and a Worked Example. *Epidemiology*. 2014;25(6):889–97. doi:10.1097/EDE.0000000000000160.
114. Edwards JK, McGrath LJ, Buckley JP, Schubauer-Berigan MK, Cole SR, Richardson DB. Occupational Radon Exposure and Lung Cancer Mortality. *Epidemiology*. 2014;25(6):829–834. doi:10.1097/EDE.0000000000000164.
115. Neophytou AM, Picciotto S, Costello S, Eisen E a. Occupational Diesel Exposure, Duration of Employment, and Lung Cancer. *Epidemiology*. 2015:1. doi:10.1097/EDE.0000000000000389.
116. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*. 2003;14(6):680–6. doi:10.1097/01.EDE.0000081989.82616.7d.
117. Taubman SL, Robins JM, Mittleman M a, Hernán M a. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *Int J Epidemiol*. 2009;38(6):1599–611. doi:10.1093/ije/dyp192.
118. Unites States Nuclear Regulatory Commission. *Standards for Protection Against Radiation*. Washington D.C.; 1991.
119. Young JG, Hernán M a., Robins JM. Identification, Estimation and Approximation of Risk under Interventions that Depend on the Natural Value of Treatment Using Observational Data. *Epidemiol Method*. 2014;3(1):1–19. doi:10.1515/em-2012-0001.
120. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*. 2009;20(1):3–5. doi:10.1097/EDE.0b013e31818ef366.

Appendix

Notation

Age is the time scale of interest from ages 18–90 collapsed into *j*=15 time periods (18–20, 21–25, 26–30, 31–35, 36–40, 41–45, 46–50, 51–55, 56–60, 61–65, 66–70, 71–75, 76–80, 81–85, 86–90) and *S* denotes the period of study entry. Y_k is defined as an indicator for cataracts at age-period *j*=*k*. W_k is defined as an indicator for work status at age-period *j*=*k*. $\bar{W}_{k-1}$is a 1-period lag of the history of work status. $X_k$is the ionizing radiation badge dose at age-period *j*=*k*. $\bar{X}_{k-1}$is a 1-period lagged variable of the cumulative exposure history up to each age-period *j*=*k*. V is a vector of baseline covariates including sex (female / male), race / ethnicity (White / Other), and birth year. $\bar{C}_k$ is defined as an indicator for censoring due to loss to follow-up at age-period *j*=*k*. Histories of time-varying covariates are denoted using an overbar such as work status history through age-period *j*=*k* is denoted $\bar{W}_k = \{W_S, W_{S+1}, \ldots, W_k\}$. The order assumed for each age-period *k* is W_k, X_k, Y_k.

Parametric g-formula for ionizing radiation and cataracts

Note, for total population N=69,798, we suppress the notation for individual i=1,… , N. In equation 1 for the parametric g-formula, the cumulative incidence of cataracts for each age-period *j* = 1, 2, … , 15 is represented as a function of the probability of cataracts and the joint distribution of exposure, working status, and baseline covariates. Because this quantity cannot be computed non-parametrically we use parametric regression models to approximate the quantities.

$$I(j) = \sum_{k=1}^{j} \sum_{v} \sum_{\bar{w}_j} \sum_{\bar{x}_j} \left\{ \Pr(Y_k = 1 | \bar{W}_k = \bar{w}_k, \bar{X}_k = \bar{x}_k, \bar{Y}_{k-1} = \bar{C}_k = 0, V = v, S \leq k) \times \right.$$

$$\left. \prod_{m=1}^{k} \begin{bmatrix} \Pr(C_m = 0 | V = v, \bar{W}_m = \bar{w}_m, \bar{X}_m = \bar{x}_m, \bar{Y}_{m-1} = \bar{C}_{m-1} = 0, S \leq m-1) \times \\ f(X_m = x_m | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{X}_{m-1} = \bar{x}_{m-1}, \bar{Y}_{m-1} = \bar{C}_{m-1} = 0, S \leq m) \times \\ Pr(W_m = 1 | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{X}_{m-1} = \bar{x}_{m-1}, \bar{Y}_{m-1} = \bar{C}_{m-1} = 0, S \leq m) \times \\ \Pr(Y_{m-1} = 0 | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{X}_{m-1} = \bar{x}_{m-1}, \bar{Y}_{m-2} = \bar{C}_{m-1} = 0, S \leq m-1) \times \\ \mathrm{f}(V = v) \end{bmatrix} \right\} \quad \textbf{(1)}$$

Under the treatment regimes where $X_k$is intervened on as specified in the intervention density (f^{int}) by replacing sampled exposures above the specified treatment regime with the specified limit, and no loss to follow-up (e.g. $\Pr(C_m = 0) = 1$) the quantity above is replaced with the formula in equation 2.

