**Evaluating Bias Caused by Screening in Observational Risk-factor Studies of Lung Cancer Nested in the PLCO Randomized Screening Trial**

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Ricky Jeffrey Jansen

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

Dr. Timothy R. Church

September 2009

**Acknowledgements**

I would like to extend my gratitude to the members on my committee for their patience, guidance, and insiteful comments. My doctoral advisor, Dr. Timothy Church, provided support for my research development, nurturing me and providing invaluable direction. Dr. Bruce Alexander taught me to always think about the public health implication of my research. Dr. George Maldonado instilled in me a desire to do the best public health studies possible. Dr. Melanie Wall taught me about the complexity of model variable interactions.

I send my thanks to the Department of Environmental Health Sciences for providing me with a strong education background and the economic means to achieve my higher education goals. The department has assembled a group of faculty members that understand and emphasize the core objectives of environmental health science. The staff was incredibly helpful and willing to assist me with any administrative paperwork or questions. I would like to specifically thank Carol Hansen for her help with various forms and scheduling issues throughout my time at the University of Minnesota.

I want to express my appreciation for everyone that was involved with the Prostate, Lung, Colorectal, and Ovarian Cancer screening trial and its supporting organization, NCI. Without trial participants and the hard work of numerous researchers and support staff, this work would not have been possible.

I would like to extend a special thanks to my wonderful wife, Laura, for sharing this journey with me and providing great support, motivation, and feedback along the way.

## Dedication

This dissertation is dedicated to my family for their love, support, and encouragement.

## Abstract

It is well-known that bias such as lead-time and length distort studies of screening efficacy whether survival or incidence is of interest. A third bias, usually called overdiagnosis bias, occurs when an individual is only diagnosed with disease before death from a different cause because he/she is screened. These forms of bias can also arise in observational studies where the proportion screened and screening rates vary by risk-factor strata. This difference in screening behaviors influences corresponding case ascertainment or case enrollment probabilities which can lead to erroneous conclusions about the size of the risk-factor effect on the disease. It has been suggested that classic confounding occurs in such risk-factor studies when screening is efficacious; therefore, it can be addressed by conventional analyses such as stratification or confounder adjustment in regression models. However, even if the test is not efficacious, screening creates changes in case ascertainment probabilities which must be addressed using alternative methods. Recurrence-time models, long used for screening programs, can be adapted to model the affect screening use has on risk-factor studies. These models can be used to study the magnitude of potential bias, but may also be adapted to provide an analytic approach to correct estimates for such bias. The risk-factor studies nested in the PLCO trial are potentially affected by such bias, and this randomized study also provides a structure within which models of screening bias may be tested and validated. To validate our model, a variety of nested case-control studies will be developed that measure the effect smoking has on lung cancer and the degree to which the bias affecting those estimates change based on the study design will be determined. This process will include a) expanding a previously developed lead-time bias model to incorporate length and overdiagnosis; b) incorporating a more flexible and realistic model of screening that can incorporate the patterns documented in the PLCO trial; c) exploring if the mathematical model is valid using varied nested study designs within PLCO and comparing resulting logistic regression estimates to simulated results; and d) using the validated models to produce correction factors for use in other nested risk-factor studies. Results indicate that the mathematical model is highly sensitive to overdiagnosis as increasing rates increase expected bias, but relatively insensitive to using different screening test sensitivities. Increasing screening behavior differential during the study, preclinical duration length, and selecting from the intervention group are associated with increasing expected screening bias. Increasing screening behavior before the study and selecting from the usual-care group are associated with a decreasing expected screening bias. Although the mathematical model couldn't be validated as a correction factor here, the results suggest using a shorter preclinical duration distribution for the model may produce more accurate screening bias values. The focus of this work was to identify if chest x-ray screening could modify the estimated risk of smoking on lung cancer diagnosis. An additional goal was to develop a usable method for adjusting observational studies of lung cancer for the bias arising from differential chest x-ray screening between ever and never smoking groups. In a boarder sense, this work has provided an explanation of the effect screening use may have on an observational risk-factor study and an example of how to implement the mathematical technique. Additionally, this

project has provided a more general method for doing sensitivity analyses on the screening related assumptions involved with these studies, whether nested in a randomized trial or sampled from the population at large.

**Table of Contents**

# List of Tables

# List of Figures

# Chapter 1: Introduction

Most literature regarding cancer screening has focused on screening efficacy where the goal is to identify the survival/mortality benefit of using a specific screening test or screening program (intended screening effect). However, the focus of this research is to look at the neglected issue of how screening can bias risk-factor studies specifically identifying how screening changes the observed association between a risk factor and the disease (unintended screening effect) and how to theoretically account for such changes.

The current practice in observational risk-factor studies has been to ignore any bias caused by screening or at the very most to treat screening as a confounder and to control for it in a traditional way. An objective of this research is to evaluate the assumption that screening bias is minimal or unimportant by developing a mathematical model which can provide the expected effect that screening has on study results. The intention is to make a model which is transparent and easily modified by user and reader alike. Following this idea, all model assumptions will be identified and sensitivity analysis for parameter values in the model will be performed so the reader can draw his/her own conclusions about their plausibility.

Evaluation of the possible effect screening has on the risk-factor-disease association can be conducted by modifying a previously developed mathematical model. The modified model will compare the outcome of interest (e.g., incidence) evaluated under specific screening and no screening theoretical situations and will yield the potential amount of screening bias affecting the risk estimate of the observational study, allowing for theoretical correction of the observed estimate. In theory, applying the bias correction to the observed risk estimate will provide a more accurate representation of the effect the risk-factor has on the disease in the specific population during a specific time period.

Data from the Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer randomized trial will be used to investigate the potential (i.e., theoretical) biasing effect chest x-ray screening has on the smoking-lung cancer risk estimate calculated in several nested case control studies. Although the smoking-lung cancer association and chest x-ray screening is the main example throughout this proposal, the concepts can be applied more generally. Chest x-ray screening could be replaced by any other form of early detection and lung cancer could be replaced with any other disease for which a progressive disease model and the form of early detection are plausible. A goal of this project is to inform researchers about the potential ways screening can affect risk-factor studies and to provide a method to address such bias. Ultimately, demonstrating the importance of this screening issue will create awareness and may encourage a movement toward the practice of presenting more valid, reliable, and comparable observational studies.

## *Specific Aims*

There are three main hypotheses for this project: 1) Can the use of screening significantly bias risk-factor estimates in observational studies? (addressed through aims 1 and 2 below); 2) How do various case-control design choices influence the theoretical amount of bias caused by screening use? (addressed through aims 2 and 3 below); 3) Can a reliable correction method be developed to adjust for potential bias caused by the use of screening in already completed risk-factor studies? (addressed through aim 4 and 5 below)

The specific aims used to address these questions are:

1) Modify an existing recurrence-time model for screening bias due to lead-time in order to incorporate length and overdiagnosis bias.
2) Simulate the effects of realistic cancer screening patterns in observational studies of cancer incidence nested within the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial using the model developed in aim 1.
3) Design and analyze a variety of nested case-control studies of the risk smoking has on the development of lung cancer
4) Investigate how these estimates and bias change based on the differential frequency of screening among the smoking strata within each design.
5) Validate the simulation model by comparing its results with the observed results of the various case-control studies.

## *Background and Significance*

In 1969, Zelen and Feinleib (1) introduced a way of conceptualizing the disease process called the progressive disease model in which the process of disease development moves systematically through three stages (Figure 1). The first stage is the disease-free stage into which a subject is born. The second stage is the pre-clinical stage in which the development of disease has begun but there are no symptoms. This stage during which the disease is screen-detectable begins with disease onset. The third stage is the clinical stage in which a person is symptomatically diagnosed with disease and eventually dies of the disease. All disease advances through these stages in sequential order unless interrupted by screening, or death from another cause. If the subject is screened during the preclinical detectable interval (the date of disease onset to date of symptomatic detection), the date of disease diagnosis is shifted from the theoretical symptomatic detection date to an earlier date of screen detection (2).

There has been uneasiness with the assumption that all preclinical disease progresses to clinical disease with several researchers providing methods to use in the absence of this assumption(3). Modifications to this original description include splitting the preclinical state into two stages, early and advance, where early stage has the potential to be cured by treatments(4, 5). The use of a Markov model with 4 stages (disease free, preclinical, clinical, and terminal) has also been suggested where a person has the potential to recover during preclinical or clinical states(6). We will use the simpler progressive disease assumption of Zelen and Feinleib in our mathematical model as our intention at this stage is to demonstrate the possible impact of screening use on an observational risk-factor estimate and not to provide the most accurate model for lung cancer disease history.

Several researchers have identified ways to estimate sojourn time or detectable preclinical period in the study population(7, 8) or within the screened group(9),but most require additional assumptions. Suggested sojourn time distribution are based on exponential(10), piecewise density function(11), study data(12-14), or a convolution of the preclinical incidence and the density of time in the preclinical phase(15). Additional studies have focused on identifying how age can affect the sojourn time distribution(8, 13, 16). In the literature, simulations have even been performed to validate some of these added assumptions about sensitivity and sojourn time(17, 18). The mathematical model used here will assume that existing registry incidence data can be adjusted to represent the preclinical incidence distribution and the method will be discussed in more detail in the following chapter.

Historically, studies have been the designed to determine the efficacy of screening tests as interventional tools where screening tests are designed to identify early stage disease during the sojourn time (19, 20). The probability of detecting the disease when disease is present (i.e. sensitivity) during this period is an important quantity when estimating the effect of screening on the study population(21). Age has been shown to effect sensitivity(8, 13, 22, 23), and the transition probabilities between stages of disease(23). Therefore, a sensitivity analysis will be completed on the sensitivity parameter and the preclinical incidence function be age dependent in the mathematical model described in further detail in the next chapter.

Unfortunately in the pursuit to determine the efficacy of cancer screening tests, several forms of bias were identified that were directly related to the screening tests themselves. Methods for adjusting these biases include the use of recurrence time distributions (1, 24, 25) and microsimulation models (MISCAN) (26) to theoretically identify modifications in observed survival time (27) and alterations in clinical course of disease(28). Additionally numerous authors have taken a more applied route to estimating lead-time bias using study data such as the number of cancers detected at the successive screenings and the number of cancers occurring in the time interval between the screening examinations when known(21, 29) or employing dependence between lead-time and post-lead-time survival(30).

In risk-factor studies of cancer, the exposures are not randomized thereby increasing the potential bias related to differential screening use between the risk-factor strata of interest and its affect on study validity. In fact, this issue can even arise in observational studies nested within randomized trials. In the example used here, the case-control studies nested within the PLCO trial were designed to estimate the risk smoking has on lung cancer incidence in this population. This population has undergone chest x-ray screening both prior to and during the trial, suggesting the risk estimates may possibly be affected by three types of screening related biases: lead-time, length, and overdiagnosis bias.

Throughout this proposal, these three types of bias will be collectively referred to as *screening bias*, which we generally define as the disproportional effect early detection has on the relationship between the risk factor strata and outcome of interest. Screening can alter the apparent causal association between the risk factor and disease where, for example, the true effect that smoking has on lung cancer diagnosis may be masked by the relationship between smoking and receiving chest x-ray. That is, if the risk of lung cancer incidence is artificially increased among smokers in a group of subjects who have been screened, the apparent positive association of smoking and lung cancer in reality may be in part a result of the association between smoking and chest x-ray.

One type of screening bias that may affect study results is called *Lead-time bias.* Lead-time (depicted by $t_C$ - $t_S$ in Figure 1) is the theoretical interval between the time a case is screen-detected and the time it would have been diagnosed symptomatically. Lead-time causes the survival time of screened compared to non-screened patients to appear longer by advancing the date of diagnosis to an earlier date by the length of the lead-time interval. When this lead-time interval overlaps either the beginning or end of the case ascertainment period (e.g. person is screen-detected before the beginning of the study, but would have been diagnosed systematically during the study), this can lead to different case-ascertainment among screened and unscreened participants (31). In an observational risk-factor study when different screening patterns exist between the strata, the lead-time interval influences case selection differently thereby modifying the observed risk estimate.

*Length bias* also plays an important role when studying either the incidence and survival of a disease. Length bias arises because as the preclinical duration of a case increases so does its chance of being screen-detected(32-34). If the preclinical duration is correlated with expected survival from the time of clinical diagnosis then the average survival time for screen-detected cases will be longer than that of clinically diagnosed cases, even in the absence of a screening benefit. For example, imagine a hypothetical situation where there are two forms of the disease with 6 patients each and the only difference between the forms is the progression rate of disease. Let's assume that disease form 1 is rapidly progressing and that form 2 has a slow progressing disease, we can see that form 2 produces an additional 2 cases solely based on a longer preclinical duration (Figure 2). This extended preclinical duration leads to an average overall longer observed survival time among the slowly progressing disease versus rapidly progressing disease cases, thus biasing the characteristics of the group of individuals identified in the study as screen-detected cases. In an observational risk-factor study when the members of one stratum screen more often than another, that stratum's case population will be comprised of a greater proportion of slower progressing disease types. Therefore, the study population will not represent an accurate illustration of the target population in terms of stratum disease association or relative disease relationship between strata.

*Overdiagnosis bias* occurs when a preclinical disease that would not symptomatically or clinically present before death is detected by screening. Overdiagnosed cases have extremely slow progressing disease (i.e., very long preclinical stage) allowing for extremely long lead-time intervals. Overdiagnosis can arise in two situations (Figure 3): 1) when the offset due to lead-time is essentially infinite, because $t_c$ never occurs, or 2) when a specific group of subjects have a preclinical duration long enough to delay clinical surfacing until the subject's death from some other cause. Both scenarios create a screen-detected case with no clinically surfacing counterpart. The existence of this bias in observational risk-factor studies further increases the discrepancies discussed in the previous two paragraphs and in the observed risk estimate.

In observational studies, screening has the potential to change case ascertainment probabilities through these mechanisms. In order to deal with this screening bias in lung cancer risk-factor studies, we examined the difference in chest x-ray screening behavior between smoking strata, and implemented the approach described in the next chapter section "Research Design and Methods". The bias evaluated in this way can theoretically be used as a correction factor, either to directly adjust the approximated smoking-lung cancer *RR* when the parameters are well known, or as a component of a sensitivity or uncertainty analysis when uncertainty surrounds the parameters. Before describing the model and using it to evaluate screening bias, we identify what research has previously been performed.

## *Prior Research/Preliminary Studies*

A 1999 article by Church(31) presented a new form of lead-time bias that affects case-ascertainment. When there are time-restrictions placed on the ascertainment period of a study, individuals that are screened could be shifted into or out of the study. For example, if a case that would be symptomatically detected during the study is screen-detected before the start of the study, that case will be left out of the study. Similarly, if a case that would be symptomatically detected after the study is screen-detected during the study, it will be enrolled in the study.

To evaluate this bias quantitatively in a screening efficacy study, Church developed a mathematical model based on recurrent time screening models with the progressive disease assumption. The model included constant sensitivity for the screening test, a log normal distribution for the preclinical duration function, and preclinical incidence and survival functions based on SEER registry data. This model was used to evaluate the possible bias that may exist in the case-control mortality study of colon cancer screening. The results stated that when the true *OR* is 1 under the null hypothesis, its estimate is 0.5 to 0.75 under plausible assumptions. The author's concluding remarks indicated this bias has the potential to significantly alter interpretations of case-control cancer screening studies; therefore sensitivity analysis should be done to increase confidence in study conclusions.

The assumption made to simplify the model was that screening has no association with other causes of death or preclinical incidence and that screening provides no benefit. Because the focus was on lead-time bias, preclinical incidence time, preclinical duration, and survival time were assumed to be independent in order to eliminate length bias in survival times. Additional assumptions were that there is an accessible incidence rate distribution that represents the target population during the target time period in the absence of screening, and that the chosen preclinical incidence and duration distributions are representative of the true unobservable distributions. These are all reasonable assumptions given the illustrative purpose of the paper.

Screening can also bias studies of cancer risk factors when screening behaviors (i.e., proportion of the population that is screened and the screening rate among those screened) vary by risk-factor level. The different screening patterns expressed by each stratum create differences in case ascertainment probabilities, which in turn can change the observed size of the association between the risk-factor and disease. Weiss(35) and Joffe(36) have suggested that classic confounding occurs in risk-factor studies when screening is efficacious; therefore, it can be addressed by conventional analyses such as stratification or including the confounders in regression. However, even when the test is not efficacious, screening bias can occur. Since these biases are more akin to ascertainment bias, adjustments must use alternative methods. To begin to quantitatively address screening bias in risk-factor studies, we have modified the previously mentioned lead-time bias model(31) to make it applicable to observational risk-factor studies and applied it to a case-control study of prostate cancer.

The approach was as follows. Imagine that a study is conducted to investigate a risk factor for prostate cancer. In this hypothetical investigation of smoking and prostate cancer, we assumed that the screening proportions and rates were different among smokers and non-smokers. The model is parameterized for the natural history of prostate cancer (with preclinical incidence and duration functions) in the target population and combined with the screening behaviors (rate and proportion) and screening sensitivity from our sampled population. In this model, if various plausible distributions for the preclinical incidence and duration are assumed with varying length for the case ascertainment period, a sense of the potential magnitude of the lead-time bias present is obtained. In Table 1, it is apparent that under the aforementioned assumptions the bias can be large, and the impact of the screening rate differential is greater than either the case ascertainment period (a) or the modal preclinical duration (b). In other words, it becomes very important to model the effects of lead-time bias in studies when a large screening differential exists between the strata of the risk factor of interest, with more significant bias arising with a shorter ascertainment period based on assumptions in this study.

Using the Minnesota and Wisconsin Prostate Cancer Study (MWPCS), a population based case-control study, as a simulation example, we have demonstrated a possible range for such bias within the smoking and physical activity risk factors. A primary goal of the MWPCS ([National Cancer Institute grant 1R01CA074103-01A2]) was to examine the associations between prostate cancer and farming and pesticide exposure, as well as examine other possible risk factors such as smoking, physical activity, alcohol consumption history, and medical history. Cases were obtained from the Minnesota and Wisconsin state cancer registries; controls were selected from each state's driver license and identification card databases for the year 2000 and frequency matched to cases on 1-year age intervals. The simulations were based on parameters estimated from the 1665 controls in order to avoid the potential confounding of prostate cancer detection with screening frequency seen among cases. Results are presented using a case-ascertainment period of 2 years, as was used in Wisconsin; simulation results using a case-ascertainment period of 1.5 years, as in Minnesota, were similar.

In order to understand the effect of lead-time bias on case-ascertainment in the MWPCS and provide theoretical corrections for its risk factor, a mathematical model was applied using parameters that represented the MWPCS population. To isolate the relative lead-time bias between risk-factor strata, simulations incorporated a joint null hypothesis of no effect of the risk factor or PSA screening on the preclinical incidence of prostate cancer. Under these conditions, the true *OR* is 1 and any systematic deviation represents bias. The two risk factors of interest used in separate simulations were cigarette smoking in those aged 40-79 ("smoking") and average hours per week of total physical activity after age 50 in those aged 50-79 ("total physical activity").

The first set of simulations examined the smoking variable stratified as "ever smoked" vs "never smoked"; the second set examined the average total physical activity variable stratified as "3 or more hours per week" vs "less than 3 hours per week." The total physical activity variable assessed moderate activities including fast walking, baseball, and volleyball to strenuous activities such as running, jogging, and football. These simulations were developed and run using the program Mathcad® 12 (Cambridge, MA)(37) and preclinical incidence was determined from an analyses of SEER data using R version 2.1.1(38), which was also used for the graphics.

A key modification to the previous model involved making functions age dependent and stratifying by risk factor level, incorporating differential screening patterns for each stratum. The questionnaire data on PSA use was used to define the corresponding age-specific screening proportions (Figure 4a and c) and rates (Figure 4b and d) for each risk-factor stratum; we used a constant sensitivity ($\xi$) of 0.89 for the screening test(39). Figures 5b and d illustrate a function for the rate of screening among those who received a screen by age in each risk-factor stratum. The study participants were asked, "How many times were you screened between 1990 and 1998." This information was then used to calculate the screening rate by using the the number of screens divided by 9. Figure 4b illustrates the screening rate function in the ever and never smoking strata and Figure 4d shows the rate for the group that reported an average of > or = 3 hours of physical activity per week and a group that reported an average of < 3 hours of physical activity per week. After applying the model over the age-specific incidence in each risk-factor stratum, the incidence ratio expected under the null hypothesis obtained through simulation was used to approximate a theoretical correction for the biased risk factor-disease estimate observed in the study.

In most situations when the study disease is rare, as with cancer, the *OR* is used to approximate a *RR* with the equation $OR_{\text{estimated}} = (a/b)/(c/d) \approx (a/(a + b))/(c/(c + d)) = RR_{\text{approximated}}$ where in a two by two table "a" represents the number of cases with the risk factor, "b" represents the number of noncases with the risk factor, "c" represents the number of cases without the risk factor, and "d" represents the number of noncases without the risk factor. The simulated *RR* (i.e., $RR_{\text{correction}}$) can be thought of as the ratio between the observed incidence rates in two risk-factor strata (e.g., smoking vs. nonsmoking, or average ≥ 3 hours of physical activity per week vs. < 3 hours per week). Under the joint null hypothesis of no association between the risk factor and disease and no association between screening and the disease, the unbiased *RR* between the strata would equal 1 with any deviation representing screening bias. Since $\text{RR}_{\text{correction}}$ is the expected ratio of the number of events seen in one stratum to those in the other under the null hypothesis, to theoretically correct $RR_{\text{approximated}}$, we multiply the incident cases in the denominator stratum (i.e., *c*) by $RR_{\text{correction}}$ (i.e., amount of simulated lead-time bias). The result is a more valid approximated *RR* under the null hypothesis which assumes that the bias affects cases proportionately and that no other forms of bias are present in the study.

The conditions in the MWPCS (i.e., short study duration and long preclinical duration for prostate cancer) are such that considerable bias can arise. In the simulations, the rate of cases shifted into and out of the ascertainment period due to screening is almost equal to the rate of cases expected in the ascertainment period in the absence of screening. Because of the implementation of PSA screening in this population, the cases observed in the ascertainment period are a different subgroup than the cases expected in the absence of screening and more importantly, present a different composition of cases than expected in the target population. Thus, even after screening is taken into account, study results should be generalized with caution.

As with all statistical analyses and simulations, it is important to consider the assumptions that underlie the procedure and decide whether they are appropriate. In our study, two major assumptions involve the age-specific preclinical incidence rate and the preclinical duration. The incidence distribution from the SEER 9 registry for the years 1973-1986 (before PSA screening was prevalent) was assumed, after shifting it by the mean of the preclinical distribution, to reasonably represent the preclinical incidence for the MWPCS population in the absence of screening. The illustrative nature of this project (interest in determining the plausibility that lead-time bias can significantly affect observational risk-factor estimates) provides justification of this assumption. There are a wide range of estimated preclinical durations for prostate cancer in the literature(40), and as a result several combinations of plausible parameters (with mode 1, 5, 10, and 20; standard deviation 1, 4, 8) were used for the preclinical distribution. The true preclinical duration for prostate cancer is unknown and probably varies based on disease and patient characteristics. Thus, presenting a range of values is more informative than presenting a single, very likely incorrect value.

In the article it was proposed that any observational risk-factor study may be affected by the voluntary use of screening in the population of interest. In fact, any factor that is correlated with changes in the date of disease diagnosis can have a similar effect. Lead-time bias increases as differential proportion screened and screening rate increase between strata of the risk factor. Simulating the MWPCS with the presented model under plausible assumptions, it was found that the observed risk estimate for the total physical activity variable in the age group 50-59 may be biased by up to 23 percent by lead-time. However, the simulations yielded lower levels of bias in the older age group and in both age groups of the smoking variable, in which bias caused the observed risk estimate to appear slightly smaller than it would in the absence of PSA screening. Thus when early detection methods are suspected to influence case-ascertainment in a risk-factor study, it is important to examine all risk-factors correlated with screening for their stratum-specific screening proportions and rates. If the screening pattern varies by stratum apply a model such as that described to evaluate the possible effect bias has on the observed risk estimate.

The logical next step, the goal and topic of this doctoral work, is to build upon these illustrative studies to create a more valid and accurate mathematical model in which to evaluate the effect screening has on observed study results. The model will be enhanced to quantify the amount of screening bias (from lead-time, length, and overdiagnosis) affecting the relationship between risk factor and disease. As an example, smoking will be the risk factor, lung cancer will be the disease, and chest x-ray will be the method of screening. The example case-control studies will be nested within the PLCO randomized trial. This nesting will provide a structure within which the enhanced mathematical model of screening bias may be tested and validated.

The 23-year PLCO randomized trial was designed so approximately 37,000 men and 37,000 women would be screened for lung, colorectal, prostate (men only), and ovarian (women only) cancers and outcomes of interest would be compared to an equal sized cohort of usual care subjects. Subjects aged 55-74 would be followed for minimum of 13 years to ascertain outcomes of interest with primary focus being disease-specific mortality reduction due to screening. At baseline, demographic characteristics, known risk-factors for studied cancers, and screening history were collected from all participants(41, 42).

For this thesis since the disease of interest is lung cancer the focus of the text will be on the effect chest x-ray use has on the observed smoking-lung cancer relationship. Investigation of this issue will begin in Chapter 2 where the mathematical model and its components which are used to evaluate screening bias will be described. The data source used here, the PLCO randomized trial, will be described further in chapter 3 and used as an illustrative example of how to parameterize the bias model. This example will also determine the possibility of screening bias being a concern in observational lung cancer studies designed within the PLCO trial. . In Chapter 4, the potential effect of the different parts of an observational study design (e.g., cohort selection, case-ascertainment length, and study years involved) on screening bias will be explored through the simulation of several nested case-control study designs. The study design simulation and regression results will be used in a comparison method to explore validation of the mathematical model which will be detailed in chapter 5.

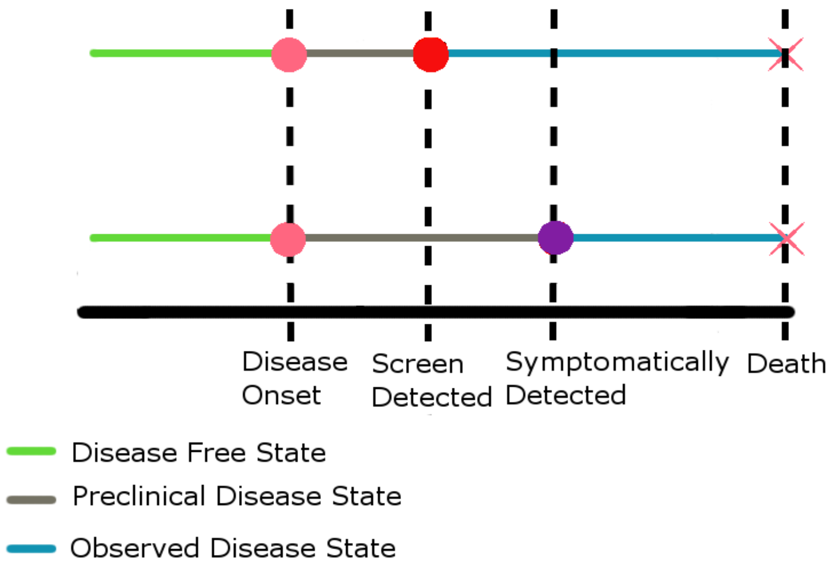



**Figure 1**. Diagram of a progressive disease model with overlaid recurrence-time model showing how screening changes the date of diagnosis of the disease and thus apparent survival. Cases pass through three states in their lifetime: disease free (from birth to date of detectable disease onset), preclinical disease (from date of detectable disease onset to date of detection), and observed disease (date of detection to death from the disease). The top line illustrates an individual's disease history given they are screened and the bottom line demonstrates the counterfactual – what would have happened had the individual not been screened.

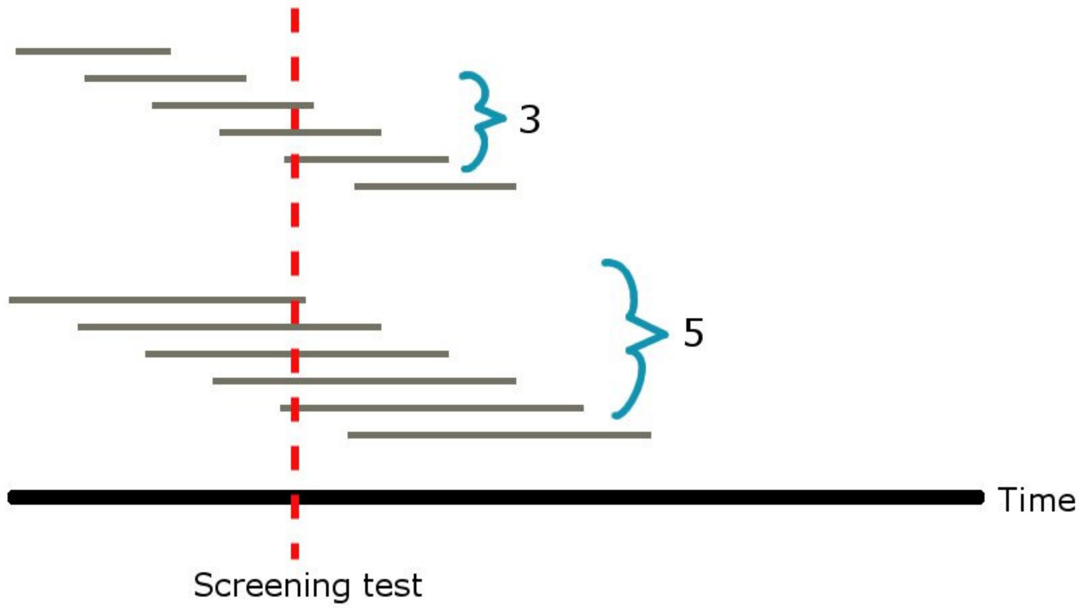



**Figure 2.** Diagram demonstrates length bias where the screening test selects a higher proportion of individuals with a long preclinical disease state compared to those with a short preclinical disease state. Assuming that length spent in the preclinical state is an indicator of overall disease progression rate, when a screening test is administered at a specific point in time, only 3 of 6 rapid progressive cases are detected (first 6 cases), while 5 of 6 slow progressive cases are detected (second 6 cases).

