Measuring the Mortality Reductions due to Cancer Screening

Zhihui (Amy) Liu

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University

Montréal, Québec

2014-07-01

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

DEDICATION

This thesis is dedicated to my parents, Qiulan Zhao and Yunsheng Liu.

ACKNOWLEDGEMENTS

I have many people to thank, for having supported me in the past years and assisted me in completing this thesis.

My deepest thanks go to my supervisor Dr. James Hanley for his mentorship, support and advice (both professional and personal). I cannot thank him more for selecting an interesting and challenging thesis topic for me. He believes in "see one, do one, teach one", and made sure that I get training in teaching (e.g. Unit 8 epidemiology teaching) and in applied health research (e.g. CKCis project), in addition to my thesis work. His passion for statistics has greatly influenced those around him. He has always made himself available, even when he was on sabbatical, thousands of miles away. He has truly been a wonderful mentor to me.

I am sincerely grateful for my co-supervisor Dr. Nandini Dendukuri, for allowing me freedom to make my own mistakes and discoveries, whilst ensuring that I remained on the right track. She has always promptly responded to my questions and concerns, and her encouragement was instrumental especially when my manuscripts received not-so-glowing reviews. I would also like to thank both of my supervisors for financially supporting my traveling to conferences.

I am continuously grateful for the CKCis project lead Dr. Simon Tanguay, for his support, guidance and confidence in me, and for sharing his clinical expertise in kidney cancer during the past three years. He kindly tries to protect my time but also pushes me to move forward. Together with the project manager Ms. Wendella Hamilton, their dedication has been a constant inspiration to me.

Highest regards go to Dr. Rebecca Fuhrer, forever the chairwoman in my mind, for her total commitment and dedication to our department.

I thank Dr. Olli Miettinen for encouraging us to always think from the first principles: *Nullius in verba.* I am privileged for having met and attended lectures from the greatest modern epidemiologist.

I thank Dr. Erin Strumpf for serving on my thesis committee and helping out with my first manuscript. Many thanks to Dr. Robert Platt for his causal inference course which was unexpectedly helpful in reshaping my thinking, and for agreeing to serve as my internal examiner. Thank you to Dr. Thomas Lumley for serving as my external examiner.

I gratefully acknowledge the financial support I received through the McGill International Doctoral Awards for my PhD studies and the CIHR Operating Grant for studying my thesis topic.

I thank our faculty members for their daily effort to provide us with excellent learning opportunities and the friendly Student Affairs Officers for making Purvis Hall a home to us all. A big thank you to Luc Villandré for translating the thesis abstract, and my friends and colleagues, such as Benjamin, Bill, Esther, Jason, Mireille, Opal, Raluca, William and Yewwei, for many philosophical discussions.

Olli Saarela's understanding, commitment and faith in me is greatly appreciated. *Tous les jours, les bons comme les mauvais.*

My mum and dad deserve not only thanks for unconditional love and support during all phases of my life, but also apologies for pursuing my studies so far away from home for so many years. To them I dedicate this thesis.

ABSTRACT

Evidence of benefits due to cancer screening is commonly reported as the mortality reduction over the entire follow-up window of a randomized screening trial. However, such a single number summary statistic is of limited use in projecting the timing, duration and magnitude of the mortality reductions that would be expected from a sustained screening program, of longer duration and possibly with a different screening regimen. Meta-analyses, by averaging such measures from trials with varying follow-up windows and screening regimens, have produced summaries that are even less meaningful.

This thesis, composed primarily of four manuscripts, presents theoretical and methodological developments for measuring the mortality reductions due to cancer screening. The objective is to project the time-specific reductions in mortality that would be produced by a sustained screening program, using data from randomized trials, with the aim to give policy makers and funders more accurate evidence on how effective screening programs are and could be.

In the first manuscript, we propose using a mortality reduction curve, instead of a single-number summary, to address the mortality impact (timing, magnitude, and duration) of a screening program. We illustrate when and how such curves from randomized trials could be computed, and how they could be used to project reduction patterns expected with different screening regimens.