$$I(j) = \sum_{k=1}^{j} \sum_{v} \sum_{\bar{w}_j} \sum_{\bar{x}_j} \left\{ \Pr(Y_k = 1 | \bar{W}_k = \bar{w}_k, \bar{X}_k = \bar{x}_k, \bar{Y}_{k-1} = \bar{C}_k = 0, V = v, S \leq k) \times \right.$$

$$\left. \prod_{m=1}^{k} \begin{bmatrix} 1 \times \\ f^{int}(X_m = x_m | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{X}_{m-1} = \bar{x}_{m-1}, \bar{Y}_{m-1} = \bar{C}_{m-1} = 0, S \leq m) \times \\ Pr(W_m = 1 | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{X}_{m-1} = \bar{x}_{m-1}, \bar{Y}_{m-1} = \bar{C}_{m-1} = 0, S \leq m) \times \\ \Pr(Y_{m-1} = 0 | V = v, \bar{W}_{m-1} = \bar{w}_{m-1}, \bar{X}_{m-1} = \bar{x}_{m-1}, \bar{Y}_{m-2} = \bar{C}_{m-1} = 0, S \leq m-1) \times \\ \mathrm{f}(V = v) \end{bmatrix} \right\} \quad \textbf{(2)}$$

Model specification in the parametric g-formula

Model 1: Work Status

Log(odds(W_k)) = β_0 + β_1*I(W_{j-1}) + β_2*($\sum_{j=1}^{k-1} X_j$) + β_3*Time + β_4*Time^2 + β_5*Sex + β_6*Race + β_7*(Birth Year) + β_8*(Birth Year)^2

A pooled logistic regression model of work indicator at age-period *j*, given work indicator at time *j*-1, the sum of previous exposure to time *j*-1, time (*j*=1,…,15), a quadratic term of time, sex (female / male), race / ethnicity (White / Other), birth year, and a quadratic term of birth year.

Model 2: Radiation Exposures

$E(\log(X_k)) = \beta_0 + \beta_1*(\sum_{j=1}^{k-1} X_j) + \beta_2*\text{Time} + \beta_3*\text{Time}^2 + \beta_4*\text{Sex} + \beta_5*\text{Race} + \beta_6*(\text{Birth Year}) + \beta_7*(\text{Birth Year})^2$

A pooled linear regression model of log-transformed Badge dose at time *j* (model of exposure only for years worked as a radiologic technologist), the sum of previous exposure to time *j-1*, time (*j*=1,…,15), a quadratic term of time, sex (female / male), race / ethnicity (White / Other), birth year, and a quadratic term of birth year.

Model 3: Cataracts

$\text{Log}(\text{odds}(Y_k)) = \beta_0 + \beta_1*I(\text{Work}_j) + \beta_2*(\sum_{j=1}^{k-1} X_j) + \beta_3*\text{Time} + \beta_4*\text{Time}^2 + \beta_5*\text{Sex} + \beta_6*\text{Race} + \beta_7*(\text{Birth Year}) + \beta_8*(\text{Birth Year})^2$

A pooled logistic regression model of cataract indicator at age-period *j*, given work-status indicator at time *j*, the sum of previous exposure to time *j*-1, time (*j*=1,…,15), a quadratic term of time, sex (female / male), race / ethnicity (White / Other), birth year, and a quadratic term of birth year.

Monte Carlo simulation and risks estimation

We fit the three parametric models using the observed dataset. We then recreate the follow-up (works–status, exposure, and cataracts) histories for a new pseudo-population using N=69,798, the original sample size, with the same baseline covariate distributions of sex, race / ethnicity, and birth year. Generally, the order of estimation is work-status, exposure, and cataracts history. For all $j=S$, assign a value of 1 for work-status. For $j>S$, estimate the probability of working in a given j using the fitted parameters of the parametric regression models. Then, using the predicted probability of work-status, draw the work-status indicator from a Bernoulli distribution. Next, if a participant works, then estimate the predicted mean exposure level for that age-period. Next, using the predicted exposure mean and mean-squared error from the linear regression model for radiation exposures draw from the Normal distribution and exponentiate for the estimated exposure level. Next, estimate the probability of cataracts in period j using the parameters of the fitted pooled logistic regression model for cataracts and the recreated exposure history. Finally, using the predicted probability of cataracts, draw the cataract indicator from a Bernoulli distribution. There is no loss-to-follow-up, and all individuals are followed until report of cataract or age 90.

Treatment Regimes

Under each treatment regime, we set an exposure limit L. For each estimated exposure in a given age-period X_k that exceeds L, we replace the sampled exposure with the value of the exposure limit L. All values of estimated exposure that are below L remain the same.

The simulated treatment regime then uses these values of exposure for the subsequent predictions of work-status, exposures, and cataracts.

Cumulative Incidence and Conditional Risks Estimation

Using the Monte Carlo simulated dataset of the new pseudo-population, we estimated the cumulative incidence of cataracts up to each selected age-period j, using the following formula where the numerator is the number of cataracts for $S \leq j$, and the denominator is the number of study participants with $S \leq j$:

$$\frac{\#\ Cataracts}{\#\ at\ risk}$$

Similarly, we estimated the conditional risks of cataracts for each selected age-period j using the same formula where the numerator is the number of cataracts that occurred in j, and the denominator contains all participants that have survived to j and not had cataracts preceding j.

To estimate the 95% confidence intervals, we repeated the entire process 500 times by sampling from the observed data with replacement, re-fitting the parametric regression models, and estimating the cumulative incidence and conditional risks of cataracts, and calculating the risk differences between the natural course and the treatment regimes. The upper and lower confidence limits of the risk differences were estimated by using the 2.5^{th} and the 97.5^{th} percentiles of the risk difference distribution from the 500 bootstrap samples.