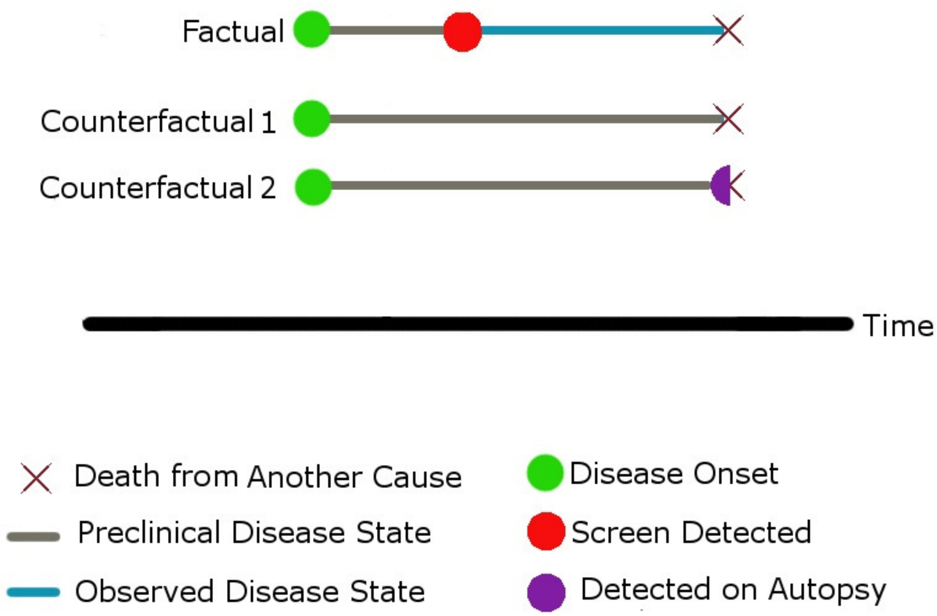



**Figure 3.** Diagram demonstrates that a subject is consider overdiagnosed when he/she would not have been symptomatically identified as a case before death from another cause had he/she not been identified as such by screening. In the absence of screening, an overdiagnosed case has such an extremely slow progressing disease that they would never have been identified as having the disease as illustrated by counterfactual 1 or would only be identified after death at autopsy as represented by counterfactual 2.

(a)

| % smokers screened | % non-smokers screened | OR (screening vs. smoking) | Case Ascertainment Period | | | | |
|---|---|---|---|---|---|---|---|
| | | | 1 yr | 1.5 yr | 2 yr | 5 yr | 10 yr |
| 55% | 65% | 0.66 | 0.94 | 0.96 | 0.97 | 0.99 | 0.99 |
| 50% | 70% | 0.43 | 0.88 | 0.92 | 0.95 | 0.98 | 0.99 |
| 40% | 80% | 0.17 | 0.76 | 0.83 | 0.88 | 0.95 | 0.97 |
| 30% | 90% | 0.05 | 0.62 | 0.73 | 0.79 | 0.92 | 0.95 |
| 25% | 95% | 0.02 | 0.55 | 0.66 | 0.73 | 0.89 | 0.94 |

(b)

| % smokers screened | % non-smokers screened | OR (screening vs. smoking) | Modal preclinical duration | | | |
|---|---|---|---|---|---|---|
| | | | 5 yr | 10 yr | 15 yr | 20 yr |
| 55% | 65% | 0.66 | 0.96 | 0.96 | 0.96 | 0.96 |
| 50% | 70% | 0.43 | 0.92 | 0.92 | 0.93 | 0.93 |
| 40% | 80% | 0.17 | 0.83 | 0.84 | 0.84 | 0.84 |
| 30% | 90% | 0.05 | 0.73 | 0.73 | 0.73 | 0.74 |
| 25% | 95% | 0.02 | 0.66 | 0.66 | 0.66 | 0.66 |

**Table 1**. Simulated *RR* for smoking vs. prostate cancer expected under the null hypothesis when the true *RR* = 1, evaluated using 5 *OR*s for screening vs. smoking and either one of (a) 5 different case ascertainment periods at preclinical duration = 5 years, or one of (b) 4 different preclinical durations with ascertainment period = 1.5 years.

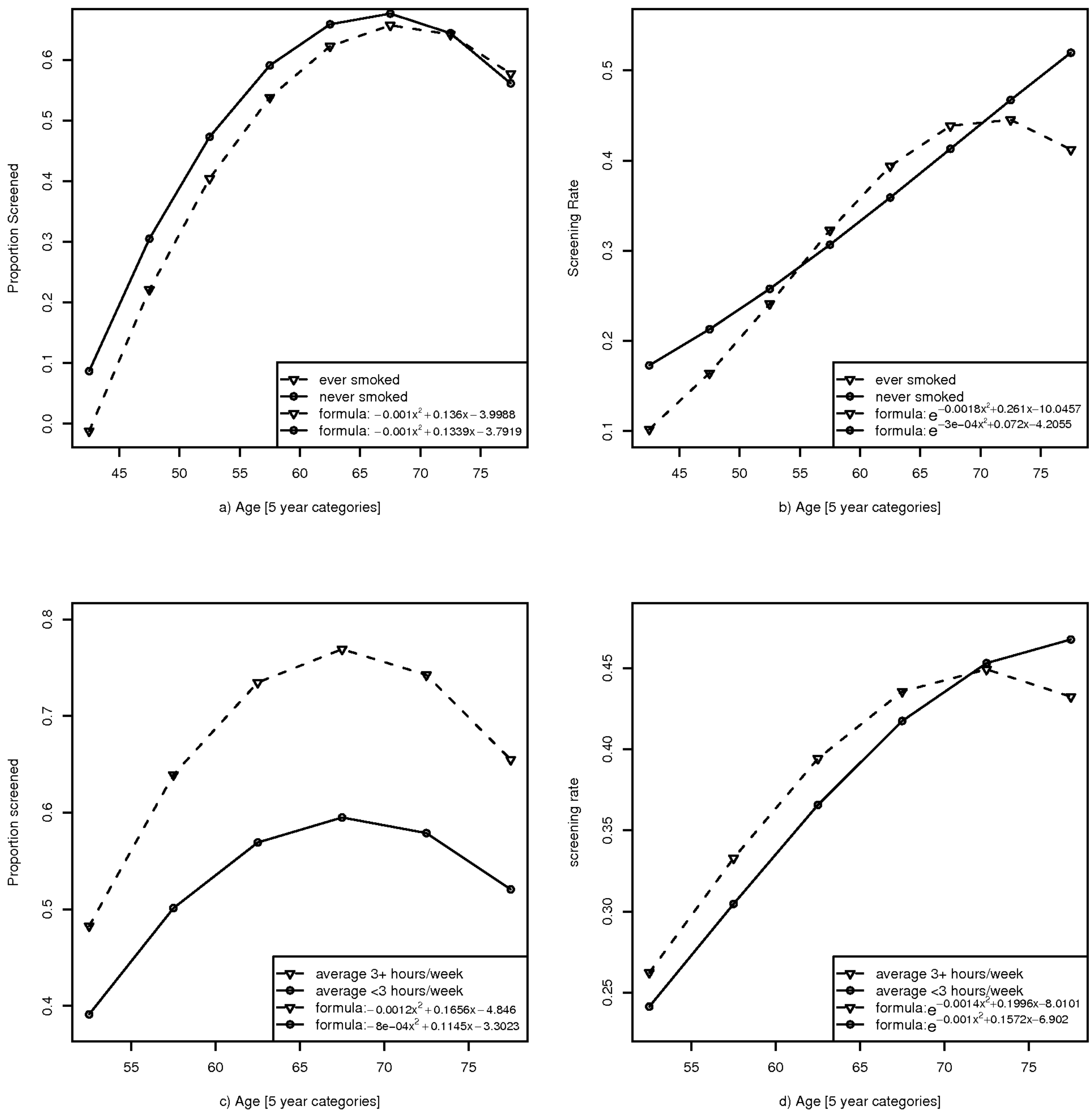



**Figure 4.** Representation of the proportion screened (number of participants who ever received a PSA test out of total number of participants) in each smoking stratum (a) and in each total physical activity stratum (c). Also represented is the screening rate per year (# of screening tests from 1990-1998 divided by 9 years) among those screened in each smoking stratum (b) and in each total physical activity stratum (d). The smoking variable separates those 40-79 that ever smoked from those that never smoked. Total physical activity variable separates the total physical activity in those 50-79 into categories of an average of 3 or more hours per week or less than 3 hours per week.

# Chapter 2: Research Design and Methods

For this doctoral work, previously collected data from the PLCO randomized control trial will be used. Interest resides in the potential for differential screening behaviors between strata to affect the observed smoking-lung cancer risk estimate and the effect study design has on the expected amount of screening bias. Because the data are from a randomized trial, the unique opportunity is present to develop nested case-control studies and identify through simulation and analysis (i.e., both theoretically and empirically) how structural differences in these studies change the amount of screening bias expected to affect the observed risk estimate (discussed in detail in chapter 4). This work will involve the well studied effect that smoking has on lung cancer as it provides the benefits of a large literature base of estimates, general consensus of a causal effect, and ability to distinguish directional effects screening has on the observed risk estimate. This project will demonstrate the impact screening can have on a known risk-factor disease estimate and provide both a mathematical model for theoretically correcting screening bias in observational lung cancer studies and a general technique to accounting for such bias in other observational studies with early detected disease.

## *Role of Counterfactual Framework*

Counterfactual is defined to say given the occurrence of an event, E, it is expect a subsequent event S will occur, however in the absence of E, S is not expected (43, 44). This concept is advantageous when thinking about the how the use of screening can influence the selection of cases into an observational study. When discussing how screening possibly affects a risk estimate, the value of the estimate should be obtained for each desired screening situation. For example, it is beneficial to evaluate the sampled population under the assumption nobody in the population is screened and compare that estimate to the situation where a fraction of the population receive screening. Essentially the difference in the two estimates is the influence screening has on the selection of cases into the sampled population and subsequent observed risk estimate.

## *Mathematical Model*

The counterfactual idea is implemented into the mathematical model as the model is used to simulate the outcome of interest in the sampled population under both unscreened and screened situations with differential screening behaviors between strata. This model is developed with the following components: preclinical duration function, preclinical incidence function, screening functions, sample specific age structure, screening test sensitivity, and enrollment period length. These components are explained in detail below.

### Preclinical Duration Function

Preclinical duration of disease for an individual is the theoretical length of time that an individual spends in the preclinical state. This preclinical period begins with disease onset and continues until symptomatic or clinical disease detection.

To create a preclinical duration function for our target population (i.e., the population we wish to generalize our results to), we want to use a distribution which will represent all the individuals in that population. A constraint on this distribution is that it must be greater than zero as it is not possible to have a negative preclinical duration. Information from the literature on the mean estimated preclinical duration for the disease, overdiagnosis estimates based on the screening test for the disease, case reports on the pathology of disease, and observational studies on the length of time between first exposure of the risk-factor and symptomatic or clinical disease detection will be used to create the function. This information should be incorporated into creating an informed, representative preclinical duration distribution for the model.

## Preclinical Incidence Function

Preclinical incidence of disease is the number of individuals that begin to develop disease out of the entire population. This preclinical incidence is unobservable and must be estimated from existing information.

To create a preclinical incidence function for our target population that is representative of that population, we must use existing information. Existing registries provide an excellent source of information on the incidence of disease. So for example, we could use this representative incidence data adjusted by the mean of the preclinical duration distribution to get a preclinical incidence distribution to use in the model (31).

## Screening Functions

The proportion screened and screening rate among those screened are functions derived from the sampled data. The questionnaire given to individuals at enrollment should contain questions about if, how many, and how often screening tests were received over a giving period of time leading up to the beginning of the study and during the study enrollment period if applicable. The screening proportion of the population would be estimated by the number that ever received a screen over the total sampled population. The screening rate would be calculated based on those who had ever received a screen as number of screens over number of years in the specific period. If there are no questions about screening behaviors in the study, screening information will have to be gathered from the literature using representative populations and the screening functions derived from that information.

## Other Important Components

Additional important model components include: age structure of sampled population, screening test sensitivity, and length of the enrollment period. For the age structure the percentage of the sampled population is calculated for each age or age category. These percentages are then applied as weights to the mathematical model to better simulate the target population assuming that the sampled population represents the age structure of the target population. The screening test sensitivity is identified from the literature preferably in a similar population. If screening test sensitivity has not been evaluated in a similar population, several literature based values should be used to test the mathematical model sensitivity to these variations. The length of the enrollment period is important to better simulate the study period and identify the potential affect the use of screening has on the selection of cases into the study.

## *Simulation*

Recurrence-time models, long used to study screening programs, can be modified to simulate potential screening bias in observational studies of risk factors associated with screening; they may also be adapted to provide adjustments to observed risk estimates. Such a progressive disease model overlaid with a simple screening model (1) (Figure 1) was used to identify the amount of bias in each nested case-control study.

The basic formulae relating lead-time to the actual disease rates in a screen efficacy study are described elsewhere(31) and partly shown below. Here $w(\cdot)$ is the preclinical incidence function, $f(\cdot)$ is the preclinical duration function where z is time of symptomatic diagnosis and x is time of detectable, preclinical disease onset, $S(\cdot)$ is the relative survival function, $t$ is a specified time, $t_E$ is end of the study period, $t_0$ is the beginning of the study period, $I(t > t_E)$ is an indicator function that equals 1 if $t > t_E$ and 0 is $t \le t_E$, and $\xi$ is the screening sensitivity.

Assuming the target population is all unscreened:

$$G_u(t) = \int_0^{\min(t,t_E)} w(x) \int_{\max(x,t_0)}^{\min(t,t_E)} f(z-x)[1-S(t-z)]dzdx$$

For a given time point, $t$, $G_u$ represents the cumulative mortality rate we expect to see during the study at that time point where detectable preclinical disease onset must occur at some time point, $t$, between birth (0) and the end of the study period ($t_E$) and symptomatic diagnosis and death must occur at some time point, $t$, during the study period (between beginning, $t_0$, and end of study period, $t_E$).

Now using the same target population, assuming a fraction of the target population is screened before ($k$) and during ($h$) the study period and that screening provides no survival benefit:

$$G_s(t) = (1-k\xi)\int_0^{t_0} w(x) \int_{t_0}^{\min(t,t_E)} f(z-x)[1-S(t-z)]dzdx$$

$$+ \int_{t_0}^{\min(t,t_E)} w(x) \int_{x}^{\min(t,t_E)} f(z-x)[1-S(t-z)]dzdx$$

$$+ I(t > t_E)h\xi\int_0^{t_E} w(x)\int_{t_E}^{t} f(z-x)[1-S(t-z)]dzdx$$

The same preclinical incidence, preclinical duration, and mortality functions are used as in the previous model. So for a given, $t$, $G_S$ represents the cumulative mortality rate expected at a study time, $t$, after incorporating that a fraction ($k$) of the individuals are screen detected before the study period and a fraction ($h$) are screen detected during the study period. For simplicity, assume that individuals are only screened once before the study and once during the study. This model has three parts where the first line identifies the cumulative mortality rate for cases in our target population expected to be screen detected (with probability $\xi$) before the beginning of the study period thus eliminating them as potential cases for the study and reducing the cumulative mortality rate expected during the study. The second line represents the cumulative mortality rate expected during the study period (same as described assuming the population is unscreened). The last line represents the cumulative mortality rate attributable to the fraction ($h$) of the population that when unscreened are noncases during the study period but now are screen detected during the study period (with probability $\xi$) and therefore add to the cumulative mortality rate expected during the study.

This model will be modified for risk-factor studies of incidence, by employing an incidence function based on representative SEER incidence data. The amount of possible bias from screening will be evaluated and an analysis of the relationship between smoking and the incidence of lung cancer will be conducted in an age-specific cohort at risk for lung cancer in the PLCO trial. To begin with, the assumption of a joint null hypothesis of no effect of the risk factor or screening on preclinical disease incidence will be used to isolate the bias affecting lung cancer incidence rates in the same population screened and unscreened. As a follow-up to this work at a later time, the results of the PLCO randomized trial will be incorporated to modify this assumption based on the estimated risk smoking has on lung cancer incidence.

Variables that have been added to the previous models and incorporated below include: $i$ which represents the number of 5-year age categories within our study population, $\omega$ is age structure of the study population and allows for better simulation of the study specific population, $k_1(x)$is the age dependent screening proportion function, and $k_2(x)$ is the age dependent screening rate function. Functions for before and during the enrollment period are represented with a "_b" and "_d" respectively, and were added to the above symbols for the preclinical incidence, screening proportion, and screening rate functions. These functions change over calendar year and so creating a before and during function averaged over calendar years is done to capture some of this variability while limiting the complexity of the model.

The mathematical formula used within each stratum when the outcome of interest is **incidence** follows:

Assuming the target population is all unscreened:

$$G_{U-R}(a) = \int_{0}^{\min(a,a_E)} [w\_b(x) + w\_d(x)] \int_{\max(x,a_0)}^{\min(a,a_E)} f(z-x)\,dz\,dx$$

For a given age, $a$, within a given risk-factor stratum, $G_{u-R}$ represents the cumulative incidence rate we expect to see during the study for that specific age where detectable preclinical disease onset must occur at some age, $a$, between birth (0) and age at end of the study period ($a_E$) and symptomatic diagnosis must occur at some age, $a$, during the study period (between age at beginning, $a_0$, and age at end of study period, $a_E$). In the model, the preclinical incidence function for before the study, $w\_b(\cdot)$, only has incidence rates up to age at the beginning of the case-ascertainment period, $a_0$, and the preclinical incidence function for during the study, $w\_d(\cdot)$, only has incidence rates for ages during the case-ascertainment period, $a_E$. The cumulative incidence rate, $G_{u-R}$, is summed over all age categories represented by our sampled population, $i$, applying the age structure of the sampled population, $\omega_i$, as weights.

$$\sum_i \omega_i \, G_{U-R}(i)$$

Assuming the target population has the same screening behavior as the sampled population (additional incidence added):

$$G_{S-R}(a) = \int_0^{a_0} w\_b(x) \left[ k\_b_1(x) * (1-\xi)^{\int_{\min(\max(screenage,x),a_0)}^{a_0} k\_b_2(y)dy} * [k\_d_1(x) * \left[ 1-(1-\xi)^{\int_{a_0}^{\min(a,a_E)} k\_d_2(y)dy} \right] \right] * \int_{\min(a,a_E)}^{\max age} f(z-x)\,dzdx$$

$$+ \int_{a_0}^{\min(a,a_E)} w\_d(x) \left[ k\_d_1(a_0) * \left[ 1-(1-\xi)^{\int_{\max(screenage,x)}^{\min(a,a_E)} k\_d_2(y)dy} \right] \right] * \int_{\min(a,a_E)}^{\max age} f(z-x)dzdx$$

$$- \int_0^{a_0} w\_b(x) \left[ k\_b_1(a_0) * \left[ 1-(1-\xi)^{\int_{\max(screenage,x)}^{a_0} k\_b_2(y)dy} \right] \right] * \int_{a_0}^{\min(a,a_E)} f(z-x)dzdx$$

The same preclinical incidence and preclinical duration functions are used as in the model for the target population under no screening. So for a given age, $a$, $G_{S-R}$

represents the cumulative incidence rate expected after incorporating the proportion screened, $k\_b_1(\cdot)$, and the screening rate among those who screen, $k\_b_2(\cdot)$ , before the study period and the proportion screened, $k\_d_1(\cdot)$ , and the screening rate among those who screen, $k\_d_2(\cdot)$ , during the study period. This model has three parts where the first line identifies the cumulative incidence rate for our target population expected to have detectable, preclinical disease onset occur at an age before the beginning of the study, who are missed by screening when screened between the age the population is recommended to start screening, represented by variable *screenage*, and age at the beginning of the study, $a_0$, but identified by screening at an age during the case-ascertainment period. In the absences of screening, these individuals would not have been identified as cases during the case-ascertainment period (symptomatically diagnosed after the end of the period). The second line represents the cumulative incidence rate expected among individuals who have disease onset occur at an age during the case-ascertainment period, who are screen detected at an age during the period where had they not been screened would have been symptomatically diagnosed at an age after the study period. Both the first and second lines identify the cumulative incidence expected to be added to a study because of the use of screening in the target population. The last line represents the cumulative incidence rate expected to be eliminated from the study because this part of the population has disease onset at an age before the beginning of the case-ascertainment period and are screen detected during that period where had they not been screened would have been eligible as a case for the study. The cumulative incidence rate, $G_{S\text{-}R}$, is summed over all age categories represented by our sampled population, $i$, applying the age structure of the sampled population, $\omega_i$, as weights (same as when we assume population is unscreened). $\sum_i \omega_i \; G_{S-R}(i)$

In addition to lead-time bias, overdiagnosis and length-biased selection will be accommodated in the example in chapter 3 by extending the tail of the preclinical duration distribution through the use of a bimodal distribution in order to incorporate that a fraction of the population has extremely long preclinical stages.

# Chapter 3: Can the use of screening significantly bias risk-factor estimates in observational studies?

In the literature, screening bias has been explored in terms of screening efficacy studies and only mentioned in the context of observational studies. Weiss (35) comments and Joffe (36) agrees that adjustment for screening must be done; however, no method or technique has been implemented or explored for handling this type of bias. In this chapter lead-time, length-biased selection, and overdiagnosis are further described in the context of how screening bias influences the selection of cases in observational studies. Additionally, the mathematical model from the previous chapter is parameterized using two example case-control studies to demonstrate the screening bias evaluation technique and potential affect screening bias has on case-control studies nested in the PLCO trial designed to estimate the risk of smoking on lung cancer incidence.

## *Screening Bias in Observational Risk-Factor Studies*

As described earlier, the use of screening in a population affects which members of that population will be selected into an observational study as a case even when screening has no benefit. In risk-factor studies, screening only creates bias when there is different screening behaviors (i.e., proportion and frequency of screening) observed between the strata of the risk-factor. The three types of screening bias previously discussed, lead-time, length, and overdiagnosis, will now be described in the context of observational studies.

The lead-time interval is the theoretical period of time by which the disease diagnosis has been advanced to an earlier date. In an observational setting, this interval becomes important when it crosses either the beginning or end of the case-ascertainment period. For example, if the lead-time interval crosses the beginning of the case-ascertainment period and ends sometime before the period ends, then the individual is diagnosed (screen-detected) before the beginning of the ascertainment period, thus is ineligible for the study. But considering the counterfactual situation, had the individual not been screened but symptomatically detected during the enrollment period, the case would have been eligible for the study. An additional situation where lead-time would be a problem occurs when the interval crosses the end of the case-ascertainment period. In this circumstance, a screen-detected cancer during the enrollment period is an eligible case, but under the counterfactual situation, the individual would only be eligible as a control.

As mentioned earlier, length bias arises because as the preclinical duration of a case increases so does the chance of being screen-detected(32-34). If it is assumed that preclinical duration is correlated with survival time, in an observational setting the stratum with the higher proportion of screen detected individuals will likely have an overall higher percentage of slower progressing disease types. When the outcome of interest is incidence in the observational study, length bias would increase the preclinical duration thereby modifying the case-ascertainment probabilities for the screened individuals.

Overdiagnosis bias, as discussed previously, occurs when a disease that would not symptomatically or clinically present before death is detected by screening during the preclinical stage. An overdiagnosed case has an extremely long lead-time interval leading to length bias when these types of cases are distributed unequally across risk-factor strata. These individuals have such a long preclinical duration stage that under the counterfactual of not being screened, either the disease would never be identified before death or the disease would only be discovered at autopsy. Both scenarios create a screen-detected case with no clinically surfacing counterpart such that an individual is only diagnosed with the disease during their lifetime because they were screened. The existence of this bias in observational risk-factor studies further increases the discrepancies discussed in the previous two paragraphs and their corresponding effects on the estimated RR.

### *Example*

The example used in this chapter will be constructed using data from the Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer randomized trial (41) in which the screening test used for lung cancer was chest x-ray. The mathematical model discussed in the previous chapter will be applied to two nested studies to evaluate the potential for screening bias in observational lung cancer studies. Plausible parameter values were selected for the model based on a combination of representative national registry and

PLCO data and a sensitivity analysis was conducted to elucidate the effect unobservable parameter values may have on the level of screening bias in the study.

## PLCO design

A primary goal of the PLCO was to examine if there was a benefit to following a specific screening protocol for four different cancers (i.e., Prostate, Lung, Colorectal, Ovarian). Enrollment for the PLCO study began in 1993 and ran through 2001 with approximately 77,000 men and 77,000 women aged 55-74 being followed for at least 13 years (42). Participants were obtained from ten sites around the United States including Denver, CO, Washington, DC, Honolulu, HI, Detroit, MI Minneapolis, MN, St. Louis, MO, Pittsburgh, PA, Salt Lake City, UT, Marshfield, WI, and Birmingham, AL.

As mention above, the focus here was to determine the potential for chest x-ray screening to influence the observed relationship between smoking and lung cancer within the PLCO data. Smoking status and screening behavior in the prior 3 years were assessed on the baseline questionnaire. After the initial chest x-ray examination, the participants were randomized into an intervention or usual-care arm. The chest x-ray screening protocol for lung cancer in the intervention arm was that individuals were scheduled to receive 3 annual screens (for a total of 4 when including the baseline screen). However, a procedural modification in December 1998 eliminated scheduling of the final reexamination screen for the never smokers in the intervention arm (45). This modification creates an inherently large difference in the proportion screened when comparing the ever to never smokers at study time T3 (Figure 5). The existence of this screening procedural modification is integral to this research as it provides an identified connection between level of screening and the estimated RR. Since all the case-control study designs are nested within the PLCO randomized trial it was expected that the estimated risk of smoking on lung cancer would be the same across study design with any difference being credited to the observed difference in screening behavior.

## *Parameterization of the Model*

In order to understand the potential for screening to bias case-ascertainment in the case-control studies nested in the PLCO trial, the described model is applied using parameters representative of the PLCO target population – the United States. To isolate lead-time and length bias, simulations incorporated a joint null hypothesis of no effect of the risk factor or screening on the preclinical incidence of lung cancer. Under these conditions, the true risk estimate is 1 and any systematic deviation from this value in the simulation represents bias. The simulations were developed and run using the program Mathcad® 12 (Cambridge, MA)(37) and analyses of PLCO data along with graphics was performed using R version 2.1.1(29) and SAS software version 9.1 (46). The components and parameterization of the model are described below.

### Preclinical duration distribution

It is known that lung cancer has different incidence rates based on different population characteristics (e.g., race, age, family history); it is also conceivable that the preclinical duration may vary based on those specific population characteristics. Because the true distribution of preclinical duration for lung cancer is unknown, a sensitivity analysis has been performed by simulating several plausible values for the mode (1, 3,5,10 years) and standard deviation (1, 3, 5 years) of the assumed lognormal distribution (Figure 6). A lognormal distribution is used as it meets the requirement to remain strictly positive overall all values.

## Preclinical incidence function

The SEER 9 registry(47) provides estimates for the incidence rate of lung cancer by age in the entire U.S. population based on nine long standing cancer registries. For the simulation, the focus was specifically on the years 1986-2005. Over this time period the age-specific incidence curve for each year (compared with the previous year's curve) had a larger peak with a sharper increase at a later age before the peak and sharper decrease at an earlier age after the peak. To limit the level of complexity of the model while still incorporating this trend, the years 1986-1995 and 1996-2005 were used separately to get two different sets of average incidence points by age representing rates before and during the PLCO study, respectively. A curve was fit to each set of average age-specific lung cancer incidence rate data points from SEER9 (Figure 7) and shifted by the mean of the preclinical duration to produce an age-specific preclinical incidence distribution to use in our simulations. It is important to note that the incidence rate points used to derive the preclinical incidence curves are assumed to have been collected in a sample of the United States population that exhibit similar screening behaviors for smokers and nonsmokers (i.e., assume same curve for both strata).

## Screening intensity function

To measure screening behavior, the PLCO participants were asked if in the three years prior to their enrollment at study time T0, they ever had received a chest x-ray and the number of times they were screened (categories: 0,1,2+) during that period. Based on the baseline questionnaire, the age-specific proportion screened (Figure 8) and the age-specific rate of screening (Figure 9) were determined for both smokers and nonsmokers. Three different constant sensitivities of 0.46, 0.66, and 0.86 were assumed for the chest x-ray screening test to incorporate literature based variations in these estimates and test model sensitivity to this parameter (48, 49). Generally, the only model parameter modification needed in a specific study when changing between risk factors within the same study is the age-stratum-specific equations for proportion screened and screening rate among those screened.

## Observed incidence

The outcome variable used for the model was observed total incidence comprised of both non-screened and screened cases. We stratified our PLCO sample by birth cohort to get categories of age at study entry and age-specific preclinical incidence rates as described in the previous section. Each risk-factor level (i.e., smokers and nonsmokers) was stratified by birth cohort in order to calculate the age-specific screening rates and proportion screened for each stratum as described in the screening intensity section.

The risk-factor-birth-cohort-specific observed incidence rate (i.e., cohort stratified based on age and smoking status) was simulated given an ascertainment period of 2 years. First, incidence assuming no screening in the population was simulated by integrating the preclinical incidence function and preclinical duration distribution as described previously. Next, the incidence under the study specific screening proportions and rates was simulated by essentially adding the cases moved into the study because of screening to the unscreened incidence and subtracting the cases moved out of the study because of screening. To get the overall observed incidence, the risk-factor-birth-cohort-specific observed incidence was summed across birth cohort strata using as birth-cohort weights, the actual birth-cohort proportions seen in the PLCO population sampled for this study.

## Theoretical relative risk correction

Typically in a case-control study, the odds ratio (OR) is used as the risk estimate, largely due to the study design restriction (e.g., selecting controls at the end of the case-ascertainment period). In most situations when the study disease is rare, as with lung cancer, the OR is used to approximate a relative risk (RR) with the equation $OR_{observed} = (a/b)/(c/d) \approx (a/(a + b))/(c/(c + d)) = RR_{estimated}$. This equation is based on a typical two-by-two table (1)

| | Lung cancer Diagnosis | Person-time at risk |
|---|---|---|
| Ever smoked | *a* | *b* |
| Never smoked | *c* | *d* |

(1)

where a is the number of cases with the risk factor, b is the number of noncases with the risk factor, c is the number of cases without the risk factor, and d is the number of noncases without the risk factor. Also, if the case-control study uses incidence density sampling where controls are selected at the failure time of the case (as done here), a RR is directly calculated rather than an OR. The simulated RR can be thought of as the ratio between the observed incidence rates in two risk-factor strata (e.g., smoking vs. nonsmoking) under the null hypothesis of no association between risk-factor or screening and disease. The unbiased RR between the strata would equal 1; any deviation represents bias. To theoretically correct our observed RR for any screening bias, we make the expected rates equal between the strata by multiplying the incident cases in the denominator stratum (i.e., *c*) by the simulated RR (i.e., expected amount of screening bias). The result is a "true" RR under the null hypothesis, which assumes that the bias affects cases proportionately and that no other forms of bias are present in the study.