In the second manuscript, instead of modelling the entire history of the cancer progression, we develop a novel probability model to address the mortality reductions, by parametrizing the conditional probability of being helped by a single round of screening, given that the cancer would have proven fatal otherwise. We (i) show that this conditional probability can be directly interpreted as the reduction in disease-specific mortality, (ii) suggest a parametric form for it, based on which we formulate a likelihood function, and (iii) extend this model to accommodate unequal allocation, less than full compliance, combination of information across trials with different regimens, as well as different regimens within a trial. Two case studies are presented using data from screening trials for lung and colorectal cancers.

A more detailed analysis of the data from the US National Lung Screening Trial is presented in the third manuscript. We demonstrate that our model can be fitted to both individual-level data and aggregated data, with very little precision lost when using the aggregated data.

All the aggregated mortality data used in this thesis were extracted via a new reconstruction technique we propose in the fourth manuscript. Using examples and an error analysis, we illustrate the extent to which, with what accuracy and precision, and in what circumstances, information can be recovered from the various electronic formats. Compared with previous approaches, one advantage of ours is that observer variation is completely eliminated and thus the extraction is completely replicable.

ABRÉGÉ

On illustre communément les bienfaits attribuables au dépistage du cancer par la réduction de la mortalité observée à travers la période de suivi d'un essai randomisé de dépistage. Toutefois, une telle statistique numérique n'est pas très utile pour prédire le moment, la durée et la magnitude de la réduction de la mortalité résultant d'un programme de dépistage soutenu, plus long et, possiblement, comportant un régime de dépistage différent. Les méta-analyses impliquent un calcul de la moyenne de ces mesures, obtenues à partir d'essais avec des périodes de suivi et des régimes de dépistage variables, et produisent par conséquent des estimés sommaires difficiles à interpréter.

Cette thèse, formée de quatre manuscrits, présente des développements théoriques et méthodologiques permettant la mesure de la réduction de la mortalité attribuable au dépistage du cancer. Nous cherchons à prévoir à travers le temps, à l'aide de données tirées d'essais randomisés, la réduction de la mortalité résultant d'un programme soutenu de dépistage du cancer. Nous souhaitons ainsi donner aux décideurs politiques et aux agences de financement une idée plus précise de l'efficacité observée et potentielle des programmes de dépistage.

Dans le 1er manuscrit, nous proposons d'utiliser une courbe de réduction de la mortalité plutt qu'une statistique numérique unidimensionelle afin de quantifier l'impact d'un programme de dépistage. Nous illustrons quand et comment une telle courbe, dérivée à partir de résultats d'essais randomisés, peut être produite, et comment on peut l'utiliser pour prédire les motifs de réduction espérés à partir de

différents programmes de dépistage.

Dans le 2e manuscrit, au lieu de modéliser l'historique entier du cancer, nous développons un nouveau modèle probabiliste quantifiant l'impact sur la mortalité, en paramétrisant la probabilité conditionnelle de bénéficier d'une seule ronde de dépistage, à condition que le cancer soit fatal autrement. Nous démontrons tout d'abord que cette probabilité conditionnelle peut être interprétée directement comme une réduction de la mortalité spécifique au cancer. Nous suggérons par la suite une formulation paramétrique pour cette probabilité, à partir de laquelle nous obtenons une fonction de vraisemblance. Enfin, nous élargissons le modèle afin de permettre une allocation inégale, une adhérence incomplète, et la combinaison d'informations provenant d'essais comportant des régimes différents ou provenant de régimes différents à l'intérieur d'un même essai. Nous présentons deux études de cas avec des données d'essais de dépistage des cancers colorectaux et du cancer du poumon.

Nous présentons dans le 3e manuscrit une analyse plus détaillée des données tirées des *US National Lung Screening Trials*. Nous démontrons que notre modèle peut être appliqué à des données individuelles ou agrégées, la perte de précision étant minime quand les données sont agrégées.

Nous avons extrait toutes les données de mortalité agrégées par l'intermédiaire d'une nouvelle technique de reconstruction que nous proposons dans le 4e manuscrit. à l'aide d'exemples et d'une analyse d'erreurs, nous illustrons la précision de l'information qu'on peut recouvrer à partir de différents formats électroniques. Comparativement aux approches précédentes, la ntre a l'avantage d'éliminer la variation entre observateurs et par conséquent, l'extraction est complètement reproductible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE: CONTRIBUTIONS OF AUTHORS

The thesis comprises an introductory part in Chapters 1–3 and four manuscripts in Chapters 4–7 containing original research. The authors' contributions to the manuscripts are as follows.

1. The research problem was conceived by James Hanley (JH) and he together with Zhihui Liu (ZL) were jointly responsible for formulating, implementing and applying the methods, and drafting the manuscript. Erin Strumpf served as an advisor, provided commentary and contributed to the editing of the manuscript.
2. The research problem was conceived by JH. ZL, JH and Olli Saarela (OS) were jointly responsible for formulating the theory and methods, and drafting the manuscript. ZL was responsible for implementing and applying the methods. Nandini Dendukuri served as an advisor, provided commentary and contributed to the editing of the manuscript.
3. The research problem was conceived by JH and ZL, who also obtained access to the NLST data. ZL was mainly responsible for implementing and applying the methods, and drafting of the manuscript. JH and OS contributed to the implementation of the methods, and commented and edited the manuscript.

4. The research problem was conceived by JH and ZL. ZL and JH were jointly responsible for formulating, implementing and applying the methods. Benjamin Rich provided critical commentary and contributed to the editing of the manuscript.

PREFACE: ETHICS APPROVAL

The manuscripts in this thesis include analyses of previously collected data from human subjects. Ethics approval for the collection of the data was obtained by the original studies. The use of the NLST data (in manuscripts 2 and 3) was guided by a data transfer agreement signed by JH, available upon request.

CHAPTER 1
Introduction

Screening for a disease is pursuit of its early, pre-symptomatic diagnosis, with the aim to reduce the probability of dying from the disease [61]. Over the last century, a range of activities, which we now think of as health or medical screening, have been developed. These activities include the use of bloodspot tests in newborn babies, the Mantoux test to screen for exposure to tuberculosis, the Beck Depression Inventory to screen for depression, ultrasound scans for abdominal aortic aneurysm, computerized tomography (CT) or magnetic resonance imaging (MRI) scans of the whole body for the worried well, as well as screening for cancers – such as pap smear to detect potentially precancerous lesions and prevent cervical cancer, mammography to detect breast cancer, colonoscopy to detect colorectal cancer, and faecal occult blood test for colorectal cancer. In many countries, elementary schools screen students periodically for hearing and vision deficiencies and dental problems. Italy launched a nationwide systematic cardiac screening program in 1982 for all competitive athletes to prevent sudden cardiac death during sports [18].

Among these, screening for cancer has been one of the most controversial activities. Canada and other countries have devoted a lot of resources to screening programs for cancer over the last 40-50 years. Despite many long and costly randomized screening trials (for breast, prostate, lung and colorectal cancers) involving large numbers of participants, we do not have good answers to the question of how

large the 'returns' (in terms of the numbers of cancer deaths averted) are for the dollars spent on actual screening programs that have been in operation. Policy makers and funders are thus faced with a wide array of uncertain and conflicting figures. Worse, the public has become confused by different advice from various authorities, although these authorities all have the same data.

The benefit, particularly reductions in cancer deaths, is studied typically by means of a randomized trial (or a meta-analysis of randomized trials), in which asymptomatic persons are randomly assigned to receive either a number of screening examinations or usual care, and then are followed up for cancer-specific deaths. Given the need to first establish proof of concept, most screening trials have been in a hypothesis-testing (zero vs. non-zero reduction) framework. Understandably, results are announced when the accumulated mortality reduction first becomes statistically significantly different from zero, as in the European Randomized Study of Screening for Prostate (ERSPC) [80] and the US National Lung Screening Trial (NLST) [72]. However, for funders, the question is not whether the reduction is 'almost definitely' nonzero – presumably by now we all know that screening saves some lives, but rather "how many fewer cancer deaths would there be every year in their country or administrative region as a result of a screening program with a specific regimen?"