## *Simulation Results*

The parameter combinations of preclinical duration distributions (mode years of 1, 3, 5, 10; standard deviation years of 1, 3, 5) for the smoked variable categorized into age groups 55-59 and 60-64; 65-69 and 70-74, respectively, produced twelve simulated risk ratios (Figure 10). The graph on the left illustrates that in the study that samples from the entire PLCO enrollment years (93-01) in the usual-care group with a case-ascertainment of T3-T5 as the mode increases, so does the bias, while the difference in bias between different standard deviation years remains relatively equal. The graph on the right illustrates that in the study sampled from those affected by the procedural modification in the intervention group with case-ascertainment T3-T5 as the mode year increases, so does the bias, while the difference in bias between different standard deviation years decreases. The result from the study done in the usual-care group follows a mostly linear pattern and the study done in with the intervention group follows more of a log pattern.

Simulation values were calculated for both datasets using a combination of study specific and study representative model parameters. The length of the case-ascertainment period, age structure, and screening proportion and rate functions were incorporated based on the dataset. Estimates of the chest x-ray screening test sensitivity for lung cancer were obtained from the literature and were used across study designs. The preclinical duration and preclinical incidence functions are unobservable and the assumptions that were used to create the distributions are stated in the "parameterization of the model" section of this chapter.

The tables below illustrate that three different chest x-ray sensitivities of 46%, 66%, and 86% were used along with twelve different combinations of the mode years (1,3,5,10) and standard deviation years (1,3,5) for a lognormal preclinical duration distribution. These combinations were repeated for each screening test sensitivity under the assumption the population had a 20% overdiagnosis rate. The overdiagnosis rate is incorporated into the simulation by modifying the representative lognormal preclinical duration distribution for the target population so that 20% of the population are assumed to be drawn from a lognormal preclinical duration distribution with a mode of 20 years and standard deviation of 3 years and the other 80% from one of the previously specified lognormal distributions. So in actuality because the screening exam sensitivity is below 100%, the simulated sample population will not exhibit an overdiagnosis rate of 20% but rather something lower in association with screening sensitivity. In other words, not all 20% of the population with the extremely long preclinical duration will be detected by screening. The association between screening sensitivity and overdiagnosis within a sample population is the probability of being screen detected multiplied by the overdiagnosis rate in the target population.

For the dataset sampled from the entire PLCO enrollment period (93-01) in the usual-care group with a case-ascertainment period of T3-T5 and with a chest x-ray screening test sensitivity of 46%, the *RR* range from 1.01 when the mode is 1 and standard deviation is 1 for the lognormal preclinical duration distribution to 1.12 when the mode is 10 and standard deviation is 5 (Table 2a). Adding a 20% overdiagnosis rate the *RR* range is 1.17 to 1.21 (Table 2c). For a screening sensitivity of 66%, the simulated *RR* values range from 1.01 to 1.12 (Table 3a) and with a 20% overdiagnosis rate to these simulations creates the *RR* range of 1.16 to 1.19 (Table 3c). When using a sensitivity of 86%, the simulated *RR* value range changes to 1.01 to 1.11 (Table 4a) and with a 20% overdiagnosis rate the range of *RR* values is 1.14 to 1.16 (Table 4c).

For the dataset sampling those affected by the procedural modification (95-01) from the intervention group with a case-ascertainment period of T3-T5 and with a chest x-ray screening test sensitivity of 46%, the *RR* range from 1.09 when the mode is 1 and standard deviation is 1 for the lognormal preclinical duration distribution to 1.97 when the mode is 10 and standard deviation is 5 (Table 2b). Adding a 20% overdiagnosis rate, the *RR* range is 2.07 to 2.04 (Table 2d). For a screening sensitivity of 66%, the simulated *RR* values range from 1.13 to 1.93 (Table 3b) and with a 20% overdiagnosis rate to these simulations creates the *RR* range of 1.99 to 1.95 (Table 3d). When using a sensitivity of 86%, the simulated *RR* value range changes to 1.18 to 1.85 (Table 4b) and with a 20% overdiagnosis rate the range of *RR* values is 1.91 to 1.86 (Table 4d).

## Discussion

When a screening test is used among subjects in an observational study, the screen-detected cases will have an earlier date of diagnosis and likely slower progressing disease compared to non-screen-detected cases resulting in screening bias. If differential screening behavior exists between risk-factor stratum, case-ascertainment may be changed differentially, and thereby misrepresenting the observed estimate from the data between the risk factor and disease. In the presence of differential screening under plausible assumptions about preclinical incidence and duration, the simulations presented here show the possibility for screening bias from chest x-ray to significantly affect the risk smoking has on the development of lung cancer.

Using two case-control studies designs nested within the PLCO randomized trial as simulation examples, a possible range for such bias within the smoking-lung cancer observed risk estimate has been given. Within these results, a relationship has emerged that as screening differential (either in proportion or rate) between strata of the variable (e.g., ever smoked vs. never smoked) increases, so does the susceptibility to this screening bias (Figure 8 and 9). Figure 10 illustrates that in general when the mode and standard deviation increase, so does the amount of bias expected to affect the observed RR. Also, the model appears to be relatively sensitive to standard deviation and even more so t0 mode variations causing simulation values to differ by about 90% kwhen comparing smallest to largest pairs of these parameters (Table 2-4). This result can be explained by considering that a disease that has a long preclinical duration in combination with an increase in screening during the study will have the potential to shift many cases into the study that would otherwise not be identified as such. There is an indication in Table 2-4c&d that incorporating overdiagnosis can have a significant effect on the RR, most of all when the preclinical duration is shortest (e.g., mode=1, standard deviation = 1). This observation fits with the previous result that a longer preclinical duration (added by overdiagnosis to the short preclinical duration group) for a disease increases the expected bias affecting the observed RR. The model is only moderately sensitive to screening test sensitivity (average difference is about 2%) comparing simulated values with equal parameters (Table 2-4) when there are small screening behavior differences (i.e., selecting from the usual-care group). The model becomes slightly more sensitivity to different screening test sensitivity variations when screening behavior difference between smokers and never smokers is increased (e.g., in the intervention group study) where the range of differences is 4% to 12% (Table 2-4).

Model sensitivity to parameter variation because of study design modification will be explored in the next chapter.

The conditions of the case-control studies nested in the PLCO trial (i.e., short study duration) and screening for lung cancer with chest x-ray (i.e., potential for overdiagnosis) are such that considerable bias can arise. Because of screening in this population, the cases observed in the ascertainment period are a different subgroup than the cases expected in the absence of screening. Thus, even if screening is theoretically accounted for with the simulated value, study results should be generalized with caution.

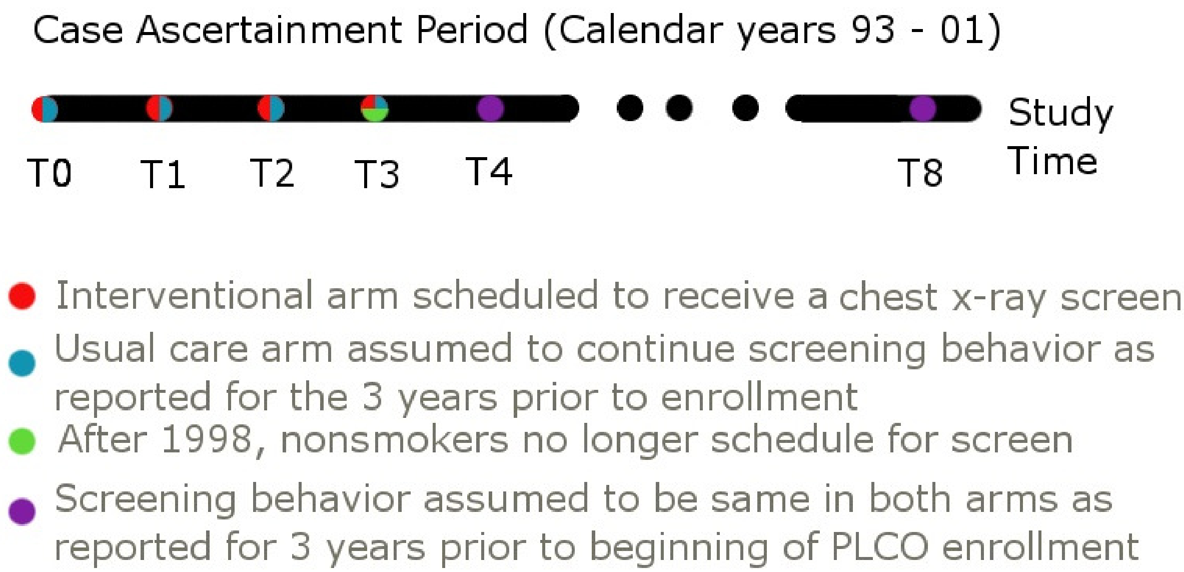



**Figure 5.** Illustration of the case-ascertainment period for the PLCO randomized trial with identification of study years and screening protocols. Individuals in the intervention arm are scheduled to receive 4 total chest x-ray screens based on the initial protocol with a 1998 modification reducing the total number screens offered to nonsmokers to 3. For the purposes of simulating screening bias here in any study year where screening information was not collected, it was assumed that individuals would continue screening behaviors as reported on the baseline questionnaire for before the beginning of the trial.

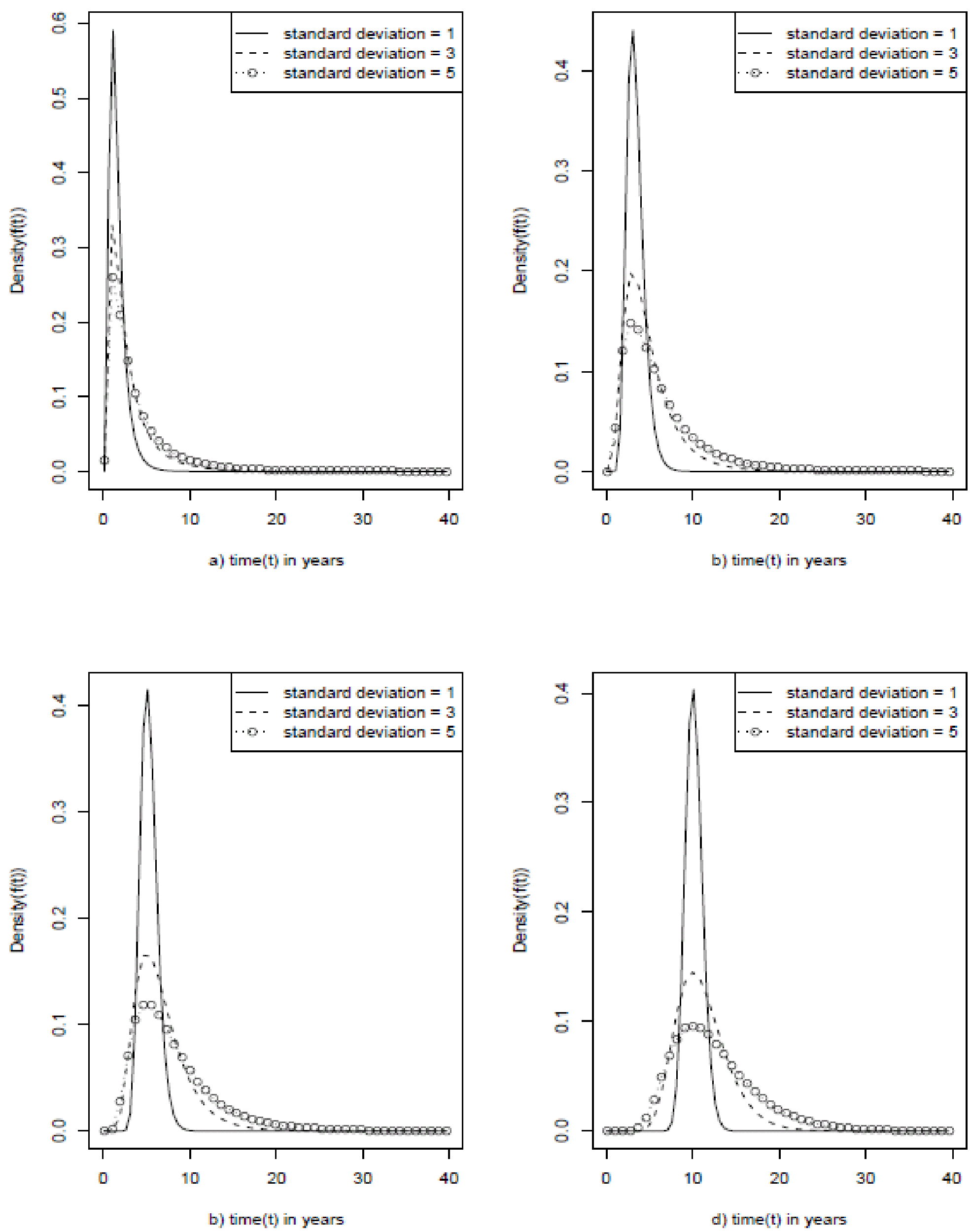



**Figure 6.** Presentation of several plausible preclinical duration distributions for lung cancer based on a log normal distribution with standard deviations of 1, 3, and 5 years for each of the following modes: 1 (a), 3 (b), 5 (c), and 10 (d). The log normal distributions are used to represent the distribution for the lengths of time individuals in our population spend in the detectable, preclinical state assuming no screening in the population. Because the preclinical duration distribution is unknown for lung cancer, the model sensitivity to variation in these parameters is explored by using the 12 different combinations. The points have not value and are just to help distinguish between the different standard deviations within each plot.

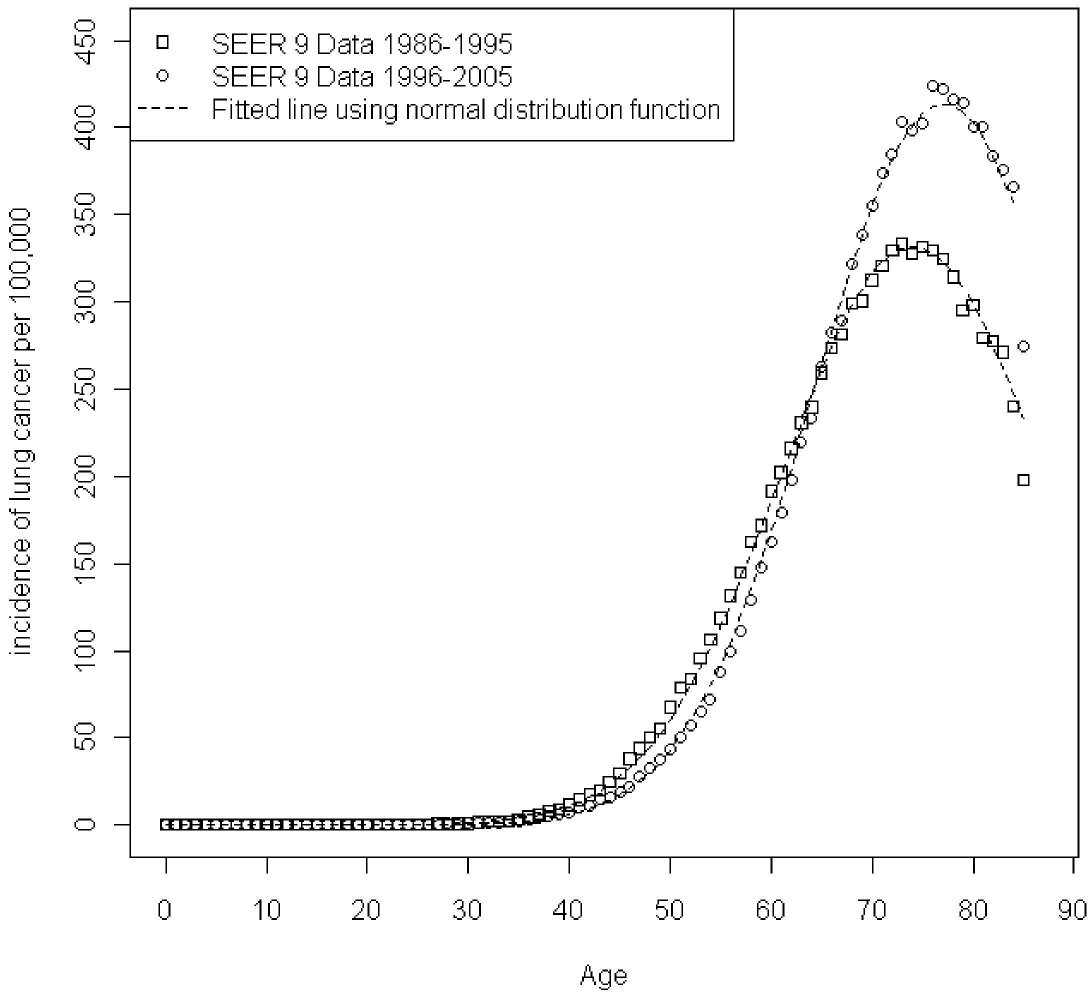



**Figure 7.** Relationship of age (5 year age groups; age range 0-85+) to incidence rate (per 100,000) of lung cancer based on average SEER 9 registry data from 1986 to 2005. A continuous age-specific incidence intensity function was fit to the point estimates from the SEER data using non-linear minimization (dotted line). The square points identify the data points used to create the preclinical incidence function in the population before the beginning of the study and the circular points identify the data points used to create the preclinical incidence function in the population during the study. Since preclinical incidence is unobservable to get a representative preclinical incidence function the continuous incidence function represented above is shifted backward by the mean of the preclinical duration distribution. For example, if the mean of the preclinical duration distribution is 5, the incidence observed for a 55 year old becomes the preclinical incidence for a 50 year old.

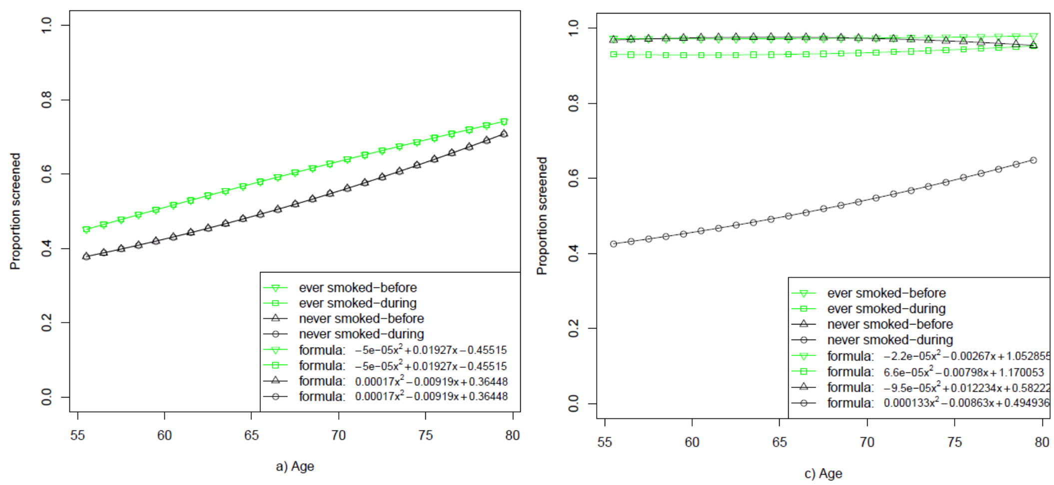



**Figure 8.** Representation of the proportion screened both before and during the study for ever smokers and never smokers separately. The age specific proportion screened (number of participants who ever received a chest x-ray test out of total number of participants at each age) is plotted for a) study sampling from all PLCO calendar enrollment years (93-01) in the usual-care arm of the PLCO during the study years T3 to T5 and c) study sampled only those affected by the procedural modifciation (95-01) in the intervention arm of the PLCO during the study years T3 to T5. Screening information from the 3 years prior to the beginning of the PLCO study and study times T0-T2 is used to calculate the proportion screened functions for before the study and screening information collected for T3 along with information from the 3 years prior to the beginning of the PLCO is used to calculate the proportion screened during the study enrollment period for each study design.

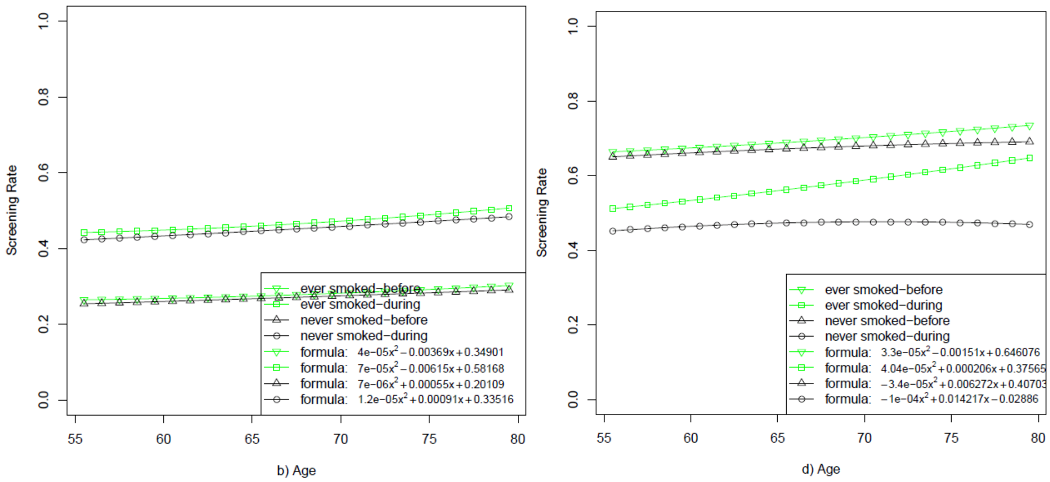



**Figure 9.** Representation of the screening rate both before and during the study for ever smokers and never smokers separately. The screening rate per year (# of screens received divided by number of years in period) among those screened is displayed for b) study sampling from all PLCO calendar enrollment years (93-01) in the usual-care arm of the PLCO during the study years T3 to T5 and d) study sampled only those affected by the procedural modifciation (95-01) in the intervention arm of the PLCO during the study years T3 to T5. Screening information from the 3 years prior to the beginning of the PLCO study and study times T0-T2 is used to calculate the screening rate functions for before the study and screening information collected for T3 along with information from the 3 years prior to the beginning of the PLCO is used to calculate the screening rate during the study enrollment period for each study design.

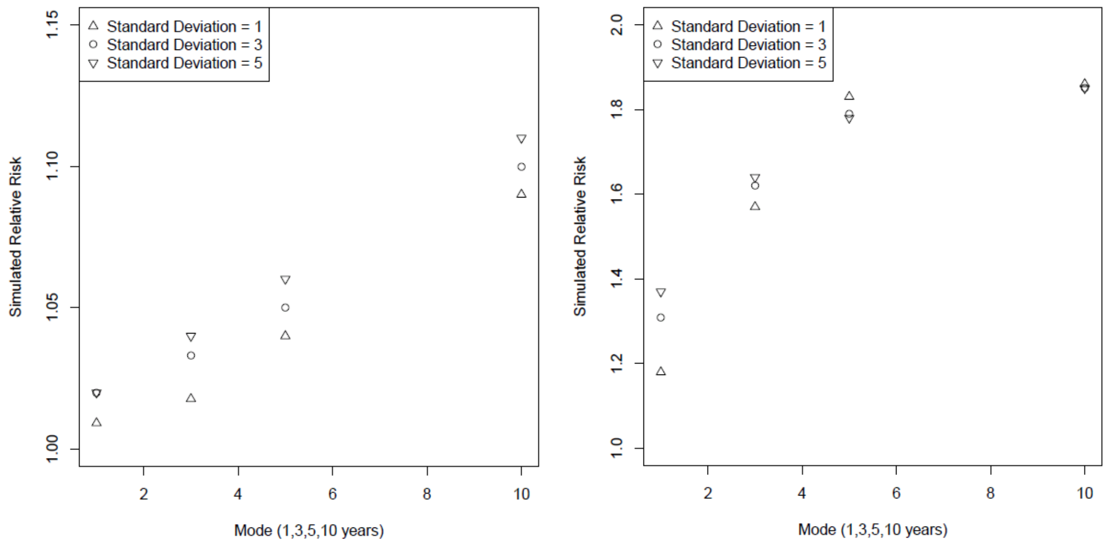



**Figure 10.** Simulated relative risks for smoking under the double null hypothesis (i.e., smoking and screening are independent of lung cancer) for studies sampling from the entire PLCO enrollment period (93-01) in the usual-care group (left) or from those affected by the procedural modification (95-01) in the intervention group (right). Both studies select cases and sample noncases between study time T3 and study time T5. The 12 relative risks were simulated using a combination of four preclinical duration distribution parameters for the mode (1,3,5,10) and three standard deviations (sd) (1,3,5). The relative risks are comparing the categories "ever smoked" to "never smoked." The simulation is based on study sample specific age distributions and screening proportion and rates among those screened.

a) Usual-care group sampled from entire enrollment period: Simulation results using smoked variable: Lognormal distribution for preclinical duration with no overdiagnosis and chest x-ray sensitivity of 46%

| | Mode = 1 | Mode = 3 | Mode = 5 | Mode = 10 |
|---|---|---|---|---|
| Standard dev. = 1 | 1.01 | 1.02 | 1.04 | 1.10 |
| Standard dev. = 3 | 1.01 | 1.03 | 1.05 | 1.12 |
| Standard dev. = 5 | 1.02 | 1.04 | 1.06 | 1.12 |

b) Intervention group sampled after procedural modification: Simulation results using smoked variable: Lognormal distribution for preclinical duration with no overdiagnosis chest x-ray sensitivity of 46%

| | Mode = 1 | Mode = 3 | Mode = 5 | Mode = 10 |
|---|---|---|---|---|
| Standard dev. = 1 | 1.09 | 1.32 | 1.64 | 1.99 |
| Standard dev. = 3 | 1.19 | 1.43 | 1.67 | 1.98 |
| Standard dev. = 5 | 1.24 | 1.49 | 1.70 | 1.97 |

c) Usual-care group sampled from entire enrollment period: Simulation results using smoked variable: Lognormal distribution for preclinical duration with 20% overdiagnosis and chest x-ray sensitivity of 46%

| | Mode = 1 | Mode = 3 | Mode = 5 | Mode = 10 |
|---|---|---|---|---|
| Standard dev. = 1 | 1.17 | 1.14 | 1.12 | 1.14 |
| Standard dev. = 3 | 1.19 | 1.19 | 1.19 | 1.19 |
| Standard dev. = 5 | 1.19 | 1.20 | 1.20 | 1.21 |

d) Intervention group sampled after procedural modification: Simulation results using smoked variable: Lognormal distribution for preclinical duration with 20% overdiagnosis and chest x-ray sensitivity of 46%

| | Mode = 1 | Mode = 3 | Mode = 5 | Mode = 10 |
|---|---|---|---|---|
| Standard dev. = 1 | 2.07 | 2.05 | 2.04 | 2.03 |
| Standard dev. = 3 | 2.07 | 2.06 | 2.05 | 2.04 |
| Standard dev. = 5 | 2.07 | 2.06 | 2.05 | 2.04 |

**Table 2.** Simulated *RR*s for the studies sampled from the entire PLCO enrollment period (93-01) in the usual-care group (a and c) and only those affected by the procedural modification (95-01) the intervention group (b and d) using a lognormal distribution for the preclinical duration with modes of 1, 3, 5, and 10 years and standard deviations of 1, 3, 5 years and a constant chest x-ray sensitivity of 46%. To test model sensitivity to overdiagnosis, in the bottom two tables a 20% overdiagnosis rate was applied by assuming that 20% of the population had a lognormal preclinical duration with mode of 20 years and standard deviation of 3 years and the other 80% of the population had a lognormal preclinical duration as indicated in the table. Because of imperfect screening sensitivity, the overdiagnosis rate in the simulated sample population is actually less than 20%.

a) Usual-care group sampled from entire enrollment period: Simulation results using smoked variable: Lognormal distribution for preclinical duration with no overdiagnosis and chest x-ray sensitivity of 66%

| | Mode = 1 | Mode = 3 | Mode = 5 | Mode = 10 |
|---|---|---|---|---|
| Standard dev. = 1 | 1.01 | 1.02 | 1.04 | 1.10 |
| Standard dev. = 3 | 1.02 | 1.03 | 1.05 | 1.11 |
| Standard dev. = 5 | 1.02 | 1.04 | 1.06 | 1.12 |

b) Intervention group sampled after procedural modification: Simulation results using smoked variable: Lognormal distribution for preclinical duration with no overdiagnosis and chest x-ray sensitivity of 66%

| | Mode = 1 | Mode = 3 | Mode = 5 | Mode = 10 |
|---|---|---|---|---|
| Standard dev. = 1 | 1.13 | 1.46 | 1.79 | 1.94 |
| Standard dev. = 3 | 1.25 | 1.54 | 1.78 | 1.94 |
| Standard dev. = 5 | 1.31 | 1.59 | 1.78 | 1.93 |

c) Usual-care group sampled from entire enrollment period: Simulation results using smoked variable: Lognormal distribution for preclinical duration with 20% overdiagnosis and chest x-ray sensitivity of 66%

| | Mode = 1 | Mode = 3 | Mode = 5 | Mode = 10 |
|---|---|---|---|---|
| Standard dev. = 1 | 1.16 | 1.13 | 1.12 | 1.13 |
| Standard dev. = 3 | 1.17 | 1.17 | 1.17 | 1.18 |
| Standard dev. = 5 | 1.18 | 1.18 | 1.18 | 1.19 |

d) Intervention group sampled after procedural modification: Simulation results using smoked variable: Lognormal distribution for preclinical duration with 20% overdiagnosis and chest x-ray sensitivity of 66%

| | Mode = 1 | Mode = 3 | Mode = 5 | Mode = 10 |
|---|---|---|---|---|
| Standard dev. = 1 | 1.99 | 1.98 | 1.97 | 1.95 |
| Standard dev. = 3 | 1.99 | 1.98 | 1.97 | 1.95 |
| Standard dev. = 5 | 1.99 | 1.98 | 1.97 | 1.95 |

**Table 3.** Simulated *RR*s for the studies sampled from the entire PLCO enrollment period (93-01) in the usual-care group (a and c) and only those affected by the procedural modification (95-01) the intervention group (b and d) using a lognormal distribution for the preclinical duration with modes of 1, 3, 5, and 10 years and standard deviations of 1, 3, 5 years and a constant chest x-ray sensitivity of 66%. To test model sensitivity to overdiagnosis, in the bottom two tables a 20% overdiagnosis rate was applied by assuming that 20% of the population had a lognormal preclinical duration with mode of 20 years and standard deviation of 3 years and the other 80% of the population had a lognormal preclinical duration as indicated in the table. Because of imperfect screening sensitivity, the overdiagnosis rate in the simulated sample population is actually less than 20%.

a) Usual-care group sampled from entire enrollment period: Simulation results using smoked variable: Lognormal distribution for preclinical duration with no overdiagnosis and chest x-ray sensitivity of 86%

| | Mode = 1 | Mode = 3 | Mode = 5 | Mode = 10 |
|---|---|---|---|---|
| Standard dev. = 1 | 1.01 | 1.02 | 1.04 | 1.09 |
| Standard dev. = 3 | 1.02 | 1.03 | 1.05 | 1.10 |
| Standard dev. = 5 | 1.02 | 1.04 | 1.06 | 1.11 |

b) Intervention group sampled after procedural modification: Simulation results using smoked variable: Lognormal distribution for preclinical duration with no overdiagnosis and chest x-ray sensitivity of 86%

| | Mode = 1 | Mode = 3 | Mode = 5 | Mode = 10 |
|---|---|---|---|---|
| Standard dev. = 1 | 1.18 | 1.57 | 1.83 | 1.86 |
| Standard dev. = 3 | 1.31 | 1.62 | 1.79 | 1.85 |
| Standard dev. = 5 | 1.37 | 1.64 | 1.78 | 1.85 |

c) Usual-care group sampled from entire enrollment period: Simulation results using smoked variable: Lognormal distribution for preclinical duration with 20% overdiagnosis and chest x-ray sensitivity of 86%

| | Mode = 1 | Mode = 3 | Mode = 5 | Mode = 10 |
|---|---|---|---|---|
| Standard dev. = 1 | 1.14 | 1.12 | 1.10 | 1.12 |
| Standard dev. = 3 | 1.15 | 1.15 | 1.15 | 1.15 |
| Standard dev. = 5 | 1.16 | 1.16 | 1.16 | 1.16 |

d) Intervention group sampled after procedural modification: Simulation results using smoked variable: Lognormal distribution for preclinical duration with 20% overdiagnosis and chest x-ray sensitivity of 86%

| | Mode = 1 | Mode = 3 | Mode = 5 | Mode = 10 |
|---|---|---|---|---|
| Standard dev. = 1 | 1.91 | 1.90 | 1.88 | 1.87 |
| Standard dev. = 3 | 1.90 | 1.89 | 1.88 | 1.87 |
| Standard dev. = 5 | 1.90 | 1.89 | 1.88 | 1.86 |

**Table 4.** Simulated *RR*s for the studies sampled from the entire PLCO enrollment period (93-01) in the usual-care group (a and c) and only those affected by the procedural modification (95-01) the intervention group (b and d) using a lognormal distribution for the preclinical duration with modes of 1, 3, 5, and 10 years and standard deviations of 1, 3, 5 years and a constant chest x-ray sensitivity of 46%. To test model sensitivity to overdiagnosis, in the bottom two tables a 20% overdiagnosis rate was applied by assuming that 20% of the population had a lognormal preclinical duration with mode of 20 years and standard deviation of 3 years and the other 80% of the population had a lognormal preclinical duration as indicated in the table. Because of imperfect screening sensitivity, the overdiagnosis rate in the simulated sample population is actually less than 20%.

# Chapter 4: How does the case-control design influence the theoretical amount of bias caused by screening use?

## *Nested Case-Control Studies*

Extensive research has been conducted on designing case-control studies. Rothman and Greenland (50) state that any case-control study can be considered nested within a source population and our designed studies will be nested within the PLCO randomized control trial. This nesting will give us a well characterized parent population, prospective measurements, and tightly monitored follow-up (51, 52), as well as a way to reliably compare the estimates across case-control designs.

In order to determine what effects study specific characteristics used in the parameterization of the mathematical model have on the simulated relative risk (RR), several nested case-control studies were designed to create different levels of screening behavior (i.e., proportion screened and rate of screening) among "ever smokers" and "never smokers". Important general components to designing a nested case-control study include case and control selection techniques, cohort stratification, and length of the ascertainment period(31, 53-56). The temporal location of the ascertainment period within the randomized trial is an aspect of the study designs that was specifically manipulated to test the model.

The selection of the cases for each design involved identifying cases with respect to diagnostic criteria (53). Incidence density sampling was implemented to select all cases and a random sample of controls weighted by person-time contributed to the denominator to enabling estimation of a RR (50). Using different combinations of case subpopulations (intervention arm, usual-care arm, or a combination) from the PLCO trial in our study designs allowed determination of the influence these modifications and resulting biases have on the estimated RRs. The four sampling schemes for selection of controls at random from cohort members at risk at the failure time of the case as identified in Robins et al. (54) are: 1) sampling controls without replacement from the noncases; 2) sampling controls with replacement from the entire risk set including the cases; 3) select controls without replacement from entire risk set; and 4) select controls with replacement from the noncases. For simplicity and unbiased sampling, we will use scheme 2 above for all our studies.

Decisions about stratifying the study population can affect the amount of screening bias we expect to see in each design. Several studies have indicated that subjects who have refused and those who have accepted screening express a different disease incidence (57-59). Connor et al (57) suggests that in order to eliminate this self-selection bias, we should limit the eligibility of the case-control study to only those offered screening. However, limiting our analysis to the intervention group of the PLCO cohort will not eliminate the differential screening behaviors present between smoking strata before nor during the study. In the PLCO differences in screening behaviors arise during the study because of a difference in compliance rates and the procedural modification that affected those scheduled to receive a third annual screen after 1998.

The length of the case ascertainment period also play an important role in determining which members of the cohort are at risk and thus eligible for a specific case-control study. In terms of follow-up time, researchers have indicated that comparing the

5-year survival of a group of subjects who receive a second mammography screen to a nonscreened group produces less bias (length and lead-time) than comparing a group of subjects who receive a first screen to a nonscreened group (60). Less bias will likely exist at the second screen because several "cases" will have been detected at the first screen (a percentage of those in their preclinical stage based on screening sensitivity) eliminating those subjects from the second screened group. Therefore less "shifting" of "cases" into or out of the study will take place during the time of the second screen. Church (31) provides several suggestions to prevent case-ascertainment bias such as eliminating the use of the date of diagnosis to define cases, extending the ascertainment window to a time before screening was implemented, and including all incident cases. When designing a case-control study, the length of the ascertainment period is generally chosen to yield the required number of cases needed to reach a specific power. However for the study designs here, the case-ascertainment period is fixed and 200 cases are randomly selected from the identified interval.

### *Goals of Nested Case-Control Study Designs*

The use of several nested case-control studies enabled the determination of how the study design influences the amount of screening bias expected as identified by the simulated *RR*. Based on this information, design suggestions were provided for future case-control risk-factor studies potentially affected by screening bias. Behavioral and nutritional studies, as well as studies of the interaction of environmental and molecular factors can benefit from these results. As shown in the case of lead-time, bias can significantly modify the observed risk estimate. Adjusting for these forms of bias is very important in validating and improving accuracy of the conclusions of the PLCO study.

The case-control studies were nested within the PLCO trial in order to identify how screening induced bias affects the smoking-lung cancer *RR* of each study. The diagram in Figure 11 demonstrates an example of possible disease histories expected within the PLCO trial displaying the consequence screening has on the disease history for each individual and the overall effect that study design may have on screening bias. When designing a nested case-control study with enrollment period T0-T3 within this diagram, notice that the relationship of the case-ascertainment period relative to the preclinical period is important. For example, if a given subject's preclinical period crosses at least one of the case-ascertainment limits (i.e., T0 or T3), screening will influence whether that subject becomes a case in the case-control study. In the diagram, screening has no influence on subjects 5, 6, or 8 as they would always be cases in the study or on subjects 10 and 14 as they would always be potential controls. Subjects 1-3 represent overdiagnosis. These individuals are only diagnosed with disease before death because they are identified by screening and only included in the study as a case because they are detected during the enrollment period. Subjects 4 and 7 are included as cases when screened, but would be eligible only as controls without screening. Alternatively, subjects 11, 12, 13, and 15 are excluded from the study because of screening, but in its absence they would be cases in the study. The screening bias described becomes a problem in risk-factor studies when the influx or efflux of cases is different between the risk factor strata.

In Table 5, the numbers for the 3 year interval prior to the beginning of the study represent percentages for the overall average screening contamination, percentage of ever smokers, and percentage of never smokers who were screen detected during the 3 years prior to the beginning of the case-ascertainment period, respectively. These numbers are for the intervention arm individuals only as we don't have screening information on the usual-care arm during the study. Therefore for the simulation it was assumed that these individuals continued their screening behavior as reported for the 3 years prior to their enrollment in the PLCO randomized trial. Notice that the percentages for the entire population in the 3 years prior to the study are around 50% with an absolute difference between smokers and nonsmokers of about 7 percentage points. Conducting a nested case-control study during this pre-study period would signify the screening behavior and associated bias that would be expected in a population based study.

The numbers for each of the first 3 screening time points ($T_0 – T_2$) indicate percentage of overall average screening compliance (among the intervention arm) for the ever smokers, never smokers, and combined, respectively (Table 5). The screening percentages for compliance at each time point are much higher overall (88.5%; 85.6%; and 84.3%) with smaller differences between screened smokers and nonsmokers when compared to screening behavior reported in the 3 years prior to the beginning of the study. Also, the nonsmokers are more compliant than the smokers during the study where as there were a higher proportion of smokers getting screened in the 3 years before the beginning of the PLCO trial. Conducting a study during the study times $T_0$-$T_2$ would aim to evaluate the effect smoking has on lung cancer in a highly screened group which displays little screening differential between “ever” and “never” smoked strata.

The overall average fraction screened in the intervention arm as a percentage is displayed for the final screening time point ($T_3$), as well as the percentage of ever smokers and never smokers that screened at the last scheduled screening (Table 5). These numbers are treated differently than the numbers for the previous study time points because of the protocol modification to eliminate this final screen after 1998 for nonsmokers. The overall average fraction screened drops to 54.3% with this study time (due to the procedural modification) having the largest absolute screening discrepancy, 52.8%, to occur between smokers and nonsmokers of any of the study times. This study time point will be used to address our artificial design requirement of maximizing the amount of screening bias within a study design. All of the proposed nested case-control study designs are described below.

## *Study Designs*

There were three main thrusts to our case-control study designs. The first design type was specifically created to maximize the amount of screening bias we expect to be affecting the smoking-lung cancer risk estimate. The second types are typical designs created to explore the amount of bias expected within population based case-control studies. These design types were evaluated to determine if it was possible to develop a third type of design in which potential screening bias effects are eliminated or minimized. Figure 12 provides a visualization of the chest x-ray screening schedule during the PLCO trial (in study time) identifying which study years were important for determining the influence of screening on the RR.

*Artificial Design type 1:* One way to artificially create a maximum amount of screening bias in the studies is to maximize the difference in screening behaviors (proportion screened and rate of screening) between the ever and never smoking strata. Due to the PLCO study design modification in the intervention arm, only smokers were required to receive a third annual screen (at T3) after December 1998, thus creating a large difference in screening behavior based on smoking status specifically at that annual screen. This inherent bias was exploited by beginning the case-ascertainment period of our case-control study with the third annual screen and extending the ascertainment window through the follow up period. Within the intervention group population, controls were selected at random with replacement from all members at risk (including future cases) at the failure time of each lung cancer case (subject either screen or symptomatically diagnosed). The simulated RR for this study compares the incidence of lung cancer among the ever smoked group to the incidence of lung cancer among the never smoked group within this sampled intervention population (with a high degree of screening differential).

*Artificial Design type 2:* This design type involves modifying the case-ascertainment screening requirement. Artificial design type 1 was repeated with the following conditions: 1) beginning the case-ascertainment period before T3 and extending through T3; 2) restricting the sampled population to those who were offered a third annual screen at T3; and 3) sampling for artificial design type 1 and 2 from the usual care-arm. It is expected that each subsequent repetition and those designs created sampling from the usual-care arm (that exhibit no effect from the procedural modification) should show a decrease in bias as the number case-ascertainment years increase. This is expected because the difference in screening behaviors (proportion screened and rate of screening) will be reduce between strata when including any study years other than T3.

*Classical Design type 1:* To fulfill the second goal of using traditional study designs, the study populations were selected initially without restriction based on intervention (i.e. identify cases and sample controls from entire cohort). The case-ascertainment period started at the beginning of the PLCO trial enrollment period (study time T0) and varied in length over this time period. Within the population, controls were again selected at random with replacement from all members at risk (including future cases) at the failure time of each lung cancer case (subject either screen or symptomatically diagnosed). The approximated RR simulated here compares the incidence of lung cancer among the ever smoked group to the incidence of lung cancer among the never smoked group within a sampled population from the entire PLCO cohort. Classical study design 1 was also implemented separately using the entire intervention arm only and entire usual care arm only. Again, we will use the contamination rates before the beginning of the PLCO trial as an approximation for the screening patterns of the usual care arm during the study (after the baseline screen at T0).

*Classical Design type 2:* Classical design type 1 was repeated after modifying how we set up the case-ascertainment period. The beginning of the ascertainment period was shifted 1 study-year forward each time until the end of the ascertainment period was T5. There is not expected to be much variation in screening bias when using the usual-care arm. If it were assumed that increased contamination occurred as the PLCO trial length grew for reasons such as subject discovery of study procedures, the simulation could be modified from what is presented here.

## Notation

Dataset names were created with 3 parts to each name. The first part takes into account how we handle the protocol change that indicated nonsmokers were no longer scheduled to receive a 4th screen at study time T3 after 1998 (Figure 12). We use "Post" to represent the fact that the dataset sample was selected after the protocol change was implemented (those who received 3rd annual screen after 1998) – therefore none of the nonsmokers in this cohort where required to receive a 4th screen at study time T3. We use "Ign" to represent the fact that we selected our dataset members from all calendar enrollment years (93-01) ignoring the protocol change, so that some of the nonsmokers in this dataset would have been scheduled to received a 4th screen at T3.

The second part of the dataset name represents whether the individuals were sampled from the intervention arm of the PLCO trial (i.e., "i"), from the usual-care arm of the PLCO trial (i.e., "c"), or from both arms (i.e., "ic") under the condition that cases be equally selected from each arm.

The last part of the dataset name represents the enrollment period (in terms of study years) of our nested case-control study. For example, "t3_t4" means that the nested case-control study begins its enrollment at study year t3 and ends its enrollment at t4, so the study duration is 2 years.

## Purpose of different datasets

The idea behind using the different datasets is to identify how the case-control study design influences the expected amount of screening bias affecting our risk estimate of interest. The nested case-control study design components to be isolated and associated with screening bias include the specific cohort that is selected by sampling individuals from the entire PLCO study, and the length of case ascertainment period. The location of the case-ascertainment period within the larger PLCO randomized trial will be manipulated for the study designs here in order to exploit the previously explained procedural modification for screening. All the datasets were created with 200 cases (if obtainable during case ascertainment period) and 4 times as many controls. This means that there is equal precision in each study and that different study designs can be compared to one another to determine design influence and to aid in the validation step of the research.

Study group numbers 1-4 (Table 6) were created to produce a large difference in screening behaviors (i.e., proportion screened and screening rate) between the never smokers and ever smokers. These studies include the study year T3 and focus on the intervention cohort as there is a "built in" large difference in screening between the two smoking strata due to the aforementioned protocol modification which affected those with a scheduled third annual screen after 1998. Within these 4 studies, we expected the amount of screening bias to decrease as the individuals receiving their third annual screen after 1998 were included (change from “Post” to “Ign”), and as more study years were added to the nested case-control enrollment period (change from “t3_t4” to “t3_t5”). It was assumed that once individuals are no longer scheduled to receive annual screens (after study time T3) they will revert back to their pre-study screening behavior where smokers and nonsmokers reported similar screening behaviors. The reported screening behavior is at least as great if not greater before the nested case-control enrollment period (t1, t2) then it is during the enrollment period (t3, t4, and sometimes t5), therefore the simulations were expected to indicate that in these studies more individuals are shifted out of the study because they have been screen detected before the enrollment period. The largest observed difference between smoking strata during the study occurs at study time T3 therefore it would be expected that our observed risk estimate be larger than it would be if differential screening behaviors weren’t present.

Study group numbers 5-8 (Table 6) were created to represent the level of bias expected in a population based case-control study. These study designs are the same as 1-4, but are carried out in the usual-care group. There are no screening histories for this usual-care group during the study, so the screening behavior reported on the baseline questionnaire (from the 3 years prior to the beginning of the study) will be used for any nested case-control study conducted within this group. Here it was anticipate that the effect of screening would not change much for the different designs. In this group any change would indicate how the length of the case-ascertainment period influences the expected amount of screening bias (e.g., what happens to the simulated amount of screening bias when increasing the enrollment period from 2 years (i.e., t3_t4) to 3 years (i.e., t3_t5)). This observed change should be almost identical when comparing "Post" and "Ign" datasets. When sampling our dataset from the usual-care group, it was assumed that the screening behavior was constant before and across the PLCO enrollment period (study time T0-T5). Therefore, only a slight variation in the level of screening bias between datasets was anticipated. Also since the screening behaviors reported by the smoking strata are not that different, little bias due to the use of screening in this population is expected with any significant variation contributed to the length change of the nested case-control enrollment period. Based on pre-PLCO study contamination rates (from the 3 years prior to the beginning of the PLCO trial), it was projected that the risk estimate comparing ever vs. never smoked would be biased to appear artificially larger.

Study group numbers 9-16 (Table 6) again involve study year T3 so there will be a large difference in the screening behaviors when comparing smokers to nonsmokers, however, T3 occurs at the end of the enrollment period of these studies (i.e., t1_t3, t2_t3). Because the screening behavior difference was larger during the study compared to before the study and a larger difference in screening behavior exists between the smoking strata, it was anticipated that more cases would be shifted into the study among the ever smokers than the never smokers leading to an observed risk estimated which was biased to be larger.

Study group numbers 17, 18, and 19 (Table 6) are set up to be a comparison for one another. In group 17 we are sampling cases and noncases from the entire PLCO cohort over the enrollment years T0 – T5. Cases were selected equally from each group.

We expected there to be little screening bias in this study since the large difference in screening patterns between the smoking strata seen at T3 will be "diluted" by the five other study years. Study groups 18 and 19 are also selected from study years T0-T5 with members from group 18 selected from only the intervention group and 19 selected only from the usual-care group. It was expected that there would be more of a screening effect in group 18 then seen in group 17 or 19 because the screening behavior difference due to the modification between smoking strata was observed in this intervention group at T3. The difference seen at T3 was expected to still be "diluted" in group 18, but more individuals would probably be selected from this study year. Any screening effect seen among study group number 19 would be largely from the length of the case ascertainment period (T0-T5) because there was assumed to be constant screening behavior before and during the study. This study group provides a good estimation of the amount of screening bias expected to affect the smoking lung cancer association in a population based case-control study. It was expected that the observed risk estimate comparing ever vs. never smoked will be larger in the total and intervention group compared to the usual-care group.

Study groups 20-27 (Table 6) were designed to look at what effect the location of the nested case-control study has on screening bias in the study. Again, half of the designs were conducted sampling from the intervention group which provided a risk estimate obtained in an extremely heightened state of screening and the usual-care group which provides a risk estimate that would be expected if conducting a population based case-control study. By creating these incremental studies (moving from T0-T2, T1-T3,

…, T3-T5) in each arm, comparisons could be made within and across groups to determine how variation in screening behaviors affect screening bias while holding other study components constant.  Among the datasets selected from the usual-care group, it was expected that the amount of bias in each study would be relatively the same since the screening behaviors are constant across study time.  Among the datasets selected from the intervention group, again any study involving the study year T3 was expected to have a large amount of screening bias while the other studies would likely have less; but still greater than that seen among the datasets sampled from the usual-care group.  It was anticipated that in both arms the T0-T2 study would produce a risk estimate that was biased to be comparatively smaller due to screening compliance being higher in the nonsmokers.  The rest of the studies (all include T3) were expected to produce observed risk estimates that were relatively larger when creating datasets from the intervention group and relatively smaller when selecting individuals from the usual-care group.

## *Combined Simulation Results*

Simulation values were calculated for both datasets using a combination of study specific and study representative model parameters.  The length of the case-ascertainment period, age structure, and screening proportion and screening rate functions were incorporated based on the dataset.  An estimate of the chest x-ray screening test sensitivity of 86% for lung cancer was obtained from the literature (48, 49) and was used across study designs.  The preclinical duration and preclinical incidence functions are unobservable and the assumptions that were used to create the distributions are stated in the "parameterization of the model" section of the previous chapter.

Chest x-ray screening in the 3 years prior to enrollment for the PLCO randomized trial was collected on a baseline questionnaire and an initial screen scheduled for all study participants. Additionally for the intervention group, it was recorded if an individual was offered and received a chest x-ray screening test at each of the first 3 study years (3 annual screens for study times T1-T3). This information was then used to develop age variable screening proportion and screening rate functions. For the usual-care group during the study and study years after T3 in the intervention group, no significant information was obtained regarding screening behavior during the study. Therefore, it was assumed these individuals would revert back to their screening behaviors as reported on the baseline questionnaire for the 3 years prior to enrollment in the PLCO trial.

This screening information was used to create the 324 simulated RR values from all 27 datasets. The relationship between simulated RR and screening behavior is illustrated in Figure 15. The linear line in these plots was created using the screening proportion and screening rate functions for the mean age of each specific dataset. Based on these results, a downward linear relationship was observed between the simulated RR and screening behavior before the study (Figure 15; top and bottom left side). An upward linear association was observed when looking at the screening behavior during the study where it is stronger when looking at screening proportion rather than screening rate among those screened.

The influence screening behavior differential between smokers and never smokers had on expected bias (i.e., 1-simulated RR) within each study design (Figure 16) was also investigated. The lines in the plot represent the differences in screening behavior (i.e., proportion screened or screening rate) both before and during each nested case-control study design. The shapes in the plot represent 1- the simulation results (i.e., screening bias expected in the corresponding study design) using the specified mode and standard deviation years for the preclinical duration distribution. This plot supports the results illustrated by the previous plots (Figure 15) in that as the screening behavior during the study increases bias also increases(e.g., study designs ign_i_t3_t4 and post_i_t3_t4). Also illustrated is that as the mode and standard deviation years increase, so does the expected bias in the study.

A purpose of designing the 27 different nested case-control studies was to determine the effect that different study designs and parameterizations have on the amount of screening bias expected in the study. Boxplots were created to demonstrate both the sensitivity of the mathematical model to design and parameter variation, and how these design modifications change the simulated RR or amount of screening bias expected in the study (Figure 17). The plot on the upper left hand side suggests that increasing the mode (1,3,5, and 10 years) also increases the expected amount of bias in the study. Although not statistically significant, it is supported by the rising level of the box, whiskers, and median RR value as the mode increases from 1 to 10. The plot on the upper right hand side also suggests an increasing trend for the standard deviation (1,3, and 5 years) and simulated RR. Again there is no statistical significant, but this relationship is supported with rising box level, whiskers, and median RR values. There is the suggestion of a decreasing association between the simulated RR and cohort selection (Figure 17; bottom left) and simulated RR and case-ascertainment length (Figure 17; bottom right). Again, these associations are not statistically significant, but are suggested by a falling box level, whiskers, and median RR value as the variable for cohort selection changes from intervention group to usual-care group and for length of the case-ascertainment period that increase from 2 to 3 to 6 years.

## *Discussion*

Twenty-seven different case-control study designs nested within the PLCO randomized trial have been developed as a method to determine a possible study design affect on screening bias within the smoking-lung cancer observed risk estimate. Out of these studies it is observed that as screening proportion and rate before the study increases the expected amount of screening bias (illustrated with simulated RR) in the study decreases (Figure 15). The opposite is suggested when looking at the screening behavior during the study such that as screening proportion or rate increases, so does the simulated RR (Figure 15). A possible explanation for this observed relationship can be found in the description of the types of screening bias given earlier in chapter 3. Recall that the use of screening before the case-enrollment period shifts cases out of the study because they are screen detected before the beginning of the enrollment period, but had they not been screened would have been clinically detected during the enrollment period. For example, when the screening behavior (i.e., proportion screened and screening rate among those screened) in the ever smokers is greater than the never smokers before the enrollment period, the exclusion of cases among the ever smokers will be greater than that among never smokers. If the RR is set up to comparing ever to never smokers, it would be expected that the observed RR during the study be decreased compared to the same RR estimated in an unscreened population.

When illustrating the association for screening behavior differential between smokers and never smokers and expected bias (1-simulated RR), the results demonstrate an increase in expected bias for increasing screening differential during the study. This demonstration of the relationship between screening behavior differential and expected bias provides evidence that the mathematical model is incorporating screening differential correctly based on theoretical definitions of screening bias as discussed in chapters 3.

The effect that different case-control study designs have on the amount of expected screening bias is illustrated in Figure 17. The boxplots for the preclinical duration parameters (Figure 17; top left and right) suggest that as the preclinical duration of lung cancer increases the expected amount of screening bias in the observed RR increases. This result was anticipated both because it was suggested in previous work (as stated in chapter 1) and as the time spent in the preclinical stage increases so does the chance of being screen detected. As the preclinical duration increases it becomes less likely that the time of screen detection and the theoretical time of symptomatic or clinical detection occur during the same interval (e.g.., observed screen detection time and counterfactual symptomatic detection time both happen before the beginning of enrollment). Case-control studies designed sampling from the usual-care group are indicated to have less expected screening bias than studies using the intervention group. Based on the screening behavior results already discussed, this outcome is expected because the intervention group is simulated to have a much higher screening proportion. As the case enrollment period increased from 2 to 3 to 6 years, the simulated RR decreased which is supported by the result of an increase in screening bias with increase in preclinical duration. When discussing the possible explanation for that bias-preclinical duration association, it was stated that as the time of screen detection and theoretical time of symptomatic detection become less likely to occur during the same interval the amount of screening bias expected increases. Here the increasing case-ascertainment period length increases the probability of the two detection times occurring during the same interval thereby reducing the expected screening bias.

Designing different case-control studies provides a method for comparison. The fact that they are nested in the same randomized trial provides additional benefits to expect any type of design bias that would occur during the study in the population should be minimized or at least very similar across studies. Using a comparison between such similarly designed studies gives us a great opportunity to explore the relative amount of bias that may be arising during a study and its sources.

Comparing the datasets that ignore the procedural modification which eliminated the third annual scheduled screen at study time T3 for nonsmokers (dataset names that begin with Ign) to those that sample only after the modification (dataset names that begin with Post), it was expected that there would be a minimal difference given all other parameters being equal. The thought was that the studies ignoring the modification would have slightly less bias because the difference in screening behavior between smokers and nonsmokers would be "diluted" with the additional sampling years. Looking at the simulated RR values calculated for each of the 27 datasets (Figure 13), the opposite trend is suggested where the studies using datasets that ignore the screening modification are expected to be affected by more screening bias than those sampled only after the modification. This unexpected relationship may be an artifact of the small number of nonsmokers affected by the procedural modification.

The location of the case-ascertainment period within the PLCO trial (i.e., T1-T3, T2-T3, etc.) was expected to affect the simulated amount of screening bias. It was anticipated that studies with enrollment periods X-T3 would have lesser amounts of bias than studies with enrollment periods T3-X or T0-T5 all other model and design parameters being equal. This was assumed that although the screening proportions and rates were higher during the period when individuals were receiving annual screens (T1-T3) the compliance rates during the study would be more alike than the proportion screened and rate at which they screen as reported on the questionnaire for during the 3-years before the beginning of the enrollment period; thus, creating less of the screening behavior differential between smoking strata needed for screening bias to occur. There is no clear trend suggested by the plot of simulated RR values (Figure 13) to infer a relationship with only a small indication within the intervention group that as the nested case-control study enrollment period (comparing enrollment periods of the same length) is shifted later in the study (i.e., form T0-T2 to T1-T3 to T2-T4 to T3-T5) the expected amount of bias increases.

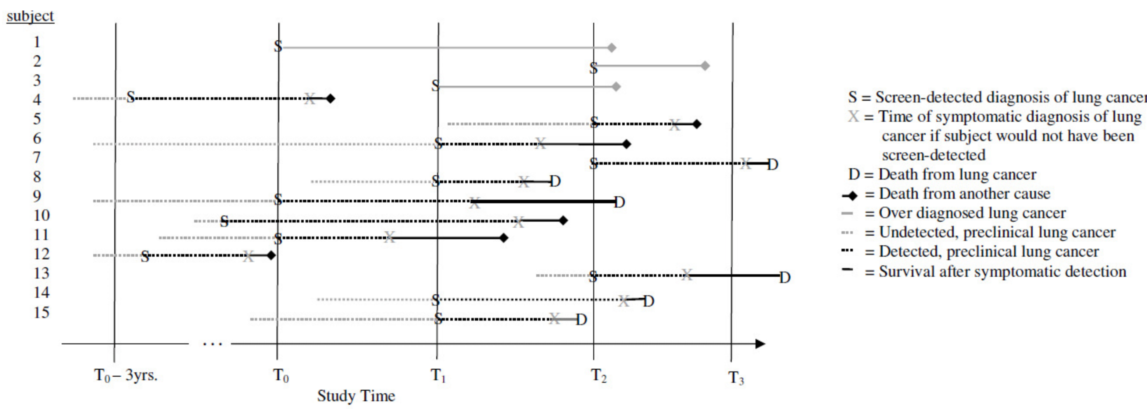



**Figure 11.** Representation of the influence screening has on the natural history of disease and effect it would have on the selection of individuals into a study with an enrollment period T0 – T3. Each subject moves through 3 subsequent states: a disease free state (from birth to detectable, preclinical disease onset), a preclinical disease state (from detectable, preclinical disease onset to date of detection), and a disease state (from date of detection to death). In the diagram, screening has no influence on subjects 5, 6, 8,9,11,13-15 as they would always be cases in the study whether screened or symptomatically detected or on subject 12 as he/she would never be in the study. Subjects 1-3 (represent overdiagnosis because would never be symptomatically detected to have disease before death from a cause other than the disease) and 7 are included in the study as cases when screened and potential control without screening. Alternatively, subjects 4 and 10 are excluded from the study because of screening, but in its absence they would be cases in the study.