Influential reports of many randomized screening trials (in cancers of the prostate, colon, and lung) have appeared recently [80, 8, 72]. The reported mortality reductions were all around 20%. Presumably, the resulting reductions depend on the type of the cancer being screened for, the screening technique used (e.g. PSA for prostate cancer, sigmoidoscopy for colorectal cancer, computed tomography for lung cancer),

the screening regimen (the number of screenings and their spacing), the characteristics of the screenees (age at the start of screening, high or low risk for a particular cancer), as well as the compliance rate. In some trials, the modest reductions are not surprising, if there was only one round of screening. However, in many of them, this 20% is merely an artifact of early-reporting rules. It is not because every single trial produces a 20% mortality reduction, but because the results were announced when the cumulative reduction reaches 20%.

As Miettinen [60] pointed out, "the research need is for estimation of meaningful *component measures* of both the good and the harm... All of this - centrally including the need for estimation rather than mere hypothesis-testing - should go without saying, but does not at present." Estimation of the 'good' is what this thesis aims to address. The objective is to project the time-specific reductions in mortality that would be produced by a sustained screening program, using data from randomized trials, in order to give policy makers and funders more accurate evidence on how effective screening programs are and could be.

While the existing data on screening effectiveness originates from randomized screening trials with only a few rounds of screening, an object of inference more relevant to decision making is the effect of a sustained screening program, with a longer duration and possibly different screening regimen, implemented in a population. This distinction between screening trials and screening programs is central to what follows. Furthermore, screening trials are fundamentally different from trials of therapeutics, as screening itself is not an intervention; the participants in a screening trial are asymptomatic, and any mortality effect can only manifest with a delay after

the screening examination. These characteristics of screening present several statistical challenges, and as we argue, the related statistical theory and methodology is presently underdeveloped, despite the importance of the problem and the substantial resources spend on cancer screening. Thus, the work in this thesis potentially has important public health implications.

The remainder of the thesis is constructed as follows.

In Chapter 2, I use a recent report on a mammography screening trial for breast cancer as an example to review and illustrate some of the principles of cancer screening.

In Chapter 3, by studying a particular micro-simulation model, I show why the prevailing approach of modelling the entire disease history is not suitable for our purpose of projecting the mortality reductions of a screening program. This serves as a motivation for us to pursue a completely different approach.

In Chapter 4, we propose using a 'mortality reduction curve', instead of a single-number mortality reduction, to address the mortality impact (timing, magnitude, and duration) of a screening program. We illustrate when and how such curves from randomized trials could be computed, and how they could be used to project the reduction patterns expected with different screening regimens.

In Chapter 5, we develop a novel probability model to address the mortality reductions, by parametrizing the conditional probability of being helped by a single round of screening, given that the cancer would have proven fatal otherwise. We (i) show that this conditional probability can be directly interpreted as the reduction in disease-specific mortality, (ii) suggest a parametric form for it, based on which we

formulate a likelihood function, and (iii) extend this model to accommodate unequal allocation, less than full compliance, combination of information across trials with different regimens, as well as different regimens within a trial. Two case studies are presented using data from screening trials for lung and colorectal cancers.

A more detailed analysis of the data from the US National Lung Screening Trial is presented in Chapter 6. We demonstrate that our model can be fitted to both individual-level data and aggregated data, with very little precision lost when using the aggregated data.

All the aggregated mortality data used in this thesis were extracted via a new data reconstruction technique that we propose in Chapter 7, based on Kaplan-Meier or Nelson-Aalen survival-type curves. Using worked examples and an error analysis, we illustrate the extent to which, with what accuracy and precision, and in what circumstances, information can be recovered from the various electronic formats in which such curves are stored. Compared with previous approaches, one advantage of ours is that observer variation is eliminated and thus the extraction is completely replicable.