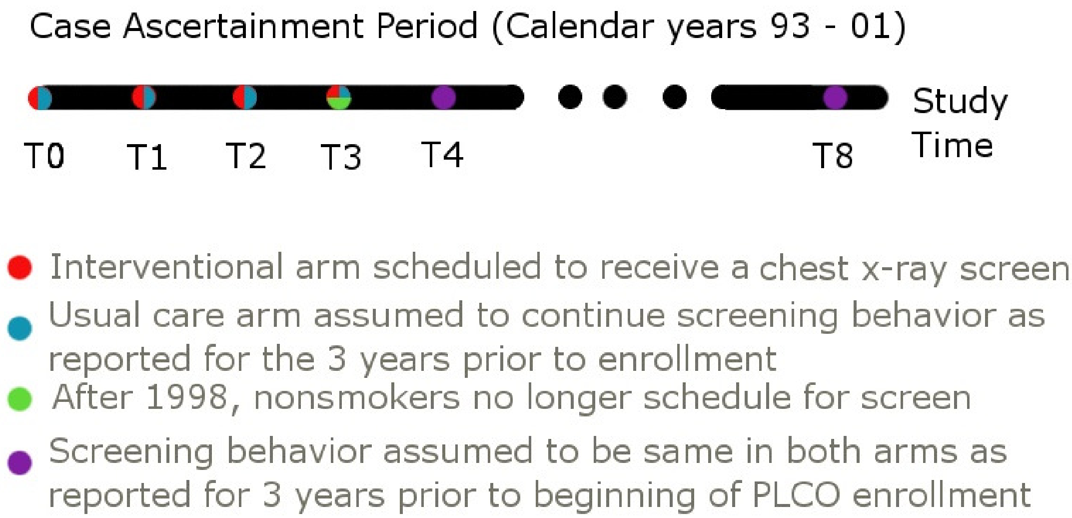



**Figure 12.** Illustration of the case-ascertainment period for the PLCO randomized trial with identification of study years and screening protocols. Individuals in the intervention arm are scheduled to receive 4 total chest x-ray screens based on the initial protocol with a 1998 modification reducing the total number screens offered to nonsmokers to 3. For the purposes of simulating screening bias here in any study year where screening information was not collected, it was assumed that individuals would continue screening behaviors as reported on the baseline questionnaire for before the beginning of the trial.

| | Average Screening Contamination Within 3 Years Prior to Enrollment (%) | Average Screening Compliance at Study Time T0 (%) | Average Screening Compliance at Study Time T1 (%) | Average Screening Compliance at Study Time T2 (%) | Average Fraction Screened at T3 (%) |
|---|---|---|---|---|---|
| Ever Smokers | 56.3 | 87.9 | 84.5 | 83.1 | 78.7 |
| Never Smokers | 49.0 | 89.2 | 86.7 | 85.7 | 25.9 |
| Combined | 52.9 | 88.5 | 85.6 | 84.3 | 54.3 |

**Table 5.** Screening contamination averaged over calendar year calculated as those who screened during 3 year before beginning of case-ascertainment period divided by total population. Screening compliance averaged over calendar year calculated using the intervention group only as those who complied with screening during the first three scheduled screens $T_0$, $T_1$, $T_2$ out of population scheduled for screening. Fraction screened over calendar year represents the percentage in the intervention group who were screened at the last scheduled screen ($T_3$). The large difference in screening at T3 is due to the procedural modification which eliminated this scheduled screen for nonsmokers after 1998.

## Nested Case-control Designs

| Study Group Number | Study Group | Study Period (inclusive) | Comment | Dataset Names |
|---|---|---|---|---|
| 1-4 | intervention group sample including and excluding those scheduled to receive a third annual screen before the 1998 protocol change | Post protocol change<br>T3 – T4, T3 – T5<br><br>Ignore protocol change<br>T3 – T4, T3 – T5 | compared the incidence among the ever smokers to the incidence among the never smokerswithin this sample population of the intervention group | Post_i_t3_t4<br>Post_i_t3_t5<br>Ign_i_t3_t4<br>Ign_i_t3_t5 |
| 5-8 | usual care group including and excluding those scheduled to receive a third annual screen before the 1998 protocol change | Post protocol change<br>T3 – T4, T3 – T5<br><br>Ignore protocol change<br>T3 – T4, T3 – T5 | compared the incidence among the ever smokers to the incidence among the never smokers within the sample population of the usual-care group | Post_c_t3_t4<br>Post_c_t3_t5<br>Ign_c_t3_t4<br>Ign_c_t3_t5 |
| 9-16 | Repeated 1-9 using different study periods (including and excluding those scheduled for screen after 1998 for intervention and usual care group separately for each new study period) | **Intervention**<br>Post protocol change<br>T1 – T3, T2 – T3<br>Ignore protocol change<br>T1 – T3, T2 – T3<br>**Usual care**<br>Post protocol change<br>T1 – T3, T2 – T3<br>Ignore protocol change<br>T1 – T3, T2 – T3 | Each subsequent repetition should show a decrease in bias as the number of required screens are decreased because the difference in screening behaviors (proportion screened and rate of screening) will be reduce between strata (largest expected bias in Post_i_t2_t3) | Post_i_t1_t3<br>Post_i_t2_t3<br>Ign_i_t1_t3<br>Ign_i_t2_t3<br>Post_c_t1_t3<br>Post_c_t2_t3<br>Ign_c_t1_t3<br>Ign_c_t2_t3 |
| 17 | Sample from entire PLCO population | T0-T5 | compare the incidence of lung | Ign_ic_t0_t5 |

| | | | | |
|---|---|---|---|---|
| | | | cancer among the ever smokers to the incidence of lung cancer among the never smokers within this sample population from the entire PLCO cohort | |
| 18 & 19 | 1) Intervention group<br>2) Usual care group (do 1 study using each group = 2 total) | **Intervention**<br>T0 – T5<br><br>**Usual care**<br>T0 – T5 | 1) to identify the smoking-lung cancer association among a population with a high proportion of screening<br>2) to identify the smoking-lung cancer association that would be expected within the general population | Ign_i_t0_t5<br>Ign_c_t0_t5 |
| 20 - 27 | Repeat 20 & 21 starting at different points in the study period | **Intervention**<br>T0 – T2;<br>T1 – T3;T2 – T4; T3 – T5<br><br>**Usual care**<br>T0 – T2;<br>T1 – T3;T2 – T4; T3 – T5 | These study design modifications will indicate the potential influence of the location of the case-ascertainment period of a nested-case control study within a randomized control trial on the amount of screening bias | **Intervention**<br>Ign_i_t0-t2<br>Ign_i_t1-t3<br>Ign_i_t2-t4<br>Ign_i_t3-t5<br><br>**Intervention**<br>Ign_c_t0-t2<br>Ign_c_t1-t3<br>Ign_c_t2-t4<br>Ign_c_t3-t5 |

**Table 6.** Naming scheme for sampled datasets used for the nested case-control designs showing the study groups and their numbers, study periods, and dataset names. The table also contains a comment describing what each set of study designs were used to estimate.

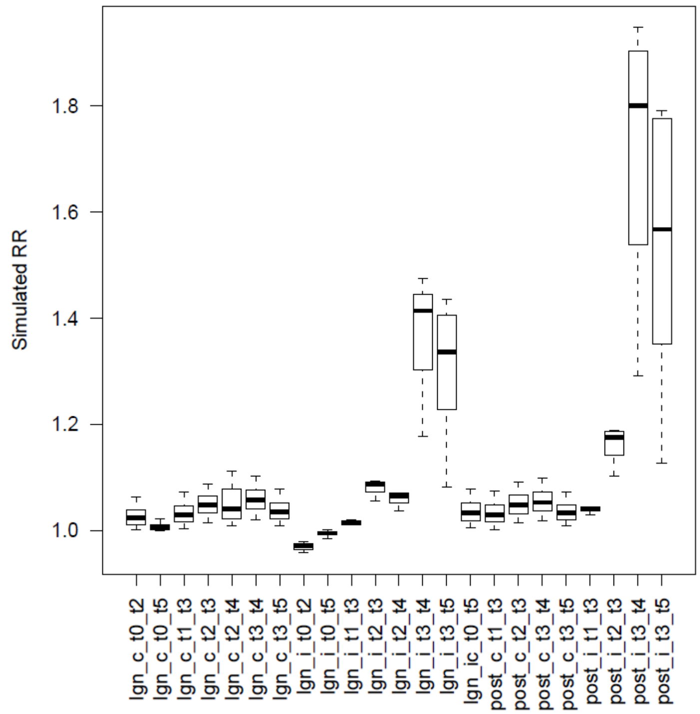



**Figure 13. G**raph illustrates simulated RR range based on preclinical duration distribution parameter variation (12 combinations of mode = 1,3,5,or 10 and standard deviations= 1,3,5) for each of the 27 study designs. The solid black line in the box represents the median, the ends of the box represent the 25$^{th}$ and 75% percentiles, and the tails extend to the 2.5$^{th}$ and 97.5$^{th}$ percentiles.

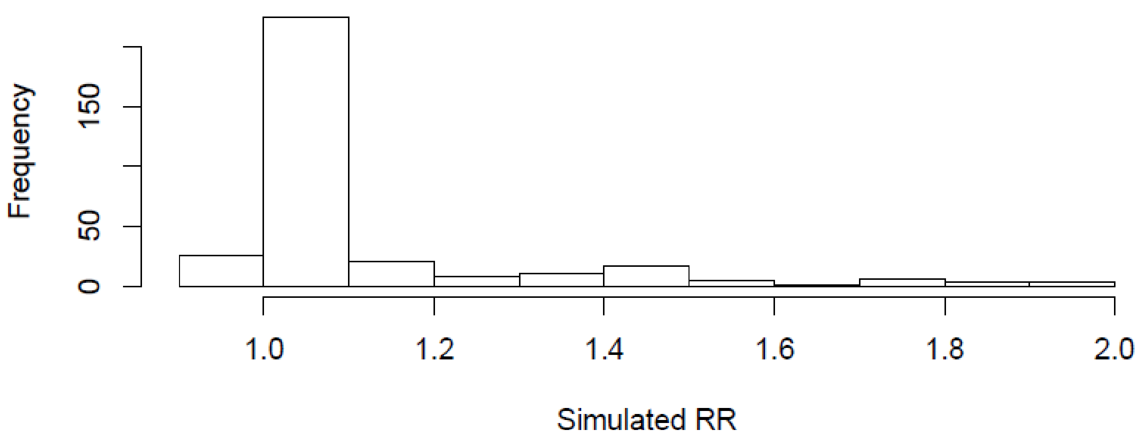



**Figure 14.** The histogram illustrates the distribution of the 324 simulated RRs (ever smoked verse never smoked risk for lung cancer diagnosis) based on the 27 different datasets when using 12 combinations of modes of 1,3,5, and 10 years and standard deviation of 1, 3, and 5 years for the preclinical duration distribution for each design. Based on the assumptions and double null hypothesis (screening and smoking effects are independent of lung cancer) of the simulation, a value of 1 indicates no expected screening bias.

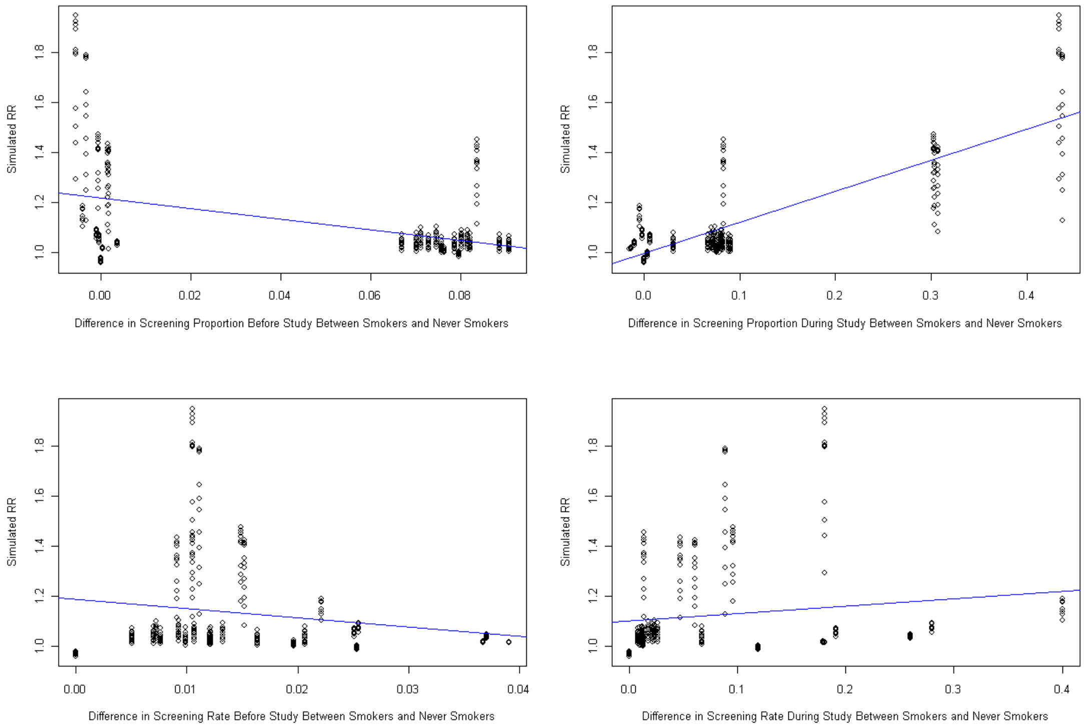



**Figure 15.** Graphs that shows the linear relationship average differential screening behavior between ever smokers and never smokers has with the simulated RR. The plots on the left illustrate the relationship between simulated RR and the difference in screening proportion before the study between ever and never smokers (top) or average difference in screening rate before the study between those smoking strata (bottom). The plots on the right demonstrate the linear relationship between simulated RR and the difference in screening proportion during the study between ever and never smokers (top) or difference in screening rate during the study between those smoking strata (bottom). Because the screening proportion and rate functions are age dependent, the mean age for each dataset was used when calculating the difference values for the plot.

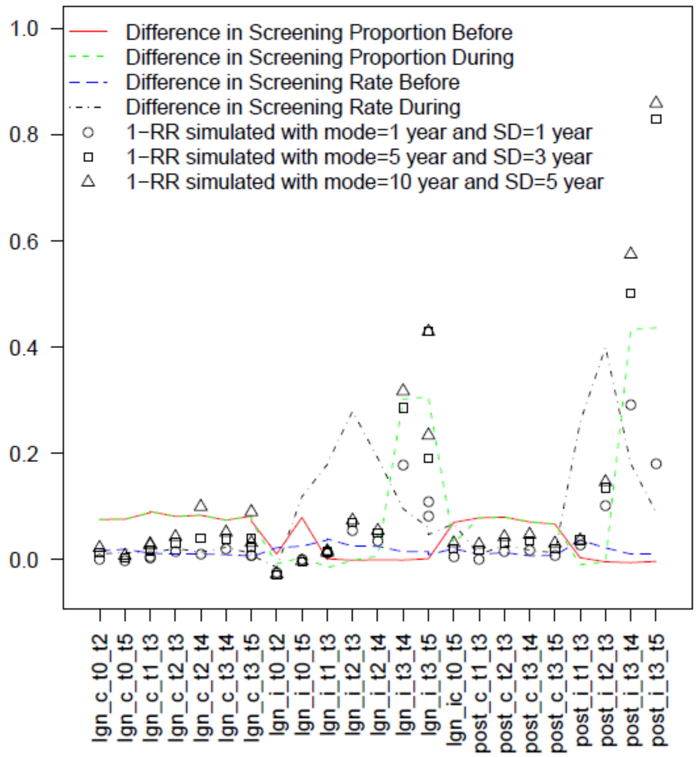



**Figure 16.** Representation of the relationship of screening behavior differential between smokers and never smokers and expected bias (i.e., 1-simulated RR) by study design. The lines in the plot represent the differences in screening behavior (i.e., proportion screened or screening rate) both before and during each nested case-control study design. The shapes represent 1- the simulation results (i.e., screening bias expected in the corresponding study design) using the specified mode and standard deviation for the preclinical duration distribution.

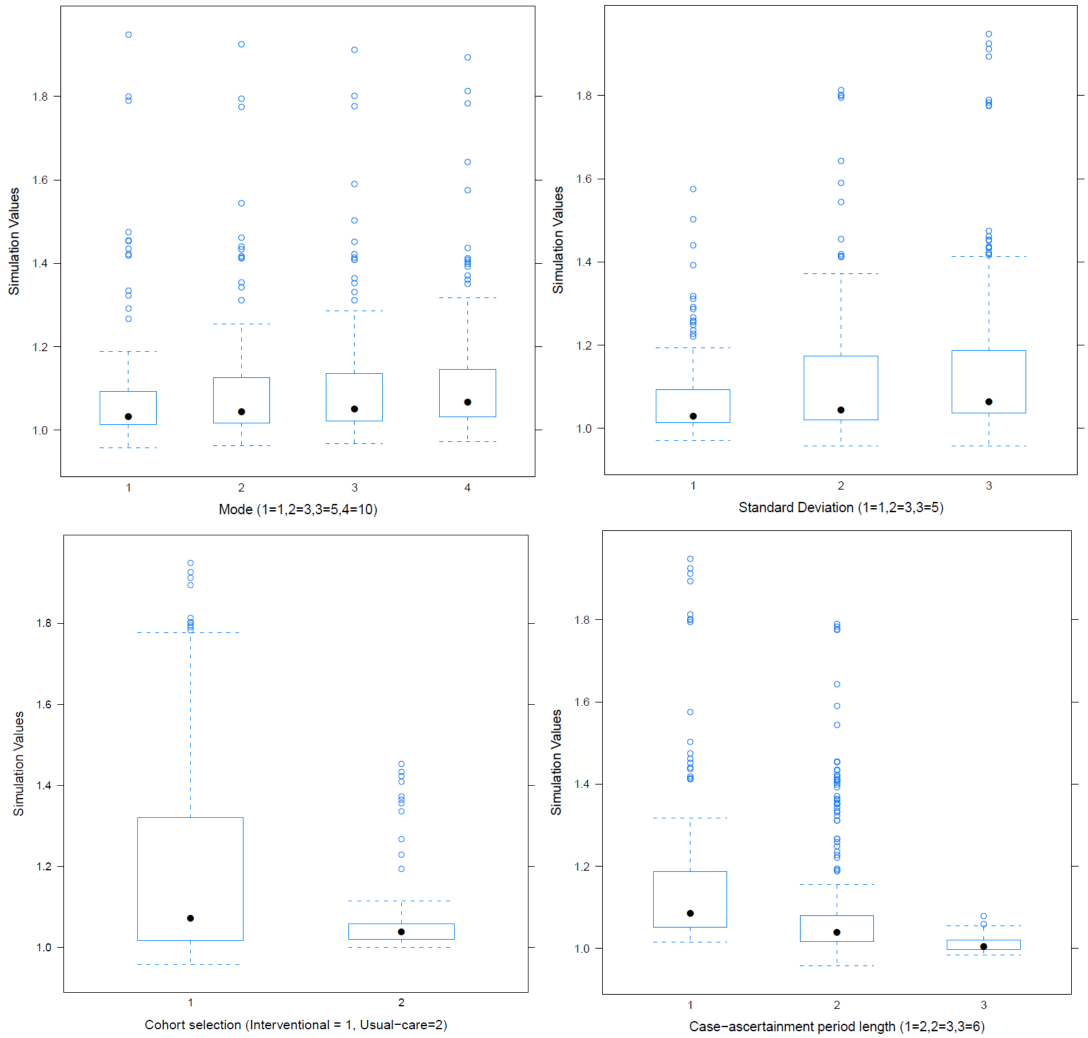



**Figure 17.** Boxplots illustrate range of simulated RRs for the mode (top left), standard deviation (top right), cohort (bottom left), and length of the case-ascertainment period (bottom right). There are 324 (12 for each of the 27 study design) simulated RR obtained by using each of the mode and standard deviation combinations. The modes used in the simulations were 1,3,5, and 10 years; the standard deviations were 1,3, and 5 years; study population sampled from either intervention group or usual-care group; and the length of the case-ascertainment period varied from 2, 3, to 6 years.

# Chapter 5: Do the simulated model results and empirical observations correspond and under what conditions?

## *Model Validation Method*

Up to this point, it has been assumed that the mathematical model predicts the amount of expected screening bias affecting each observational risk-factor study accurately. Within the limits of accuracy afforded by the sample size, this assumption can be explored for case-control studies of lung cancer nested within the PLCO randomized trial because of the comparability between studies offered by nesting. For factors it is designed to study, the randomized trial theoretically eliminates all potential sources of baseline confounding prior to the beginning of the study offering an unbiased estimate of the risk attributed to the variable under study (with regard to lung cancer in the PLCO trial, this factor is the developed chest x-ray screening protocol of annual exams compared to usual screening behavior). When the variable of interest is changed, as when conducting studies nested within a randomized trial, an unbiased estimate should still be attainable after accounting for the measured variables within the randomized trial (i.e., in this example it is screening). A mathematical model has been developed and used here to account for the measured screening variable and the differential behavior demonstrated between the ever smoked and never smoked groups.

To begin the validation process for the mathematical model which was created to determine the amount of screening bias present in an observational study, empirical comparisons between each combination of the 27 different study designs (24 unique designs) were made (see appendix). For easy of illustration, only 4 study designs representing the extreme forms of potential bias (i.e., little to no bias and large amount of bias) will be shown in the chapter. The model would be considered valid when the amount of screening bias the simulation suggested correlated with the expected screening bias within a study. The expected bias was determined by a relative comparison (ratio) between the logistic regression estimated *RR* from two nested case-control studies. This ratio was then compared to the corresponding simulated screening bias ratio; with a final ratio that equals 1 providing evidence to support the simulated model and its assumptions hold. Further work must be completed to determine the distribution for acceptable values expected around this single point estimate for complete validation of the mathematical model.

With each nested case-control study design the actual data were analyzed to provide an estimate of the *RR* between smoking and lung cancer (i.e., $RR_{\text{observed}}$). For each study design, the dataset was sampled 100 times repeating the logistic regression analysis each time. A median *RR* was calculated for all 100 datasets (i.e., $RR_{\text{median}}$). The theoretical *RR* was simulated under previously identified assumptions about disease history adjusting model parameters such as case-ascertainment period length and age distributions to represent each of the 27 different study designs, denoted as $RR_{\text{simulated}}$. The ratio between estimates of $RR_{\text{median}}$ from two different study designs was compared to the corresponding ratio of the $RR_{\text{simulated}}$ for those two study designs; with a final ratio (i.e., *V*) of 1 representing a valid simulation method and model for the nested studies. A comparison was made between all study designs in an attempt to identify a predictable pattern. The equations in the process are as follows.

The two by two table was set up in the following way:

| | Lung cancer | All members at risk |
|---|---|---|
| Ever smoked | *a* | *b* |
| Never smoked | *c* | *d* |

$$RR_{observed} = \frac{a/b}{c/d}$$

$$ORDER(RR_{oberved-1,} \dots RR_{observed-100}) = \frac{(RR_{observed-50} + RR_{observed-51})}{2} = RR_{median}$$

$$\frac{RR_{median-study\ 1}}{RR_{median-study\ 2}} = RR_{median-1-2}$$

$$Simulation\ Value = f(x, \dots)$$

$$\frac{f(x, \dots)_{ever\ smoked}}{f(x, \dots)_{never\ smoked}} = RR_{simulated}$$

$$\frac{RR_{simulated-study\ 1}}{RR_{simulated-study\ 2}} = RR_{simulated-1-2}$$

$$\frac{RR_{simulated-1-2}}{RR_{median-1-2}} = V$$

A variance for each *V* was estimated as a way to determine if the value was close enough to 1 to indicate that the mathematical model was valid. The delta method for moments of random variables (61) was applied here to estimate the variance of *V*. The variance formula for the *V* is present below. We also calculated a chi-squared type value, $\chi^2$, as a metric of closeness between $RR_{\text{simulated-1-2}}$ and $RR_{\text{median-1-2}}$ using the formula $\chi^2 = (RR_{\text{median-1-2}} - RR_{\text{simulated-1-2}})^2 / RR_{\text{simulated-1-2}}$. The variance estimate is subtracted from this chi-squared type value to determine if the value is in acceptable range of 0 (i.e., mathematical model predicts amount of screening bias $RR_{\text{observed}}$ perfectly).

For linear functions g of a random variable X

$$E[f(X)] = f(E[X])$$

The function is often approximated with a truncated Taylor series expansion such as

$$f(x) \approx f(a) + f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2!} + \cdots$$

Applying the delta method to a function of two random variables X and Y about values $x_0$ and $y_0$

$$f(x,y) = f(x_0, y_0) + \frac{\partial f(x,y)}{\partial x}|_{x_o y_o}(x - x_0) + \frac{\partial f(x,y)}{\partial y}|_{x_o y_o}(y - y_0) + \cdots$$

If the function is a ratio of two random variables, f(x,y) = x/y, then,

$$\frac{\partial f(x,y)}{\partial x} = \frac{1}{y}$$

$$\frac{\partial f(x,y)}{\partial y} = \frac{-x}{y^2}$$

And the first moment or mean will be approximately:

$$E\left[\frac{X}{Y}\right] \approx \frac{\bar{x}}{\bar{y}} \approx \frac{\mu_x}{\mu_y} + \frac{-\mu_x}{\mu^2{}_y}\bar{y} + \frac{1}{\mu_y}\bar{x}$$

And the second moment or variance will be approximately:

$$Var\left(\frac{X}{Y}\right) \approx \frac{\mu_x^2}{\mu_y^4}Var(Y) + \frac{1}{\mu_y^2}Var(X) - \frac{2\mu_x}{\mu_y^3}Cov(X,Y)$$

And after substituting the sample variance into the equation get:

$$Var\frac{\bar{x}}{\bar{y}} \approx \frac{\mu_x^2}{\mu_y^4}\frac{\sigma_y^2}{n} + \frac{1}{\mu_y^2}\frac{\sigma_x^2}{n} - \frac{2\mu_x}{\mu_y^3}\frac{\rho\sigma_x\sigma_y}{n}$$

With the estimated variance of a ratio estimator given by:

$$\widehat{Var}\frac{\bar{x}}{\bar{y}} \approx \frac{1}{n}\left[\frac{\bar{x}^2}{\bar{y}^4}s_y^2 + \frac{1}{\bar{y}^2}s_x^2 - \frac{2\bar{x}}{\bar{y}^3}\hat{\rho}s_x s_y\right]$$

For our case, the first step was to get the mean $RR_{observed}$ of the 100 samples of each of the 27 study designs. Then the estimated variance equation above was applied to each of the 377 ratio combination of observed *RR* ratios (i.e., Ign_c_t3_t5 $RR_{\text{observed}}$/Ign_i_t2_t3 $RR_{\text{observed}}$, Ign_c_t3_t5 $RR_{\text{observed}}$/Ign_i_t1_t3 $RR_{\text{observed}}$, etc.) such that n = 100, = mean of 100 samples of $RR_{\text{observed}}$ and $s_x^2$ = standard deviation for the first study, $\bar{y}$ = mean of 100 samples of $RR_{\text{observed}}$ and $s_y^2$ = standard deviation for the second study. The datasets were considered to be approximately independent and therefore the covariance factor is 0. The equation that was used to estimate the variance of *V* (ratio of $RR_{\text{simulated}}$ for two studies / ratio of $RR_{\text{observed}}$ for same two studies) can now be given. Because $RR_{\text{simulated}}$ is a constant the following property of a variance can be used

$$Var[cX] = c^2 Var(X)$$

Such that

$$Var[V] \approx \left(\frac{RR_{simulated\ study\ 1}}{RR_{simulated\ study\ 2}}\right)^2 \Bigg/ \frac{Var[Mean\ RR_{observed\ study\ 1}]}{Var[Mean\ RR_{observed\ study\ 2}]}$$

$$\approx \left(\frac{RR_{simulated\ study\ 1}}{RR_{simulated\ study\ 2}}\right)^2 \Bigg/ \frac{\frac{1}{n}\left[\frac{\bar{x}^2_{study1}}{\bar{y}^4_{study1}} s^2_{y-study1} + \frac{1}{\bar{y}^2_{study1}} s^2_{x-study1}\right]}{\frac{1}{n}\left[\frac{\bar{x}^2_{study2}}{\bar{y}^4_{study2}} s^2_{y-study2} + \frac{1}{\bar{y}^2_{study2}} s^2_{x-study2}\right]}$$

The predicted amount of bias in a study (i.e., $RR_{\text{simulated}}$) was simulated under the following assumptions: 1) screening and smoking effects are independent; 2) screening and smoking are not associated with lung cancer; and 3) the relative change in bias doesn't depend on the underlying incidence rate or histological type. The explanation for the method of how the simulated *RR* value can provide a theoretical correction for the observed *RR* is illustrated below.

In the target population, there is one screening pattern for smokers ($S_1$) and another screening pattern for nonsmokers ($S_2$) that will affect the number of lung cancer cases observed for the sampled ever-smoked group (*a*) and the sampled never-smoked group (*c*).

$a = E(C|S_1)$ and $c = E(C|S_2)$

Under the three assumptions, if *C* represents the distribution of lung cancer cases in the population then,

$$RR_{observed} = \frac{a/b}{c/d} = \frac{a/b}{(c * B')/d} = 1$$

Where $B$ is the screening bias factor:

$$B = \frac{a}{c}$$

When comparing $RR_{\text{observed}}$ from two study designs under our 3 assumptions,

$$\frac{RR_{observed-1} * B_1}{RR_{observed-2} * B_2} = 1 \quad \frac{RR_{observed-1}}{RR_{observed-2}} = \frac{B_2}{B_1}$$

In other words, if all the assumptions hold then any difference in the ratio of the study-observed $RR$s should be explained with the ratio of the bias terms assuming all other forms of bias are either nonexistent or equal between the two studies. Both $RR_{observed}$ are calculated from the data and if one of the bias terms (either $B_1$ or $B_2$) is known with a great degree of certainty, the other bias value can be estimated with the following equation.

$$\frac{RR_{observed-1}}{RR_{observed-2}} * B_1 = B_2 = B_{expected}$$

If the $RR$ among the unscreened strata is used for comparison to all other $RRs$, $B_1$ would be expected to be 1; this $RR_{observed}$ is expected to contain no screening bias. Therefore, relevance is given to the above equation. In this situation, $B_{expected}$ can provide an additional way to validate our simulation model such that $B_{expected} \approx RR_{simulated}$. When the screening bias factor, $B$, is unknown in both studies, the parameters are estimated with $RR_{simulated}$ and a validation method such as described above must be used as is the case here.

## *Results*

### Combined Simulation Results

Parameterization of the mathematical model is as described in Chapter 4. The $RR_{simulated}$ values calculated for chapter 4 are used here for model validation and displayed in Table 6. The range of simulated *RR* values is 0.96 for the study design sampling any intervention group individuals at risk for lung cancer from enrollment date through second year of the study (i.e.,Ign_i_t0_t2) to 1.95 for the study design sampling intervention group individuals at risk for lung cancer from the third year of the study through the fourth year (post_i_t3_t4). Across study designs, the average *RR* is 1.13 and the median *RR* is 1.04 (Table 7). Looking only at the results from the simulations using a lognormal distribution for the preclinical duration with mode of 1 year and a standard deviation of 1 year, the group has a mean of 1.05 and a median of 1.01. Restricting the data to the results of the simulations that used a mode of 5 years and standard deviation of 3 years and again with a mode of 10 years and standard deviation of 5 years, the mean and median *RR*s are 1.14 and 1.04, 1.17 and 1.08, respectively.

### Combined Logistic Results

A logistic regression model (2) was used to get a *RR* for the risk smoking has on the development of lung cancer after adjusting for age (Table 8; Figure 12 and appendix

Table a-8) where $\pi(x)$ = case, $\alpha$ = the y-intercept, $x_1$ = categorical smoking variable (ever or never smoked) with $\beta_1$=risk of smoking on lung cancer diagnosis for given age, and $x_2$ = continuous age variable (55-74) with $\beta_1$=risk of age on lung cancer diagnosis for given smoking category, and $\varepsilon$ = the error associated with each $\beta$ estimate.

$$logit[\pi(x)] = log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i \quad (2)$$

A simple model is acceptable here because the goal is to illustrate possible affects that screening use has on the selection of cases into a case-control study and not to produce the most accurate estimate of the risk smoking has on the development of lung cancer. The focus here is on the *RR* of smoking, age is only included in the logistic model because it is stratified in the simulation. The 4 median RRs range from 12.53 to 16.12 with average Wald 95% CIs from (12.04,13.18) to(16.12,17.41), and empirical 95% CLs from(8.93,12.31) to (10.24,26.56) (Table 7; Figure 18). These estimates were calculated for the smoke variable from each of the 4 datasets. The values for all 27 median are similar with *RR*s range from 10.52 to 16.34 with average Wald 95% CIs from (10.03,11.18) to(16.14,17.41), and empirical 95% CLs from(8.92,12.41) to (10.89,25.43) (appendix Table a-7; Figure a-18) The medians along with the empirical 95% confidence limits were obtained by finding the values based on ordering the 100 samples of each dataset. The average Wald 95% confidence intervals were calculated by using the mean *RR*s and their mean standard errors for the 100 samples as illustrated in the following formula (3):

$$Average\ Wald\ 95\%\ CI = mean(RR_{observed}) \pm 1.96 * mean(SE) \quad (3)$$

## Comparison between Simulation and Logistic Regression

As mentioned above, the mathematical model will be explored for validity in estimating the expected amount of screening bias due to use of chest x-ray in a population sampled for a case-control study of lung cancer. During this process we will try to identify the optimal set of parameters used here for the mathematical model. Using the comparison method (i.e., comparing the ratio of simulated *RR* for two studies to the ratio of their observed *RR*s) will require the assumption that the only difference between two observed *RR* values estimated with logistic regression is due to screening bias. This means that all other forms of bias are assumed either to be negligible or equal across study design. The assumption here is that any change in the observed *RR* from one study to the next is due to screening (if there is additional bias it is equal across study designs) therefore the ratio of the $RR_{observed}$ for the two studies should be equal to the ratio of the $RR_{simulated}$. Making these assumptions is acceptable to the authors for the purposes of developing and testing the mathematical model, but in the case where estimating risk becomes the goal these other bias types would need to be explored and accounted for in the analyses.

The graph comparing the $RR_{\text{simulated}}$ values to $RR_{\text{observed}}$ values (Figure 20 and appendix Figure a-21) demonstrates the correlation between the two values where a linear diagonal line would represent perfect correlation. For each study design, there is one estimated $RR_{\text{observed}}$ and 12 $RR_{\text{simulated}}$ due to using several different mode years (1,3,5,10) and standard deviation years(1,3,5) for the lognormal preclinical duration distribution. So, each point represents the combination of $RR_{\text{simulated}}$ and $RR_{\text{observed}}$ where the vertical spread of the points represents the various $RR_{\text{simulated}}$ values for each $RR_{\text{observed}}$ value. In this plot there doesn't appear to be any observable pattern moving from one $RR_{\text{observed}}$ value to the next suggesting there is no correlation between the two *RR* values. This same conclusion results when looking at the relationship between $RR_{\text{simulated}}$ and $RR_{\text{observed}}$ by selected mode and standard deviation year combinations (plots not shown).

The ratio of the $RR_{\text{simulated}}$ ratio for a pair of study designs to the $RR_{\text{observed}}$ ratio for those same designs (i.e., *V*) (Figure 21; left column) are variable with a range from about 0.5 to 2. When looking at all 27 study designs, these result are supported where values are highly variable across study design with a range from about .45 to 2.18 and even within study design the smallest range was about 0.5 and largest range about 1.25. Across all two study design comparisons, the mean *V* is 1.07 and the median is 1.03. Here a value of 1 represents perfect agreement between the $RR_{\text{simulated}}$ and $RR_{\text{observed}}$ ratios for a two study design comparison. In these plots, the vertical range for each study design across the x-axis represents the different *V* values for each comparison to the second study design (e.g., at Ign_i_t3_t5 on the x-axis, one point represents the *V* for Ign_i_t3_t5 and Ign_c_t3_t5 comparison, another point represents *V* for the Ign_i_t3_t5 and post_c_t3_t5 comparison, etc.). The first plot on the left hand side of Figure 21 is of *V* vs study design using a short preclinical duration for lung cancer in our simulation model (mode =1 and standard deviation =1). The second plot on the left hand side of Figure 21 is of *V* vs study design using a medium preclinical duration for lung cancer in our simulation model (mode =5 and standard deviation =3). The third plot on the left hand side of Figure 21 is of *V* vs study design using a long preclinical duration for lung cancer in our simulation model (mode =10 and standard deviation =5). Notice that for all three plots part of the boxplot for each study design crosses 1 (same is true when looking at all 27 designs), giving possible validation to the mathematical model if the same set of parameters is responsible.

The Chi-squared type value plot (Figure 21 and Figure a-22; right columns) provides an idea about how well the simulation model predicts changes in observed *RR*s between two study designs with a value of 0 indicating perfect prediction. As mentioned above, it is assumed that any change in the observed *RR* from one study design to the next is due to screening and that is why the mathematical model presented here, if valid, should predict the observed change in the *RR* values. The summary statistics for the 4 designs are almost identical to the values for the 27 designs over all mode and standard deviations year combinations for the preclinical duration distribution. The Chi-squared values have a range from about 0 up to about .8 with a mean of 0.06 and median of 0.02 across study design. Moving down the right hand column, the plots change from illustrating the chi-sq type statistic for comparisons between study designs with short preclinical durations (mode and standard deviation = 1 year) to medium preclinical duration (mode=5 years and standard deviation = 3 years), to long preclinical durations (mode=10 years and standard deviation = 5 years). As would be expected and is suggested in the plots in the left hand column of Figure 21, the chi-squared type values become more spread out for increasing preclinical duration.

All the estimated variances for *V* occur under a value of .003 with the mean value for the 27 datasets of 0.0008 and a median value of 0.0006 (raw data not shown). Because the estimated variance is so small, repeating the present plots taking into account this estimated variance would not create any noticeable change and therefore those plots will not be presented.

Sum of the Chi-squared type statistic by preclinical duration distribution parameterization (mode years = 1, 3, 5, 10 and standard deviation years = 1, 3, 5) calculated for each of the 6 combinations of the 4 selected study designs types is illustrated in Table 10. The sum of this statistic is used to see how well the ratio of the simulated amount of screening bias between two study designs correspond with the ratio of observed RR between two study designs. This chi-squared type statistic provides a way to find the best parameterization from those we used with a value of 0 for the statistic indicating perfect prediction. The range of sums goes from 0.28 when using a mode and standard deviation of 1 year for the preclinical duration distribution to 1.34 when using a mode of 3 years and standard deviation of 5 years.

## *Discussion*

To validate the mathematical model which was created to determine the amount of screening bias present in an observational study, empirical comparisons between each combination of the 27 different study designs (24 unique designs) were made with four selected designs given in this chapter. The model was considered to give a valid value when the ratio (i.e., $V$) of the simulated risk ratios ($RR_{\text{simulated}}$) for two study designs was equal to the ratio of the observed risk ratios ($RR_{\text{observed}}$) estimated with a logistic model for the same two study designs. The $V$ values were highly variable across study design and after breaking down the results based on preclinical duration in order to identify the best set of model parameters, it is suggested that using a shorter preclinical duration in case-control studies of the risk smoking has on the development of lung cancer results in values closer to that which is expected. The shorter preclinical duration is supported by the distribution of chi-squared type statistics with many more values close to 0 which indicates that the model is predicting what is expected. However, it must me noted here that until further research is conducted to determine what the distribution for this Chi-squared type statistic should be, a full validation of the mathematical model can't be completed.

To give utility to the validation technique used here of calculating $V$ values, it was assumed that any change in the observed $RR$ from one study to the next is due to screening (if there is additional bias it is equal or negligible across study designs) therefore the ratio of the $RR_{\text{observed}}$ for the two studies should be equal to the ratio of the $RR_{\text{simulated}}$. However, this assumption doesn't appear to hold based on Figure 19 and a-19 showing that changes in the $RR_{\text{observed}}$ are not completely driven by screening behavior (proportion screened and rate of screening) differential between smokers and never smokers. Because screening behavior differential is essential for screening bias to occur in observational studies and is the basis of the mathematical model developed here, the PLCO data may not provide the needed situation to validate that the model can predict the amount of screening bias affecting the $RR_{\text{observed}}$.

In the appendix, Figure a-20 representation of the difference in screening rate between smokers and never smokers versus difference in screening proportion between smokers and never smokers for each of the 27 study designs. The plot on the right is the screening behavior difference before the study and the plot on the left is screening behavior difference during the study. The four selected study designs from chapter 5 are identified within each plot providing an illustration that there is one point per study design type. We have chosen the 4 study design types to represent the extremes of the range of potential bias where the datasets sampling from the usual-care group have low levels of screening bias and datasets sampling from the intervention group have high levels of screening bias. From the plot and the identified points it appears that the combination of a small difference in screening behavior before the study and a large difference in proportion screened during the study results in high levels of screening bias (looking at points selected from the intervention group). We also see that a larger difference in proportion screened before the study and smaller difference in the screening behavior during the study results in low levels of screening bias (looking at points selected from the usual-care group). Because the difference values in screening behavior before the study are so small, the observation that studies selected from the usual-care group have higher difference in screening proportion doesn't support a conclusion that large differences in screening proportion before the study reduce bias. Thus, these results correspond to what we expect, that as the differential in screening behavior between risk factor strata increases so does the screening bias.

It appears that differential screening behaviors demonstrated between smokers and never smokers in these nested case-control studies of lung cancer diagnosis do influence observed *RR*s. However, there are likely other types of bias (besides screening bias) also influencing these observed *RR*s differentially making validation of the mathematical model using the described empirical comparison technique difficult. The results indicate that using a shorter preclinical duration in the simulation may provide more accurate screening bias prediction in these studies, but there are some indications (i.e., $V$ not close enough to one or chi-squared type statistic not close enough to 0) to the contrary. It is recommended that the mathematical model be fully validated by developing a distribution for the chi-squared type statistic and using additional techniques and additional data before using the simulation results as a method to provide an unbiased observed *RR* with respect to screening bias.

| study_design | Simulated RR | Mode | Standard deviation |
|---|---|---|---|
| lgn_c_t3_t5 | 1.01 | 1 | 1 |
| Study population sampled from the usual-care group during all PLCO calendar enrollment years (93-01) from between third and fifth study year | 1.02 | 1 | 3 |
| | 1.04 | 1 | 5 |
| | 1.02 | 3 | 1 |
| | 1.03 | 3 | 3 |
| | 1.05 | 3 | 5 |
| | 1.02 | 5 | 1 |
| | 1.04 | 5 | 3 |
| | 1.06 | 5 | 5 |
| | 1.03 | 10 | 1 |
| | 1.06 | 10 | 3 |
| | 1.08 | 10 | 5 |
| lgn_i_t3_t5 | 1.08 | 1 | 1 |
| Study population sampled from the intervention group during all PLCO calendar enrollment years (93-01) from between third and fifth study year | 1.27 | 1 | 3 |
| | 1.42 | 1 | 5 |
| | 1.16 | 3 | 1 |
| | 1.31 | 3 | 3 |
| | 1.41 | 3 | 5 |
| | 1.19 | 5 | 1 |
| | 1.33 | 5 | 3 |
| | 1.41 | 5 | 5 |
| | 1.23 | 10 | 1 |
| | 1.35 | 10 | 3 |
| | 1.40 | 10 | 5 |
| post_c_t3_t5 | 1.01 | 1 | 1 |
| Study population sampled from the usual-care group during PLCO calendar enrollment years (95-01) from between third and fifth study year | 1.02 | 1 | 3 |
| | 1.03 | 1 | 5 |
| | 1.02 | 3 | 1 |
| | 1.03 | 3 | 3 |
| | 1.04 | 3 | 5 |
| | 1.02 | 5 | 1 |
| | 1.04 | 5 | 3 |
| | 1.05 | 5 | 5 |
| | 1.03 | 10 | 1 |
| | 1.05 | 10 | 3 |
| | 1.07 | 10 | 5 |

| | | | |
|---|---|---|---|
| post_i_t3_t5 | 1.13 | 1 | 1 |
| Study population sampled from the intervention group during all PLCO calendar enrollment years (95-01) from between third and fifth study year | 1.46 | 1 | 3 |
| | 1.79 | 1 | 5 |
| | 1.25 | 3 | 1 |
| | 1.54 | 3 | 3 |
| | 1.78 | 3 | 5 |
| | 1.31 | 5 | 1 |
| | 1.59 | 5 | 3 |
| | 1.78 | 5 | 5 |
| | 1.39 | 10 | 1 |
| | 1.64 | 10 | 3 |
| | 1.78 | 10 | 5 |

**Table 7.** Simulated relative risks for smoking of four selected study designs under the double null hypothesis (screening and smoking effects are independent of lung cancer) for datasets sampled from the usual-care group (indicated under "Study Design" column with "c") and sampled from the intervention group (indicated under "Study Design" column with "i"). The relative risks were simulated using four preclinical duration distribution parameters for the mode (1,3,5,10) and three standard deviations (1,3,5). For the smoking variable, the relative risk is comparing the categories "ever smoked" to "never smoked." The simulations are based on study sample specific age distributions and screening proportion and rates among those screened and population based representations of age specific incidence.

| | Mode=1,StDev=1 | Mode=5,StDev=3 | Mode=10,StDev=5 | Combined |
|---|---|---|---|---|
| Mean *RR* | 1.06 | 1.25 | 1.33 | 1.21 |
| Median *RR* | 1.05 | 1.19 | 1.24 | 1.16 |

**Table 8.** Mean and Median simulated relative risk (*RR*) values at selected preclinical duration lognormal distribution parameterization combinations for the mode and standard deviation (StDev) of (1,1), (5,3), and (10,5) across the four selected study designs.

| Study Group Number | Dataset Name | Smoke Risk Ratio (median)* | Average Wald 95% Confidence Interval (lower, upper)# | Resampling 95% Confidence Interval (lower, upper) |
|---|---|---|---|---|
| 2 | Post_i_t3_t5 | 13.43 | (11.13,16.40) | (10.97,17.52) |
| 4 | Ign_i_t3_t5 | 13.07 | (10.80,16.06) | (8.93,20.31) |
| 6 | Post_c_t3_t5 | 12.53 | (9.99,15.23) | (9.41,16.83) |
| 8 | Ign_c_t3_t5 | 16.12 | (14.04,19.49) | (10.24,26.56) |

* $logit[\pi(x)] = log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta_1 x_1 - \beta_2 x_2 + \varepsilon_i$ where $x_1$ = smoking variable and $x_2$ = age variable

#

$Average\ Wald\ 95\%\ CI = mean(RR_{observed}) \pm 1.96 * mean(SE)$

**Table 9.** Logistic regression results for the simple model estimating the risk ever smoking has on the development of lung cancer after adjusting for age. Only the results for the smoking variable are presented because it is the variable of interest here with an adjustment for age added to the logistic model to correspond with the simulation. The PLCO data are sampled and logistic regression applied to the 100 samples of each dataset. The median risk ratio (*RR*) of the 100 samples in each of the four selected study designs is presented along with the average Wald 95% confidence interval based on mean of the 100 *RR*s and their standard errors and empirical confidence limits based on the 2.5% and 97.5% of the range of 100 *RR* values.

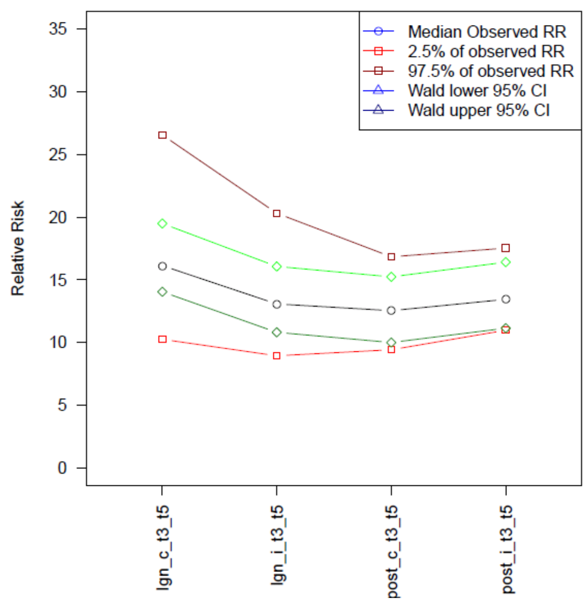



**Figure 18.** Graph illustrates the Logistic regression results for the four selected study designs from Table 7 for the simple model estimating the risk ever smoking has on the development of lung cancer after adjusting for age. The median risk ratio (*RR*) of the 100 samples in each dataset is presented along with the average Wald 95% confidence interval based on mean of the 100 *RR*s and their standard errors and empirical confidence limits based on the 2.5% and 97.5% of the range of 100 *RR* values for each study design.

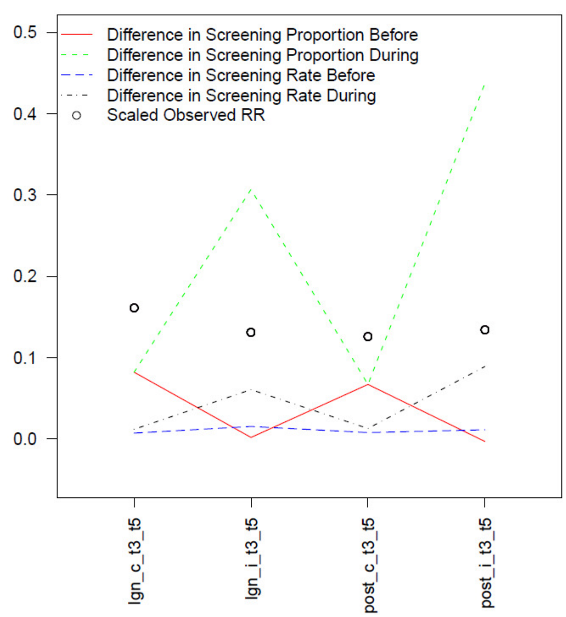



**Figure 19.** Representation of the relationship of screening behavior differential between smokers and never smokers and scaled observed *RR* (divided by 100) by four study design. The lines in the plot represent the differences in screening behavior (i.e., proportion screened or screening rate) both before and during each nested case-control study design. The points represent scaled observed RR estimated with logistic regression model (divided by 100).

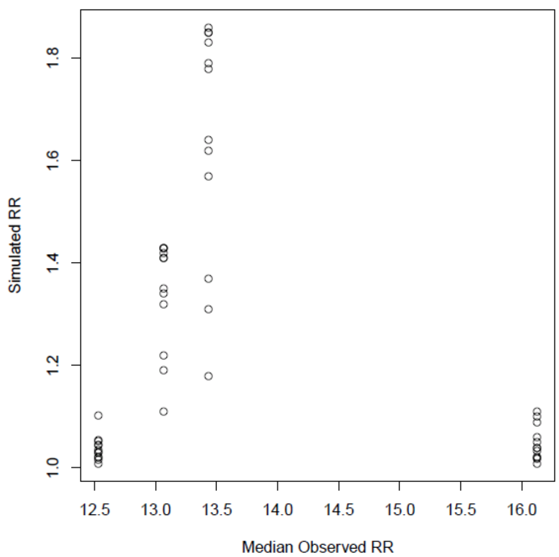
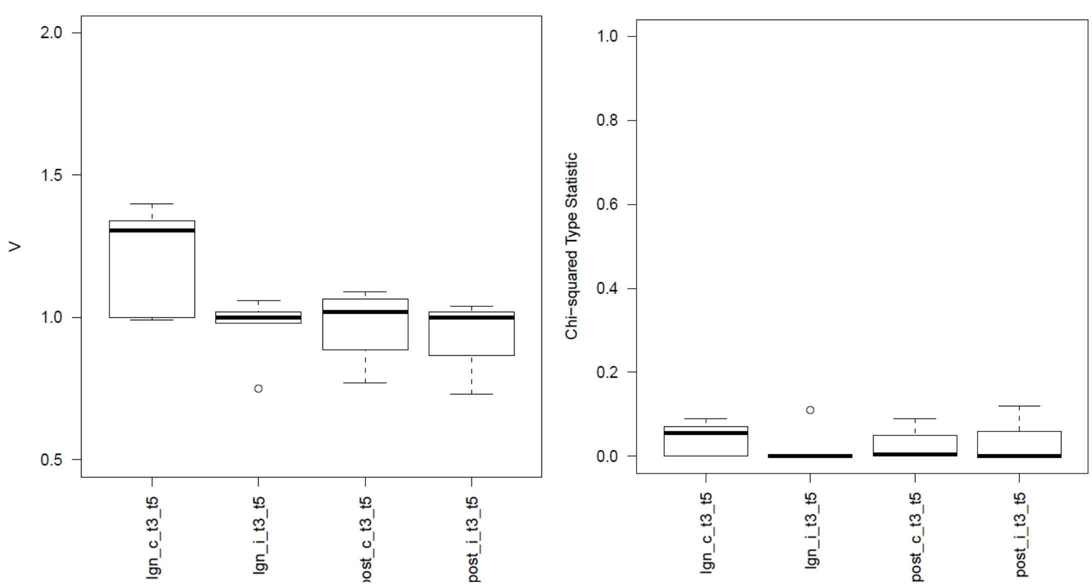



**Figure 20.** Graph of the observed *RR* versus the simulated *RR* for each of the four study designs created to show any correlation between the two *RR* values. The position on the x-axis represents the observed *RR* estimated using a logistic regression model, one calculated for each of the four study designs where vertical range represents the 12 different simulated *RR*s (obtained through combination of mode (1,3,5,10) and standard deviation (1,3,5) year model parameterizations for the preclinical duration distribution) for that study design.



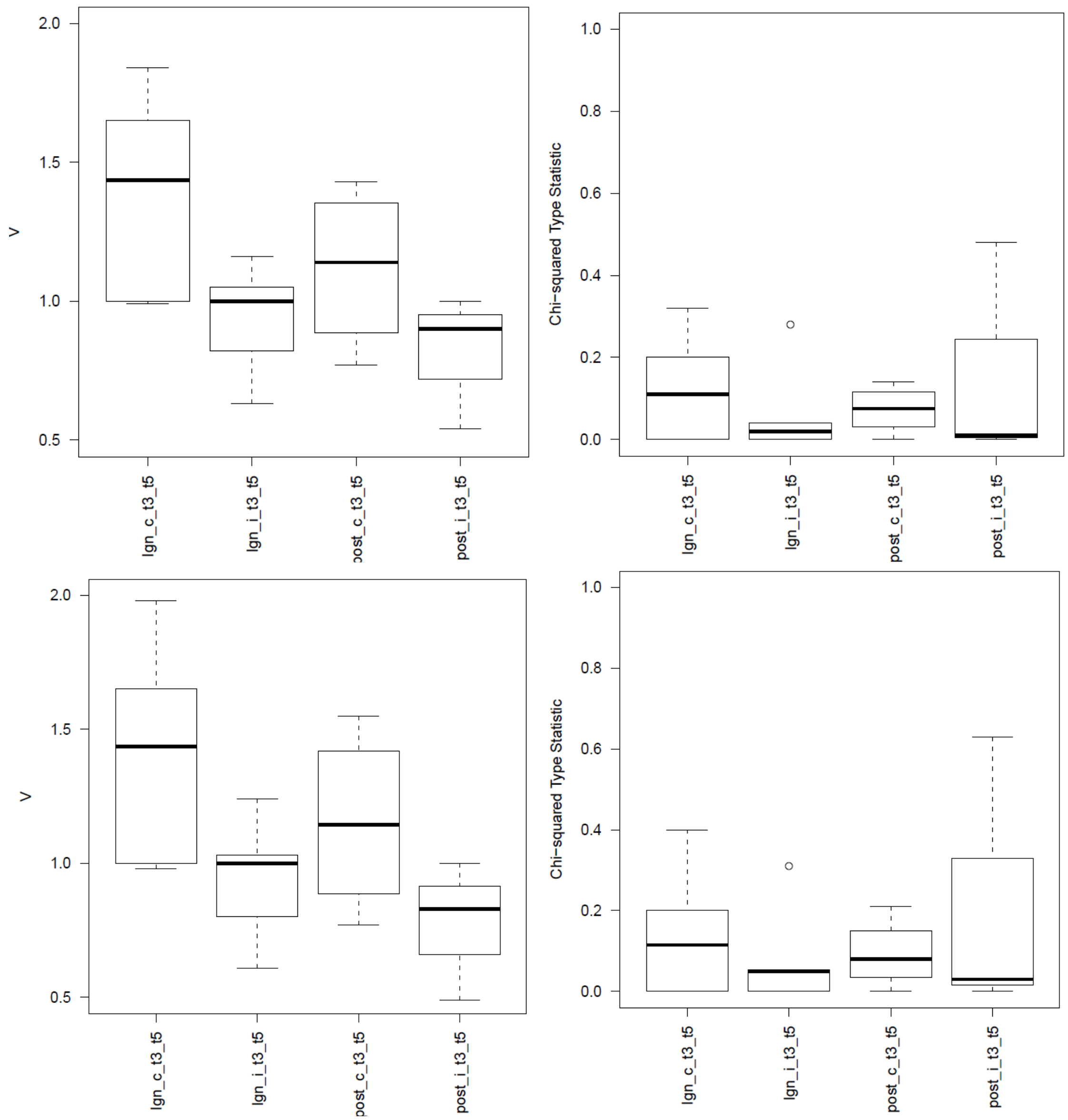



**Figure 21.** Representation of the range of $V$ (i.e., ratio of the ratio of two $RR_{\text{simulated}}$ to the ratio of two $RR_{\text{observed}}$) for each of the four study design combinations (left side). Each combination is obtained by comparing two study designs (e.g., ign_c_t3_t5 to ign_i_t3_t5, etc.) for a total of 6 pairs. These pairs were then evaluated under 3 different model parameterizations (1) mode=1, standard deviation = 1; 2) mode= 5, standard deviation = 3; 3) mode= 10, standard deviation = 5) for a total of 18 $V$s which are represented using a boxplot. Figure 20 (right side) uses same technique with Chi-squared type value (($RR_{\text{observed}}$ ratio-$RR_{\text{simulated}}$ ratio)$^2$/$RR_{\text{simulated}}$ ratio) for each of the study design combinations.

| Mode | Standard Deviation | Sum of Chi-sq type Statistic |
|---|---|---|
| 1 | 1 | 0.29 |
| 1 | 3 | 0.78 |
| 1 | 5 | 1.43 |
| 3 | 1 | 0.43 |
| 3 | 3 | 0.92 |
| 3 | 5 | 1.34 |
| 5 | 1 | 0.52 |
| 5 | 3 | 0.98 |
| 5 | 5 | 1.29 |
| 10 | 1 | 0.63 |
| 10 | 3 | 1.03 |
| 10 | 5 | 1.23 |

**Table 10.** Sum of the Chi-squared type statistic by preclinical duration distribution parameterization (mode years = 1,3,5,10 and standard deviation years = 1,3,5) calculated for each of the 6 combinations for the 4 selected study designs types. The sum is used to see how well the ratio of the simulated amount of screening bias between two study designs correspond with the ratio of observed RR between two study designs. This statistic provides a way to find the best parameterization from those we used with a value of 0 for the statistic indicating perfect prediction.

## Assumptions and Limitations

The assumptions of the mathematical model under the null hypothesis include:

1) Chest x-ray screening is ineffective and has no association with other causes of death or preclinical incidence
2) Preclinical incidence time and preclinical duration are independent
3) Preclinical incidence and preclinical durations distributions are accurate representations of "true" population distributions
4) Chest x-ray screening and smoking are not associated with lung cancer

To date there has not been any definitive evidence to indicate that chest x-ray is an effective screening tool in the detection of lung cancer, or is associated with other causes of death. A few studies have indicated that extensive use of chest x-ray maybe associated with increased risk of lung cancer, but it is unknown if the radiation spawns the lung cancer or just accelerates the development of pre-existing cancerous cells. The PLCO randomized trial has been designed to address the effectiveness of chest x-ray as a screening tool for lung cancer and the results can be incorporated to verify or suggest revision of this assumption.

In the progressive disease model we described from Zelen and Feinleib(1) , we cannot observe the preclinical incidence time or the preclinical duration. Within the study time limits (in the PLCO a span of 23 years), there would only be an association between the preclinical incidence time and preclinical duration if the disease or classification of lung cancer changed over this time period (thereby advancing the date of symptomatic diagnosis and creating a shorter preclinical duration). This occurrence is unlikely given the study followed a similar diagnostic protocol throughout.

The target population of the PLCO study was the U.S. population, so the SEER database which also represents the U.S. population was used to get 5-year age specific incidence rates. Previously published data on a range of sensitivity estimates for the U.S. population was also incorporated. It is plausible to assume that these distributions are representative of the PLCO sample because of the same target populations and the large number of individuals (i.e., 1000) in each study or database.

One of the main interests is investigating the screening effect on the smoking-lung cancer relationship within the control group and the intervention group after final scheduled screen. To accomplish this, the screening behavior information within these groups during the first 5 years of the study needs to be known. However the only information is screening in the 3 years prior to the beginning of the study for the usual-care group and additionally 2-3 annual screens for the interventional group. So the assumption that the screening patterns for these 3 years prior to enrollment are representative of any unobserved screening behavior during the study was used. There is additional partial screening data on the control group during the study, and although very limited, could be incorporated to adjust this assumption. This additional data would provide stronger support for this assumption that the screening behavior from before the study is representative of the entire group during the study.

The goal of a risk-factor study is to identify the causal relationship between the risk factor and disease. Using the counterfactual framework, the same population would need to be observed during the same time period with the same life histories only allowing the exposure of interest to change from the first to second observation. This is impossible, thus a substitute must be used for the unobservable counterfactual situation. In each of the case-control designs the interest is in estimating the effect smoking has on lung cancer after removing the screening effect. The Mathematical model attempts to accomplish this by simulating the same population under screening and no screening conditions. If as indicated theoretically, the simulation results are used to correct the RR, it would only be a causal measure for the effect of smoking on lung cancer if it is assume that there are no other forms of bias in the study and all assumptions are accurate for our data (including our adjustment for screening bias). It is unlikely that our nested case-control studies are free of all other forms of bias and thus the simulation corrected RR would only be unbiased for forms of screening bias identified here and not a causal parameter for the effect of smoking on lung cancer. In order to obtain that causal parameter, the simulation corrected RR would need to be combine with corrections for all the other forms of bias (such as misclassification and sampling) as demonstrated in Maldonado and Greenland(44).

## Summary

The focus of this research was to look at the neglected issue of how screening can bias risk-factor studies specifically identifying how chest x-ray screening changes the observed association between smoking and lung cancer (unintended screening effect). Evaluation of the possible screening effect on the smoking-lung cancer association was conducted by modifying a previously developed mathematical model. The modified model compared the outcome of lung cancer diagnosis (e.g., incidence) evaluated under the specific screening situations of several case-control studies nested in the PLCO randomized trial between smokers and never smokers to yield the potential amount of screening bias affecting the risk estimate. In theory applying the bias as a correction to the observed risk estimate would provide a more accurate representation of the effect smoking has on lung cancer diagnosis in a representative sample of the United States around the 21st century.

Although the smoking-lung cancer association in the presence of chest x-ray screening is the main example throughout this proposal, the concepts can be applied more generally. Chest x-ray screening could be replaced by any other form of early detection and lung cancer could be replaced with any other disease for which a progressive disease model and the form of early detection are plausible. A goal of this project is to inform researchers about the potential ways screening can affect risk-factor studies and to provide a method to address such bias.

The three main hypotheses for this project were: 1) Can the use of screening significantly bias risk-factor estimates in observational studies?; 2) How do various case-control design choices influence the amount the risk estimate is biased by screening use?; 3) Do the simulated model results and empirical observations correspond and under what conditions?

The first hypothesis was tested using a modified recurrence-time model for screening bias due to lead-time, length, and overdiagnosis bias. This model was used to simulate the effects of realistic cancer screening behaviors in case-control studies of lung cancer incidence nested within the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. When a screening test is used among subjects in an observational study, the screen-detected cases will have an advanced date of diagnosis and likely slower progressing disease compared to non-screen-detected cases resulting in screening bias. If differential screening behavior exists between risk-factor strata, case-ascertainment maybe changed differentially, thereby misrepresenting the observed measure of association between the risk factor and disease. In the presence of differential screening under plausible assumptions about preclinical incidence and duration, the simulations presented showed the possibility for screening bias from chest x-ray to affect the risk smoking has on the development of lung cancer by up to 85%.

Within these results, a relationship emerged that as screening differential (either in proportion or rate) between strata of the variable (e.g., ever smoked vs. never smoked) increases, so does the susceptibility to this screening bias. In general when the preclinical duration (i.e., mode and standard deviation) increases, so does the amount of bias expected to affect the observed RR. Also, the model appears to be relatively sensitive to standard deviation and mode variations causing simulation values to differ by about 30%

in some situation when comparing smallest to largest pairs of these parameters. There is an indication that incorporating overdiagnosis can have a significant effect on the RR, most of all when the preclinical duration is shortest (e.g., mode=1, standard deviation = 1). The screening rate was likely high enough to negate slight variations in the screening test sensitivity.

To answer question 2, twenty-seven (24 unique) case-control study designs nested within the PLCO randomized trial have been developed as a method to determine a possible study design effect on screening bias within the smoking-lung cancer observed risk estimate. Out of these studies, a relationship between screening behavior has surfaced that as screening proportion and rate before the study increases the expected amount of screening bias (illustrated with simulated RR) in the study decreases. The opposite is suggested when looking at the screening behavior during the study such that as screening proportion or rate increases, so does the simulated RR. It was observed that as the preclinical duration of lung cancer increases the expected amount of screening bias in the observed RR increases. This result was anticipated both because it was suggested in previous work and as the time spent in the preclinical stage increases so does the chance of being screen detected. As the preclinical duration increases it becomes less likely that the time of screen detection and the theoretical time of symptomatic or clinical detection occur during the same interval (i.e., observed screen detection time and counterfactual symptomatic detection time both happen before the beginning of enrollment). Case-control studies designed to sample only from the usual-care group were indicated to have less expected screening bias than studies using the interventional group. As the case enrollment period increase from 2 to 3 to 6 years, the simulated RR decreases. This trend is supported by the result of increase screening bias and increased preclinical duration. Here the lengthened case-ascertainment period increases the probability of the two detection times occurring during the same interval thereby reducing the expected screening bias.

Comparing the datasets that ignore the procedural modification in the number of offered screens to nonsmokers (dataset names that begin with “Ign”) to those that sample only after the modification (dataset names that begin with “Post”), the results suggested that the studies using datasets that ignore the screening modification are expected to be effected by more screening bias than those sampled only after the modification.

There is no clear trend suggested by the plot of observed RR or simulated RR values to infer a relationship for the location of the case-ascertainment period within the PLCO trial (i.e., T1-T3, T2-T3, etc.) with only a small indication within the interventional group that as the nested case-control studies enrollment period (if enrollment period is of same length) is shifted later in the study (i.e., form T0-T2 to T1-T3 to T2-T4 to T3-T5) the expected amount of bias actually increases.

To validate the simulation model can be used as a correction method (hypothesis 3), the logistic regression results were compared with the results from simulated the mathematical model for the 27 nested case-control studies. The model was considered to give a valid value when the ratio (i.e., V) of the simulated risk ratios ($RR_{simulated}$) for two study designs was equal to the ratio of the observed risk ratios ($RR_{observed}$) estimated with a logistic model for the same two study designs. The V values were highly variable across study design and after breaking down the results based on preclinical duration in order to identify the best set of model parameters, it was suggested that using a shorter preclinical duration in case-control studies of the risk smoking has on the development of lung cancer results in values closer to that which is expected. The shorter preclinical duration is supported by the distribution of chi-squared type statistics with many more values close to 0 which indicates that the model is predicting what is expected.

To give utility to the validation technique used here of calculating V values, it was assumed that any change in the observed RR from one study to the next is due to screening (if there is additional bias it is equal or negligible across study designs) therefore the ratio of the $RR_{observed}$ for the two studies should be equal to the ratio of the $RR_{simulated}$. However, this assumption doesn't appear to hold based on the plot illustrating that changes in the $RR_{observed}$ are not completely driven by screening behavior (proportion screened and rate of screening) differential between smokers and never smokers. Because screening behavior differential is essential for screening bias to occur in observational studies and is the basis of the mathematical model developed here, the PLCO data may not provide the needed situation to validate that the model can predict the amount of screening bias affecting the $RR_{observed}$. It is recommended that the mathematical model be fully validated by developing a distribution for the chi-squared type statistic and using additional techniques and additional data before using the simulation results as a method to provide an unbiased observed *RR* with respect to screening bias.

It appears that differential screening behaviors demonstrated between smokers and never smokers in these nested case-control studies of lung cancer diagnosis do influence observed RRs. However, there are likely other types of bias (besides screening bias) also influencing these observed RRs differentially making validation of the mathematical model using the described empirical comparison technique difficult. The results indicate that using a shorter preclinical duration in the simulation may provide more accurate screening bias prediction here, but there are some indications (i.e., V not close enough to one or chi-squared type statistic not close enough to 0) to the contrary. It is recommended that the mathematical model be validated using additional techniques and additional data before using the simulation as a method to "correct" the observed RR for screening bias. Even if the mathematical model had been validated, however, screening use in a population causes the observed cases during the ascertainment period to be a different subgroup than the cases expected in the absence of screening. Thus, even if screening is adjusted for with the simulated value, study results should be generalized with caution.

## References


1. Zelen M, Feinleib M. On the theory of screening for chronic diseases. Biometrika 1969;56:601-614.
2. Albert A, Gertman PM, Louis TA, Liu S-I. Screening for the Early Detection of Cancer - II. The Impact of Screening on the Natural History of Disease. Math Biosci 1978;40:61-109.
3. Albert A, Gertman PM, Louis TA. Screening for the Early Detection of Cancer - 1. The Temporal Natural History of a Progressive Disease State. Math Biosci 1978;40:1-59.
4. Flehinger BJ, Kimmel M. The natural history of lung cancer in a periodically screened population. Biometrics 1987;43:127-144.
5. Brookmeyer R, Day NE. Two-stage models for the analysis of cancer screening data. Biometrics 1987;43:657-69.
6. Yamaguchi N, Tamura Y, Sobue T, et al. Evaluation of cancer prevention strategies by computerized simulation model: an approach to lung cancer. Cancer Causes Control 1991;2:147-55.
7. Day NE. The assessment of lead time and length bias in the evaluation of screening programmes. Maturitas 1985;7:51-58.
8. Gyrd-Hansen D, Sogaard J, Kronborg O. Analysis of screening data: Colorectal cancer. International Journal of Epidemiology 1997;26:1172-1181.
9. Walter SD, Day NE. Estimation of the duration of a pre-clinical disease state using screening data. American Journal of Epidemiology 1983;118:865-886.
10. Prevost TC, Launoy G, Duffy SW, Chen HH. Estimating sensitivity and sojourn time in screening for colorectal cancer: A comparison of statistical approaches. American Journal of Epidemiology 1998;148:609-619.
11. Shen Y, Parmigiani G. A model-based comparison of breast cancer screening strategies: mammograms and clinical breast examinations. Cancer Epidemiol Biomarkers Prev 2005;14:529-32.
12. Day NE, Walter SD. Simplified Models of screening fro chronic disease: Estimation procedures from mass screening programmes. Biometrics 1984;40:1-14.
13. Paci E, Duffy SW. Modelling the analysis of breast cancer screening programmes: Sensitivity, lead time and predictive value in the Florence District Programme (1975-1986). International Journal of Epidemiology 1991;20:852-858.
14. Shen Y, Huang X. Nonparametric estimation of asymptomatic duration from a randomized prospective cancer screening trial. Biometrics 2005;61:992-9.
15. Pinsky PF. Estimation and prediction for cancer screening models using deconvolution and smoothing. Biometrics 2001;57:389-95.
16. Cong XJ, Shen Y, Miller AB. Estimation of age-specific sensitivity and sojourn time in breast cancer screening studies. Stat Med 2005;24:3123-38.

17. Oortmarssen GJV, Habbema JDF, Lubbe JTN, Maas PJVD. A model-based analysis of the hip project for breast cancer screening. Int J Cancer 1990;46:207-213.
18. Shen Y, Zelen M. Robust modeling in screening studies: estimation of sensitivity and preclinical sojourn time distribution. Biostatistics 2005;6:604-14.
19. Baker SG, Chu KC. Evaluating screening for the early detetion and treatment of cancer without using a randomized control group. Journal of the American Statistical Association 1990;85:321-327.
20. Baker SG, Erwin D, Kramer BS, Prorok PC. Using observational data to estimate an upper bound on the reduction in cancer mortality due to periodic screening. BMC Medical Research Methodology 2003;3.
21. Straatman H, Peer PGM, Verbeek ALM. Estimating lead time and sensitivity in a screening program without estimating the incidence in the screened group. Biometrics 1997;53:217-229.
22. Cong XJ, Shen Y, Miller AB. Estimation of age-specific sensitivity and sojourn tim ein breast cacner screening studies. Statistics in Medicine 2005;24:3123-3138.
23. Wu D, Rosner GL, Broemeling L. MLE and bayeian inference of age-dependent sensitivity and transition probability in periodic screening. Biometrics 2005;61:1056-1063.
24. Prorok PC. The theory of periodic screening I: Lead time and proportion detected. Advances in applied probability 1976;8:127-143.
25. Prorok PC. The theory of periodic screening II: Doublly bounded recurrence times and mean lead time and detection probability estimation. Advances in applied probability 1976;8:460-476.
26. Draisma G, Boer R, Otto SJ, et al. Lead time and overdetection due to prostate-specific antigen screening: Estimates from the European Randomized Study of Screening for Prostate Cancer. Journal of the National Cancer Institute 2003;95:868-878.
27. Xu J-l, Prorok PC. Non-parametric estimation of the post-lead-time survival distribution of screen-detected cancer cases. Statistics in Medicine 1995;14:2715-2725.
28. Morrison AS. The effects of early treatment, lead time and length bias on the mortality experienced by cases detected by screening. International Journal of Epidemiology 1982;11:261-267.
29. Chen JS, Prorok PC. Lead time estimation in a controlled screening program. American Journal of Epidemiology 1983;118:740-751.
30. Xu JL, Fagerstrom RM, Prorok PC. Estimation of post-lead-time survival under dependence between lead-time and post-lead-time survival. Stat Med 1999;18:155-62.
31. Church TR. A novel form of ascertainment bias in case-control studies of cancer screening. J Clin Epidemiol 1999;52:837-47.
32. Moss SM. Case-control studies of screening. Int J Epidemiol 1991;20:1-6.
33. Prorok PC, Connor RJ, Baker SG. Statistical consideration in cancer screening programs. Urologic Clinics of North America 1990;17:699-708.

34. Sasco A. Lead time and length bias in case-control studies for the evaluation of screening. J Clin Epidemiol 1988;41:103-104.
35. Weiss NS. Adjusting for screening history in epidemiologic studies of cancer: why, when, and how to do it. Am J Epidemiol 2003;157:957-61.
36. Joffe MM. Invited Commentary: Screening as a Nuisance Variable in Cancer Epidemiology: Methodological Considerations. American Journal of Epidemiology 2003;157:962-964.
37. Mathsoft® Engineering and Education I. Mathcad® 12. Cambridge: Mathsoft® Engineering and Education, Inc., 2004.
38. Quaglia A, Vercelli M, Puppo A, et al. Prostate cancer in Italy before and during the 'PSA era': survival trend and prognostic determinants. Eur J Cancer Prev 2003;12:145-52.
39. Auvinen A, Maattanen L, Finne P, et al. Test sensitivity of prostate-specific antigen in the Finnish randomised prostate cancer screening trial. Int J Cancer 2004;111:940-3.
40. Etzioni R, Cha R, Feuer EJ, Davidov O. Asymptomatic incidence and duration of prostate cancer. Am J Epidemiol 1998;148:775-85.
41. Gohagan JK, Prorok PC, Hayes RB, Kramer BS. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. Control Clin Trials 2000;21:251S-272S.
42. Prorok PC, Andriole GL, Bresalier RS, et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. Control Clin Trials 2000;21:273S-309S.
43. Pearl J. Causation, action, and counterfactuals. Theoretical Aspects Of Rationality And Knowledge: Proceedings of the 6th conference on Theoretical aspects of rationality and knowledge. The Netherlands, 1996:51-73.
44. Maldonado G, Greenland S. Estimating Causal Effects. International Journal of Epidemiology 2002;31:422-429.
45. O'Brien B, Nichaman L, Browne JE, Levin DL, Prorok PC, Gohagan JK. Coordination and management of a large multicenter screening trial: the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. Control Clin Trials 2000;21:310S-328S.
46. SAS software. Cary, NC: SAS Institute Inc., 2000-2004.
47. Dubin N. Benefits of screening for breast cancer: Application of a probabilistic model to a breast cancer detection project. Chronic Disease 1978;32:145-151.
48. Freedman M. State-of-the-art screening for lung cancer (part 1): the chest radiograph. Thorac Surg Clin 2004;14:43-52.
49. Doi K, MacMahon H, Katsuragawa S, Nishikawa RM, Jiang Y. Computer-aided diagnosis in radiology: potential and pitfalls. European Journal of Radiology 1997;31:97-109.
50. Rothman K, Greenland S. Case-control studies. In: Greenland S, ed. Modern Epidemiology. Philadelphia: Lippincott-Raven, 1998:93-114.
51. Suissa S, Edwardes MDd, Boivin J-F. External comparisons from nested case-control designs. Epidemiology 1998;9:72-78.

52. Katz M. Designing a study. In: Katz M, ed. Study design and statistical analysis. New York: Cambridge University Press, 2006:25-31.
53. Friis R, Sellers T. Study designs: ecologic, cross-sectional, case-control. In: Sellers T, ed. Epidemiology for public health practice. Massachusetts: Jones and Bartlett Publishers, 2004:234-348.
54. Robins JM, Gail MH, Lubin JH. More on "biased selection of controls for case-control analyses of cohort studies". Biometrics 1986;42:293-299.
55. Woodward M. Case-control studies. In: Woodward M, ed. Epidemiology: study design and data analysis. Boca Raton: Chapman and Hall/CRC, 1999:234-285.
56. Schlesselmann J. Case-control studies: design, conduct, analysis. New York: Oxford University Press, 1982.
57. Connor RJ, Prorok PC, Weed DL. The case-control design and the assessment of the efficacy of cancer screening. Journal of Clinical Epidemiology 1991;44:1215-1221.
58. Friedmann D, Dubin N. Case-control evaluation of breast cancer screening efficacy. American Journal of Epidemiology 1991;133:974-984.
59. Gullberg B, Andersson I, Janzon L, Ranstam J. Screening Mammography. Lancet 1991;337:244.
60. Shwartz M. Estimates of lead time and length bias in a breast cacner screening program. Cancer 1980;46:844-851.
61. Oehlert GW. A Note on the Delta Method. The American Statistician 1992;46:27-29.

# Appendix

## *Chapter 5 figures and tables for all datasets*

| | Mode=1,StDev=1 | Mode=5,StDev=3 | Mode=10,StDev=5 | Combined |
|---|---|---|---|---|
| Mean RR | 1.05 | 1.14 | 1.17 | 1.13 |
| Median RR | 1.01 | 1.04 | 1.08 | 1.04 |

**Table a-8.** Mean and Median simulated relative risk (RR) values at selected preclinical duration lognormal distribution parameterization combinations for the mode and standard deviation (StDev) of (1,1), (5,3), and (10,5).

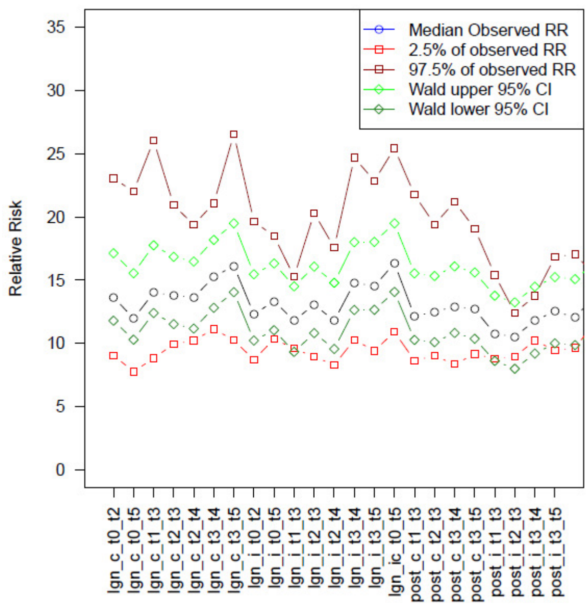



**Figure a-18. G**raph illustrates the Logistic regression results from Table 7 for the simple model estimating the risk ever smoking has on the development of lung cancer after adjusting for age. The median risk ratio (*RR*) of the 100 samples in each dataset is presented along with the average Wald 95% confidence interval based on mean of the 100 *RR*s and their standard errors and empirical confidence limits based on the 2.5% and 97.5% of the range of 100 *RR* values for each study design.

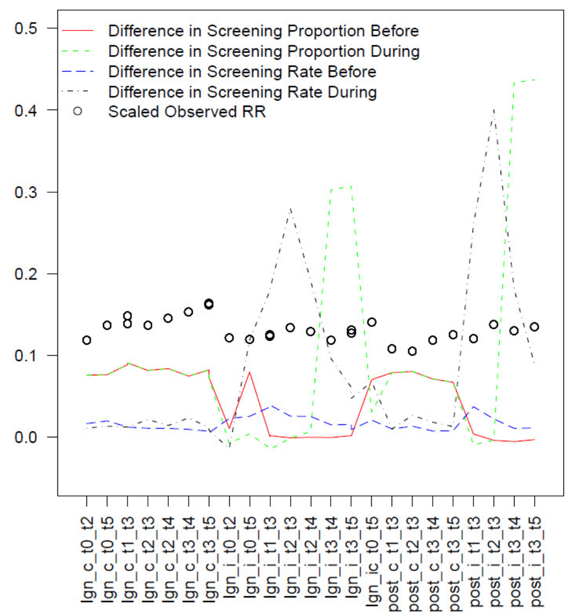



**Figure a-19.** Representation of the relationship of screening behavior differential between smokers and never smokers and scaled observed *RR* (divided by 100) by each of the 27 study designs. The lines in the plot represent the differences in screening behavior (i.e., proportion screened or screening rate) both before and during each nested case-control study design. The points represent scaled observed RR estimated with logistic regression model (divided by 100).

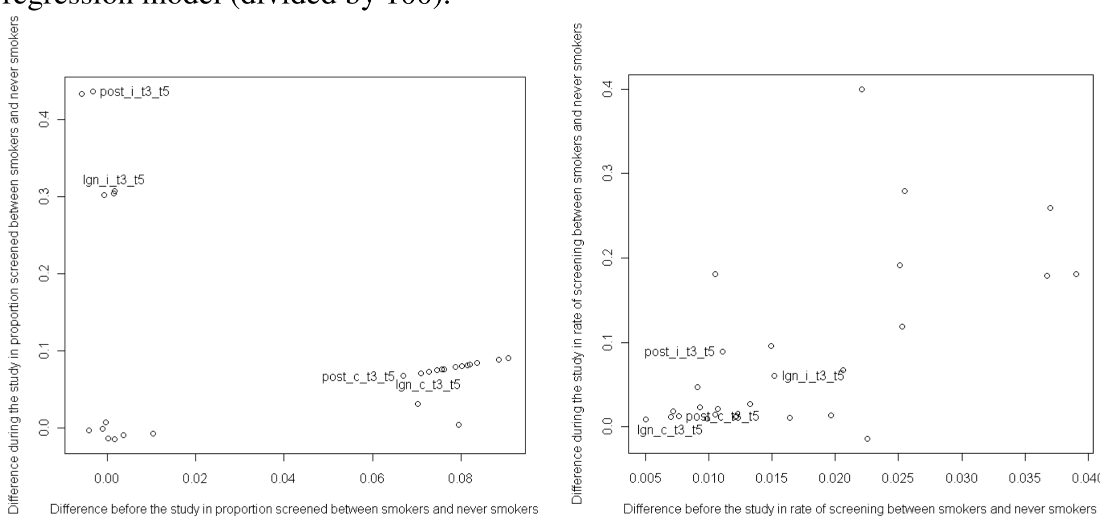



**Figure a-20.** Representation of the difference in screening proportion and rate between smokers and never smokers comparing the difference before to the difference during for each of the 27 study designs. The plot on the right is the screening proportion difference before the study versus during the study and the plot on the left is screening rate difference before the study versus during the study. The four selected study designs from chapter 5 are identified within each plot providing an illustration that there is one point per study design type.

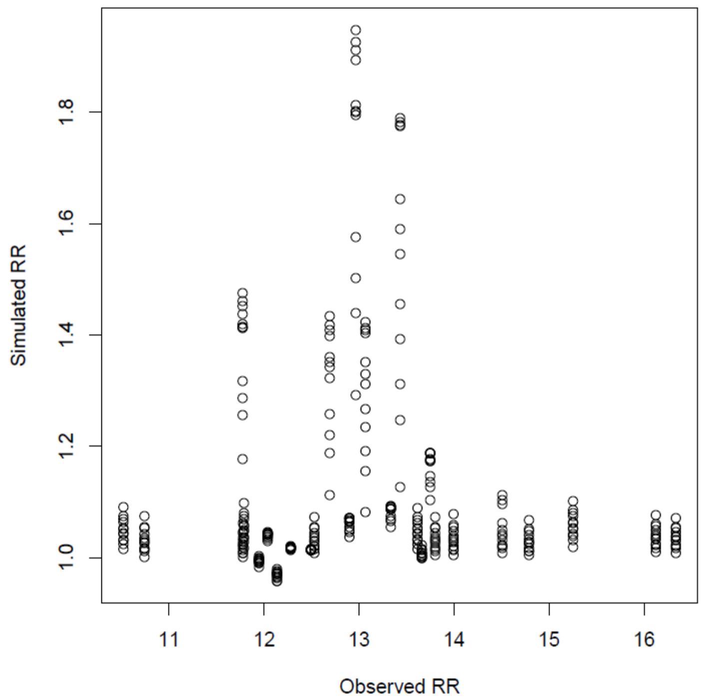
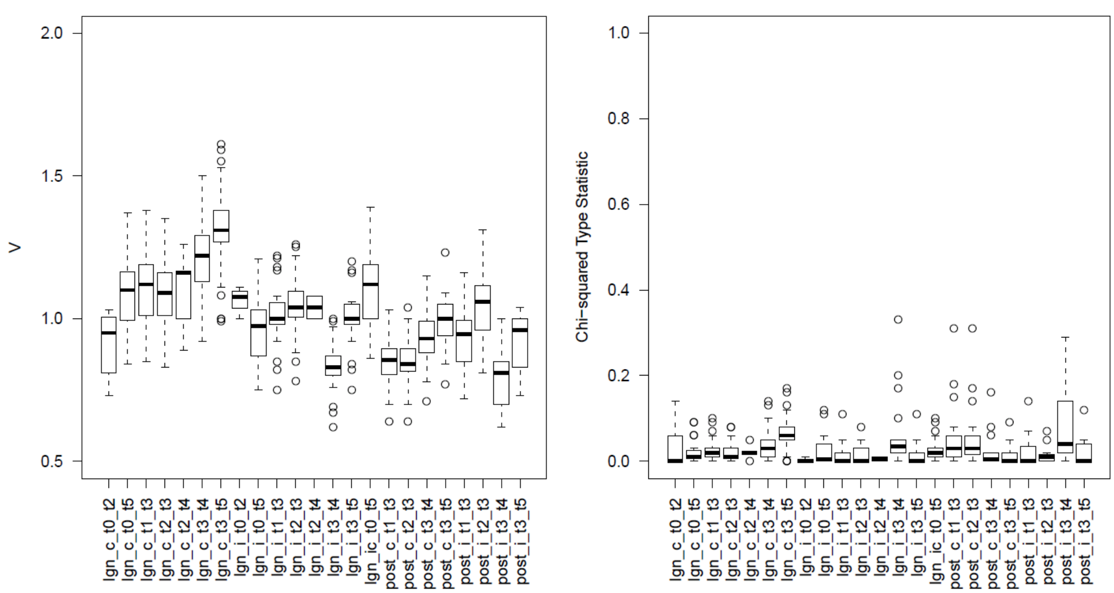



**Figure a-21.** Graph of the observed RR verse the simulated RR for each of the 27 study designs created to show any correlation between the two RR values. The position on the x-axis represents the observed RR estimated using a logistic regression model, one calculated for each of the 27 study designs where vertical range represents the 12 different simulated RRs (obtained through combination of mode (1,3,5,10) and standard deviation (1,3,5) year model parameterizations for the preclinical duration distribution) for that study design.



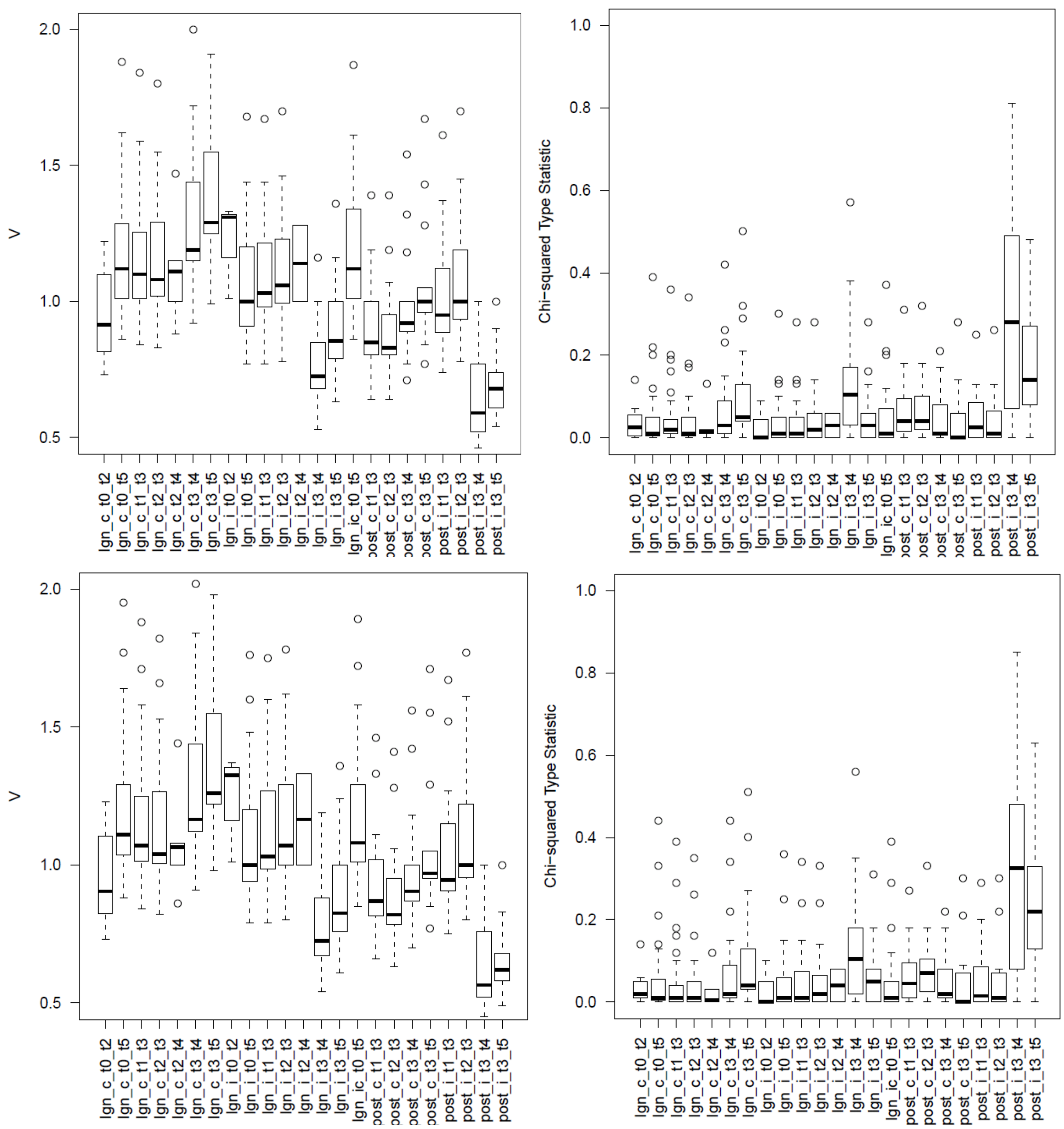



**Figure a-22.** Representation of the range of V (i.e., ratio of the ratio of two $RR_{simulated}$ to the ratio of two $RR_{observed}$) for each of the study design combinations (left side). Each combination is obtained by comparing two study designs (e.g., ign_c_t0_t2 to ign_c_t0_t5, etc.) for a total of 378 pairs. These pairs were then evaluated under 3 different model parameterizations (1) mode=1, standard deviation = 1; 2) mode= 5, standard deviation = 3; 3) mode= 10, standard deviation = 5) for a total of 1134 *V*s which are represented using a boxplot. Figure 20 (right side) uses same technique with Chi-squared type value (($RR_{observed}$ ratio-$RR_{simulated}$ ratio)$^2$/$RR_{simulated}$ ratio) for each of the study design combinations.

## *Sample Mathcad worksheet for lgn_c_t0_t5, the usual-care group using all cases from T0 to T5 regardless of randomization date*

**Number of Age categories (for 55-74)**

$i := 0, 1 .. 3$

**middle of each 5-year age categories**

$t_o := 60.5, 65.5 .. 75.5$

**Sensitivity of screening test**

$\xi := .86$

**Maximum age**

$max_{age} := 120$

**Age suggested to begin Screening**

$screen_{age} := 50$

**Screening pattern values**

$g :=$

| | 0 |
|---|---|
| 0 | $-1.232\cdot10^{-4}$ |
| 1 | 0.028 |
| 2 | -0.657 |
| 3 | $3.81\cdot10^{-5}$ |
| 4 | $-8.305\cdot10^{-4}$ |
| 5 | 0.598 |
| 6 | $-1.232\cdot10^{-4}$ |
| 7 | 0.028 |
| 8 | -0.657 |
| 9 | $2.54\cdot10^{-5}$ |
| 10 | $-5.537\cdot10^{-4}$ |
| 11 | 0.399 |
| 12 | $1.429\cdot10^{-4}$ |
| 13 | $-4.961\cdot10^{-3}$ |
| 14 | 0.25 |
| 15 | $4.25\cdot10^{-6}$ |

### Determine the parameters for the preclinical duration distribution f(x)

$mu := 2 \quad sd := .1$ **Starting values**

Given

$$\sqrt{\left(e^{sd^2}-1\right)\cdot e^{2\cdot mu+sd^2}} = 10$$ **Standard deviation**

$$e^{mu-sd^2} = 5$$ **Mode**

$x := \mathrm{Find}(mu, sd)$

$x =$

$x_0 =$

$mu := x_0$

$x_1 =$

$sd := x_1$

Mode $\quad e^{mu-sd^2} =$

$$Mean := e^{mu+\frac{sd}{2}}$$

$$Median := e^{mu}$$

StdDev $\quad \sqrt{\left(e^{sd^2}-1\right)\cdot e^{2\cdot mu+sd^2}} =$

**This vector represents the middle number for five year age categories at T3**

$$t_0 := \begin{pmatrix} 60.5 \\ 65.5 \\ 70.5 \\ 75.5 \end{pmatrix}$$

**PLCO study weights for age groups**

$$\omega := \begin{pmatrix} 0.330 \\ 0.307 \\ 0.227 \\ 0.136 \end{pmatrix}$$

**This value is the end of study period**

$d := t_0 + 6$

**This is a log-normal formula for the preclinical duration distribution - how long do we expect a person to have prostate cancer before being diagnosed clinically**

$$f(x) := \frac{1}{\sqrt{2\cdot\pi}\cdot sd\cdot x}\cdot \exp\left[-\frac{\left(\frac{\log(x)}{\log(e)}-mu\right)^2}{2\cdot sd^2}\right]$$

### Define the parameters of incidence intensity function (determined using nonlinear minimization function in R).

$a\_b := 13.24607 \quad c\_b := 75.56094 + Mean$

$b\_b := 3.51654 \quad d\_b := -2.00611$

$$w\_b(x) := \left(\frac{1}{a\_b\cdot\sqrt{2\cdot b\_b}}\right)\cdot \exp\left[-\frac{1}{2\cdot\left(a\_b^2\right)}\cdot\left[\left[(x-c\_b)^2\right]+d\_b\right]\right]$$

$a\_d := 12.549125 \quad c\_d := 77.475978 + Mean$

$b\_d := 2.755585 \quad d\_d := -2.092053$

$$w\_d(x) := \left(\frac{1}{a\_d\cdot\sqrt{2\cdot b\_d}}\right)\cdot \exp\left[-\frac{1}{2\cdot\left(a\_d^2\right)}\cdot\left[\left[(x-c\_d)^2\right]+d\_d\right]\right]$$

### Define the parameters of the screening intensity function.

**Before start of study period (T0)**

**Chest X-Ray Proportion screened in each age category** | **Screening rate among those screened**

**smoking screening rate** $k\_b_3(x) := \mathrm{if}\left(x < screen_{age}, 1, g_0 x^2 + g_1 x + g_2\right)$ $\quad k\_b_1(x) := g_3 x^2 + g_4 x + g_5$

**nonsmoking screening rate** $k\_b_4(x) := \mathrm{if}\left(x < screen_{age}, 1, g_{12} x^2 + g_{13} x + g_{14}\right)$ $\quad k\_b_2(x) := g_{15} x^2 + g_{16} x + g_{17}$

**During study period (T0-T5)**

**Chest X-Ray Proportion screened in each age category** | **Screening rate among those screened**

**smoking screening rate** $k\_d_3(x) := \mathrm{if}\left(x < screen_{age}, 1, g_6 x^2 + g_7 x + g_8\right)$ $\quad k\_d_1(x) := g_9 x^2 + g_{10} x + g_{11}$

**nonsmoking screening rate** $k\_d_4(x) := \mathrm{if}\left(x < screen_{age}, 1, g_{18} x^2 + g_{19} x + g_{20}\right)$ $\quad k\_d_2(x) := g_{21} x^2 + g_{22} x + g_{23}$

---

### Estimate the cumulative incidence without screening and with screening under the null hypothesis

***Incidence without screening***

**Incidence (during the study period) from cancers initiated before study start**

$$G_{u1} := \sum_i \left[\omega_i\left[\int_0^{t_{0_i}}\int_{t_{0_i}}^{d_i} (w\_b(x))\cdot f(z-x)\,dz\,dx\right]\right]$$

**z is time of clinical surfacing and**
**x is the beginning of the preclincial duration**

**Incidence (during the study period) from cancers initiated after study start but before the end of the study period**

$$G_{u2} := \sum_i \left[\omega_i\cdot\left(\int_{t_{0_i}}^{d_i}\int_x^{d_i} w\_d(x)\cdot f(z-x)\,dz\,dx\right)\right]$$

**Total incidence (during the study period) without screening in use**

$D := G_{u1} + G_{u2}$

***Incidence with ever smoked screening characteristics***

**This formula represents the amount of people who would develop cancer during the study period but are detected through screening to have cancer before the start of the study period and thus are left out of the study**

$$b_{11} := \sum_i \left[\omega_i\cdot\int_0^{t_{0_i}}\int_{t_{0_i}}^{d_i} (w\_b(x))\cdot f(z-x)\cdot k\_b_3(x)\cdot\left[1-(1-\xi)^{\int_{\max(screen_{age},x)}^{t_{0_i}} k\_b_1(y)\,dy}\right]dz\,dx\right]$$

---

This formula represents the amount of people who would develop cancer after the study period but are detected through screening to have cancer before the end of the study period and thus are included in the study

$$b_{12} := \sum_i \left[ \left[ \omega_i \cdot \int_0^{t_{0_i}} \int_{d_i}^{max_{age}} [(w\_b(x)) \cdot f(z-x)] \cdot \left[ k\_b_3(x) \cdot (1-\xi)^{\int_{\min(\max(screen_{age},x),t_{0_i})}^{t_{0_i}} k\_b_1(y)\,dy} \cdot \left[ k\_d_3(x) \cdot \left[ 1-(1-\xi)^{\int_{t_{0_i}}^{d_i} k\_d_1(y)\,dy} \right] \right] \right] dz\,dx \right] \ldots \right.$$
$$\left. + \omega_i \cdot \int_{t_{0_i}}^{d_i} \int_{d_i}^{max_{age}} w\_d(x) \cdot f(z-x) \cdot k\_d_3(x) \cdot \left[ 1-(1-\xi)^{\int_{\max(screen_{age},x)}^{d_i} k\_d_1(y)\,dy} \right] dz\,dx \right]$$

Total incidence due to screening among those who ever smoked

$$B_1 := b_{11} + b_{12}$$

### *Incidence with never smoked screening characteristics*

This formula represents the amount of people who would develop cancer during the study period but are detected through screening to have cancer befo the start of the study period and thus are left out of the study

$$b_{21} := \sum_i -\left[ \omega_i \cdot \int_0^{t_{0_i}} \int_{t_{0_i}}^{d_i} (w\_b(x)) \cdot f(z-x) \cdot k\_b_4(x) \cdot \left[ 1-(1-\xi)^{\int_{\max(screen_{age},x)}^{t_{0_i}} k\_b_2(y)\,dy} \right] dz\,dx \right]$$

This formula represents the amount of people who would develop cancer after the study period but are detected through screening to have cancer befor the end of the study period and thus are included in the study

$$b_{22} := \sum_i \left[ \left[ \omega_i \cdot \int_0^{t_{0_i}} \int_{d_i}^{max_{age}} (w\_b(x)) \cdot f(z-x) \cdot \left[ k\_b_4(x) \cdot (1-\xi)^{\int_{\min(\max(screen_{age},x),t_{0_i})}^{t_{0_i}} k\_b_2(y)\,dy} \cdot \left[ k\_d_4(x) \cdot \left[ 1-(1-\xi)^{\int_{t_{0_i}}^{d_i} k\_d_2(y)\,dy} \right] \right] \right] dz\,dx \right] \ldots \right.$$
$$\left. + \omega_i \cdot \int_{t_{0_i}}^{d_i} \int_{d_i}^{max_{age}} w\_d(x) \cdot f(z-x) \cdot k\_d_4(x) \cdot \left[ 1-(1-\xi)^{\int_{\max(screen_{age},x)}^{d_i} k\_d_2(y)\,dy} \right] dz\,dx \right]$$

Total incidence due to screening among those who never smoked

$$B_2 := b_{21} + b_{22}$$

## RR ever smoked (1) vs never smoked (2)

Relative Risk of smokers vs nonsmokers shows the effect of lead time (correction for the observed measure of association)

$$RR_{1v2} := \frac{D + B_1}{D + B_2} \qquad RR_{1v2} = \blacksquare$$

### *Sample R code*

```
####################################################################
## Functions to use to find screening patterns for all datasets ##
####################################################################



pack_yr_dist.f<-function(case_control, smoked,
smoked_stop_20,filename,dataset_name,...){
##plot histograms to show distribution of pack-years among cases and controls defining
nonsmokers as "true" never smokers and then as never smokers plus smokers who quit >
20 years ago##
	dataset_name$case_smoked<-ifelse(case_control==1,ifelse(smoked==1,1,0),0)
	dataset_name$case_nonsmoked<-
ifelse(case_control==1,ifelse(smoked==0,1,0),0)
	dataset_name$case_smoked_20<-
ifelse(case_control==1,ifelse(smoked_stop_20==1,1,0),0)
	dataset_name$case_nonsmoked_20<-
ifelse(case_control==1,ifelse(smoked_stop_20==0,1,0),0)
	pdf(file=filename)
	attach(dataset_name)
	hist(pack_yr[case_smoked==1],main=paste("Histogram of pack years among
cases that smoked"),xlab="Pack-years",ylim=c(0,4000))
	hist(pack_yr[case_nonsmoked==1],main=paste("Histogram of pack years among
cases that didn't smoke"),xlab="Pack-years",xlim=c(0,150),ylim=c(0,4000))
	hist(pack_yr[case_smoked_20==1],main=paste("Histogram of pack years among
cases that smoked","\n","(excluding those that quit >20 years ago)"),xlab="Pack-
years",ylim=c(0,4000))
	hist(pack_yr[case_nonsmoked_20==1],main=paste("Histogram of pack years
among cases that didn't smoke","\n","(including those that quit >20 years
ago)"),xlab="Pack-years",ylim=c(0,4000))
	detach(dataset_name)
	dev.off()
}

prop_scrn_eq<-
function(age_min,age_max,age_by,age,age_squared,screened,coeff1,coeff2,coeff3,filena
me,filenamesink,group,...){
##fit a nonlinear function - second degree polynomial equation - to proportion of
smoking group that screened##
	scrn<-screened
	age<-age
	age2<-age_squared
```

```
	p1<-coeff1
	p2<-coeff2
	p3<-coeff3
	nlfit<-nls(scrn~x1*age2+x2*age+x3, start=list(x1=p1,x2=p2,x3=p3))
	age_cat<-seq(age_min,age_max, age_by)
	age_cat2<-age_cat*age_cat
	est_b<-0
	est_b[1:3]<-summary(nlfit)$parameters[1:3]
	#divert output to file named filenamesink
	sink(file ="prop and rate equations.csv", append = TRUE, type = "output", split =
FALSE)
	cat(filenamesink, "\n", "Estimate 1", "Estimate 2", "Estimate 3", "\n")
	cat(est_b[1],"\n",est_b[2],"\n",est_b[3],"\n")
	sink()
	eqt_b<-est_b[1]*age_cat2+est_b[2]*age_cat+est_b[3]
	pdf(file=filename,width = 6, height = 6)
	#par(mfrow=c(2,2))
	#cat(filename,"\n")
	plot(age, screened, main=filename,xlim=c(55,85))
	points(age_cat,eqt_b,col="blue",type="l")
	#text(66.5,.401,expression({list(est_b[1])*{x^2}+list(est_b[2])*{x}+list(est_b[3])
}),pos=4,cex=.8)
	dev.off()
}

rate_scrn_eq<-
function(age_min,age_max,age_by,age,age_squared,coeff1,coeff2,coeff3,filename,filena
mesink,scrn_rate,num_scrned,...){
##fit a nonlinear function - second degree polynomial equation - to screening rate in
specified smoking group
	scrn_rt<-scrn_rate[num_scrned>0]
	print(length(scrn_rt))
	age<-age[num_scrned>0]
	#if(length(scrn_rt)==1) scrn_rt<-rep(1,length(age))
	age2<-age_squared[num_scrned>0]
	p1<-coeff1
	p2<-coeff2
	p3<-coeff3
	nlfit<-nls(scrn_rt~x1*age2+x2*age+x3, start=list(x1=p1,x2=p2,x3=p3))
	age_cat<-seq(age_min,age_max, age_by)
	age_cat2<-age_cat*age_cat
	est_b<-0
	est_b[1:3]<-summary(nlfit)$parameters[1:3]
	#divert output to file named filenamesink
```

```
	sink(file = "prop and rate equations.csv", append = TRUE, type = "output", split =
FALSE)
	cat(filenamesink, "\n","Estimate 1", "Estimate 2", "Estimate 3", "\n")
	cat(est_b[1],"\n",est_b[2],"\n",est_b[3],"\n")
	sink()
	eqt_b<-est_b[1]*age_cat2+est_b[2]*age_cat+est_b[3]
	pdf(file=filename,width = 6, height = 6)
	#par(mfrow=c(2,2))
	#cat(filename,"\n")
	plot(age, scrn_rt, main=filename,xlim=c(55,85))
	points(age_cat,eqt_b,col="blue",type="l")
	#text(66.5,.401,expression({list(est_b[1])*{x^2}+list(est_b[2])*{x}+list(est_b[3])
}),pos=4,cex=.8)
	dev.off()
}
######################################################################
## code to find table for screening pattern by ever vs. never smoked across all 100
samples ##
######################################################################
## dataset ign_c_t3_t5 ##
#########################

################################################################
## Copy the following into R to run program on specific dataset ##
################################################################
#setwd("Q:/PhD work/PhD study designs")
#source("Rfunctions")
#setwd("Q:/PhD work/PhD study designs/ign_c_t3_t5")
#source("R Code ign_c_t3_t5.r")
################################################################


#load in functions
setwd("Q:/PhD work/PhD study designs")
source("Rfunctions")

#set global options
options(scipen=2) #displays significant digits to ^-6 then uses scientific notation
options(expressions=5000) #increase limit of memory for stored local variables

## study design: ign_c_t3_t5 ###

setwd("Q:/PhD work/PhD study designs/ign_c_t3_t5")
```

```
data_set<-read.table(file="ign_c_t3_t5.csv",sep=",",header=TRUE) ##read in the data
from a CSV file##
attach(data_set) ##allow me to use just variable names as displayed in dataset##
pack_yr_dist.f(case_control, smoked,
smoked_stop_20,"pack_yr_dist_ign_c_t3_t5.pdf",data_set)

########################
## Logistic Regression ##
########################

case<-ifelse(case_control==1,1,0)
OR_age<-rep(0,100)
OR_smoked<-rep(0,100)
age_upperCL<-rep(0,100)
age_lowerCL<-rep(0,100)
smoked_upperCL<-rep(0,100)
smoked_lowerCL<-rep(0,100)
smoked_stderr<-rep(0,100)
nsmk_case<-rep(0,100)
for(i in 1:100){
log_result<-
glm(case[smoked<99&sample_num==i]~age_t3[smoked<99&sample_num==i]+smoked
[smoked<99&sample_num==i],family=binomial(logit))
age_est<-summary(log_result)$coefficient[2]
smoked_est<-summary(log_result)$coefficient[3]
age_stderr<-summary(log_result)$coefficient[5]
smoked_stderr[i]<-summary(log_result)$coefficient[6]
age_zvalue<-1.96
smoked_zvalue<-1.96
age_upperCL[i]<-exp(age_est+age_zvalue*age_stderr)
age_lowerCL[i]<-exp(age_est-age_zvalue*age_stderr)
smoked_upperCL[i]<-exp(smoked_est+smoked_zvalue*smoked_stderr)
smoked_lowerCL[i]<-exp(smoked_est-smoked_zvalue*smoked_stderr)
OR_age[i]<-exp(age_est)
OR_smoked[i]<-exp(smoked_est)
nsmk_case[i]<-sum(case[smoked==0&sample_num==i])
}
mean(OR_smoked[1:100]+1.96*mean(exp(smoked_stderr[1:100])))
mean(OR_smoked[1:100]-1.96*mean(exp(smoked_stderr[1:100])))


hist(nsmk_case)

##histogram for simulated RRs from 100 repeat samples
```

```
hist(OR_smoked, main="Sampling from entire PLCO enrollment period in usual-care group", xlab="Observed RR")
abline(v=mean(OR_smoked))
abline(v=median(OR_smoked),col="red")
abline(v=quantile(OR_smoked,prob=c(0.025)),col="blue")
abline(v=quantile(OR_smoked,prob=c(0.975)),col="forestgreen")
legend("topright",legend=c("Mean","Median","2.5% limit","97.5% limit"),lty=c(1),col=c("black","red","blue","forestgreen"))

###output variance of smoked OR to csv file in study designs folder
setwd("Q:/PhD work/PhD study designs")
sink(file="variations for smoked ORs.csv", append=TRUE, type="output")
cat(paste("ign_c_t3_t5",mean(OR_smoked),var(OR_smoked),(var(OR_smoked))/100,sep=","),"\n")
sink()
setwd("Q:/PhD work/PhD study designs/ign_c_t3_t5")

mean_upperCI<-mean(OR_smoked)+1.96*mean(smoked_stderr)
mean_lowerCI<-mean(OR_smoked)-1.96*mean(smoked_stderr)

sink(file ="quantiles for 100 samples.csv", append = FALSE, type = "output", split = FALSE)
cat("OR age", "\n")
print(quantile(OR_age,probs=c(0,0.025,0.25,0.5,0.75,0.975,1)))
cat("OR smoked", "\n")
print(quantile(OR_smoked,probs=c(0,0.025,0.25,0.5,0.75,0.975,1)))
cat("Upper Wald Confidence Limit age", "\n")
print(quantile(age_upperCL,probs=c(0,0.025,0.25,0.5,0.75,0.975,1)))
cat("Lower Wald Confidence Limit age", "\n")
print(quantile(age_lowerCL,probs=c(0,0.025,0.25,0.5,0.75,0.975,1)))
cat("Upper Wald Confidence Limit smoked", "\n")
print(quantile(smoked_upperCL,probs=c(0,0.025,0.25,0.5,0.75,0.975,1)))
cat("Lower Wald Confidence Limit smoked", "\n")
print(quantile(smoked_lowerCL,probs=c(0,0.025,0.25,0.5,0.75,0.975,1)))
cat("Upper Wald Confidence Interval smoked", "\n",mean_upperCI,"\n")
cat("Lower Wald Confidence Interval smoked", "\n",mean_lowerCI,"\n")
sink()

detach(data_set)


###################
# Smoking strata ##
###################
```

```
##before study period screening pattern functions for Smokers
table(data_set$smoked, data_set$screened_b)
dat_smoked<-subset(data_set,smoked==1&case_control==2&screened_b<99)
table(dat_smoked$smoked, dat_smoked$screened_b)
print(length(dat_smoked$screened_b))
attach(dat_smoked)
par(mfrow=c(2,2))
prop_scrn_eq(55,80,1,age_t3,age_t3_sq,screened_b,.0002,.0015,1,
"prop_smk_scrn_before.pdf","equations_for_prop_scrn_Smokers_before.txt",graph_title
="proportion of ever smokers screened before the study
usual-care group")
rate_scrn_eq(55,80,1,age_t3,age_t3_sq,.0002,.0015,1,
"rate_smk_scrn_before.pdf","equations_for_scrn_rate_Smokers_before.txt",screen_rate_
b,num_screened_b,graph_title="rate of screening among ever smokers before the study
usual-care group")



##during study period screening pattern functions for Smokers
table(dat_smoked$smoked, dat_smoked$screened_d)
print(length(dat_smoked$screened_d))
prop_scrn_eq(55,80,1,age_t3,age_t3_sq,screened_d,.0002,.0015,1,
"prop_smk_scrn_during.pdf","equations_for_prop_scrn_Smokers_during.txt",graph_title
="proportion of ever smokers screened during the study
usual-care group")
rate_scrn_eq(55,80,1,age_t3,age_t3_sq,1,1,1,
"rate_smk_scrn_during.pdf","equations_for_scrn_rate_Smokers_during.txt",screen_rate_
d,num_screened_d,graph_title="rate of screening among ever smokers during the study
usual-care group")
par(mfrow=c(1,1))
detach(dat_smoked)

######################
# Non-smoking strata ##
######################

##before study period screening pattern functions for Nonsmokers
table(data_set$smoked, data_set$screened_b)
#here change smoked variable and dataset name (1)
dat_nonsmoked<-subset(data_set,smoked==0&case_control==2&screened_b<99)
##change dataset name (2)
table(dat_nonsmoked$smoked, dat_nonsmoked$screened_b)
##change dataset name (1)
```

```
print(length(dat_nonsmoked$screened_b))
##change dataset name (1)
attach(dat_nonsmoked)
par(mfrow=c(1,2))
##change file names for next two lines (4)
prop_scrn_eq(55,80,1,age_t3,age_t3_sq,screened_b,.0002,.0015,1,
"prop_nonsmk_scrn_before.pdf","equations_for_prop_scrn_Nonsmokers_before.txt",gra
ph_title="proportion of never smokers screened before the study
usual-care group")
rate_scrn_eq(55,80,1,age_t3,age_t3_sq,.0002,.0015,1,
"rate_nonsmk_scrn_before.pdf","equations_for_scrn_rate_Nonsmokers_before.txt",scree
n_rate_b,num_screened_b,graph_title="rate of screening among never smokers before
the study
usual-care group")



##during study period screening pattern functions for Nonsmokers
#change dataset name in next two lines (3 total)
table(dat_nonsmoked$smoked, dat_nonsmoked$screened_d)
print(length(dat_nonsmoked$screened_d))
##change file names for next two lines (4)
prop_scrn_eq(55,80,1,age_t3,age_t3_sq,screened_d,.0002,.0015,1,
"prop_nonsmk_scrn_during.pdf","equations_for_prop_scrn_Nonsmokers_during.txt",gra
ph_title="proportion of never smokers screened during the study
usual-care group")
rate_scrn_eq(55,80,1,age_t3,age_t3_sq,1,1,1,
"rate_nonsmk_scrn_during.pdf","equations_for_scrn_rate_Nonsmokers_during.txt",scree
n_rate_d,num_screened_d,graph_title="rate of screening among never smokers during
the study
usual-care group")
par(mfrow=c(1,1))
##change dataset name (1)
detach(dat_nonsmoked)
```

***Sample SAS code***

```
/*After opening the sas file in browser mode, just copy
code from below.  Then can run programs on data.
*/

filename in 'Q:\PhD work\PhD study designs\entire_c_t0_t5';
libname out 'Q:\PhD work\PhD study designs\entire_c_t0_t5';
OPTIONS FIRSTOBS=1;


data work.Entire2_c_t0_t5;
     set Out.entire_c_t0_t5;

     screened_b=99;
     if XRAY >0 and XRAY < 3 then screened_b=1;
     if XRAY = 0 then screened_b=0;



     screened_d = 0;
     if had_screen0 = 1 then screened_d=1;
     if had_screen1 = 1 then screened_d=1;
     if had_screen2 = 1 then screened_d=1;
     if had_screen3 = 1 then screened_d=1;
     if XRAY >0 and XRAY<3 then screened_d=1;

     smoked=99;
     if cig_stat = 0 then smoked=0;
     if cig_stat >0 and cig_stat<3 then smoked=1;

     nsmk_nscrn_b = 0;
     if screened_b=0 and smoked=0 then nsmk_nscrn_b=1;

     nsmk_scrn_b = 0;
     if screened_b=1 and smoked=0 then nsmk_scrn_b=1;

     smk_nscrn_b = 0;
     if screened_b=0 and smoked=1 then smk_nscrn_b=1;

     smk_scrn_b = 0;
     if screened_b=1 and smoked=1 then smk_scrn_b=1;

     diff_nsmk_b = nsmk_scrn_b - nsmk_nscrn_b;
```

```
diff_smk_b = smk_scrn_b - smk_nscrn_b;

nsmk_nscrn_d = 0;
if screened_d=0 and smoked=0 then nsmk_nscrn_d=1;

nsmk_scrn_d = 0;
if screened_d=1 and smoked=0 then nsmk_scrn_d=1;

smk_nscrn_d = 0;
if screened_d=0 and smoked=1 then smk_nscrn_d=1;

smk_scrn_d = 0;
if screened_d=1 and smoked=1 then smk_scrn_d=1;

diff_nsmk_d = nsmk_scrn_d - nsmk_nscrn_d;

diff_smk_d = smk_scrn_d - smk_nscrn_d;

cigs = 0;
if cigpd_f=1 then cigs = 5;
if cigpd_f=2 then cigs = 15;
if cigpd_f=3 then cigs = 25;
if cigpd_f=4 then cigs = 35;
if cigpd_f=5 then cigs = 50;
if cigpd_f=6 then cigs = 70;
if cigpd_f=7 then cigs = 90;

years = 0;
if cig_years="M" then years = 0;
if cig_years="N" then years = 0;
if cig_years>0 then years = cig_years;

pack=cigs/20;

pack_yr=pack*years;

stop_c = 0;
if cig_stop="A" then stop_c=0;
if cig_stop="F" then stop_c=0;
if cig_stop="M" then stop_c=0;
if cig_stop="N" then stop_c=0;
if cig_stop >0 then stop_c = cig_stop;

stop_20 = 0;
if stop_c<21 and stop_c>0 then stop_20 = 1;
```

```
	smoked_stop_20 = 0;
	if smoked=1 and stop_20=1 then smoked_stop_20=1;
	if smoked=99 then smoked_stop_20=99;

	age_t3 = age;

	age_t3_sq = age_t3 * age_t3;

	num_screened_b = 0;
	if XRAY >=0 and XRAY <3 then num_screened_b = XRAY ;
	/*if XRAY=3 or XRAY=0 then num_screened_b =
had_screen0 + had_screen1;*/

	screen_rate_b = 0;
	if screened_b=1 then screen_rate_b = num_screened_b/2;
	/*assume XRAY variable answer for 2 years before study
instead of 3*/

	num_screened_d = 0;
	if 0<XRAY<3 then num_screened_d = XRAY*2 + had_screen0
+ had_screen1 + had_screen2 + had_screen3;
	if XRAY=3 or XRAY=0 then num_screened_d = had_screen0
+ had_screen1 + had_screen2 + had_screen3;

	screen_rate_d = 0;
	if screened_d=1 then screen_rate_d = num_screened_d/6;

run;

proc export data=work.Entire2_c_t0_t5
	outfile='Q:\PhD work\PhD study
designs\entire_c_t0_t5\entire_c_t0_t5.csv'
	dbms=csv
	 replace;
	delimiter=',';
run;
```