Chapter 8 is a summary.

CHAPTER 2
Mammography Screening: a Controversy that Refuses to Die

In this chapter, we review some history of cancer screening, and the controversies around it, especially relating to randomized screening trials.

2.1 Cancer screening - an orientation

In the medical community, one of the earliest advocates of screening is the British physician Horace Dobell who gave a series of lectures in 1861 encouraging doctors to periodically check everyone irrespective of their health status [77, p. 1]. In 1900, Dr. George Gould presented a paper at the American Medical Association meeting, recommending annual health checks to Americans, just like "ranchers check their cattle, merchants check their stock, generals check their armies and government check their budgets" [77, p. 3]. Although many doctors initially felt that it was a waste of their time and expertise, the practice of the periodic health examinations was endorsed by the American Medical Association in 1922 and had become standard by the 1950s, driven by life-insurance companies [77, p. 6].

The fact that abnormalities were found in the tests was sufficient to convince many people that screening was needed. Whether or not the benefit outweighed the unintended harm and associated costs had not been a concern [77, p. 7]. By 1957, authoritative bodies such as the US Commission on Chronic Illness had recommended screening for diabetes, glaucoma and cancers of the mouth, skin, breast, cervix and rectum, based on expert opinions such as "increasing numbers of physicians and

other health personnel have come to the conclusion that screening tests can be an effective device in secondary prevention of chronic illness" [77, p. 11], without any particular evidence.

When in the 1960s two randomized trials of multiphasic health examinations failed to show benefits in general health and in mortality [77, p. 15-16], the need to properly evaluate the benefits and harms from screening started to become more obvious. A randomized screening trial is a study which randomly assigns asymptomatic persons either to be screened (and treated when diagnosed) or to receive usual care (i.e. not screened and treated only when clinically diagnosed based on symptoms). The expectation is that some of the screened persons would develop abnormalities that can be picked up by screening at a less advanced stage so that the associated early treatment is more successful (than delayed treatment).

To be beneficial, screening-associated early treatment must be proven to lead to a lower mortality or morbidity compared with usual care; reduction in cancer specific mortality is considered the definitive criterion for evaluating the effectiveness of cancer screening. Randomized trials have been used to study the effectiveness of, for instance, mammography screening for breast cancer, prostate-specific antigen (PSA) screening for prostate cancer, fecal occult blood (FOB) test for colorectal cancer, and low-dose computed tomography (CT) screening for lung cancer.

A commonly used measure of benefit is the proportional reduction in disease-specific mortality. We illustrate this using a modified version of Web Figure 1 of Hanley [37]. Figure 2–1 is a schematic figure showing the numbers of cases of (prostate)

cancers that came to attention and that proved fatal over the lifetime of a hypothetical closed population of a given size: c_0 diagnosed cases and d_0 deaths in the absence of a screening program, and c_1 cases and d_1 deaths in the presence of screening. The time spans, screening regimens, and participation rate are deliberately left vague, to keep the example simple at this stage.

Assuming that all cancer would eventually be clinically diagnosed, overdiagnosis due to screening is represented by $(c_1 - c_0)/c_1$. This thesis concentrates on measuring the benefits of screening, and we do not address measuring the harms of screening, including overdiagnosis. This does not mean that the harms of screening would not be real or important, and in eventual public health decision making both harms and benefits would need to be considered.

The number of cancers that would have proved fatal if left untreated, f, is unobservable. This is because cancer often demonstrates a spectrum of behaviours – "some tumours are inherently benign, genetically determined to never reach the fully malignant state, and some tumours are intrinsically aggressive, and intervention at even an early, pre-symptomatic stage might make no difference to the prognosis of a patient" [70, p. 292], and consequently, physicians cannot distinguish between cancers that would and would not have proved fatal if left untreated. As a result, most of the c_0 and c_1 cancers would be treated when they came to clinical attention. (While overdiagnosis does not necessarily cause harm in itself, the subsequent overtreatment does.)

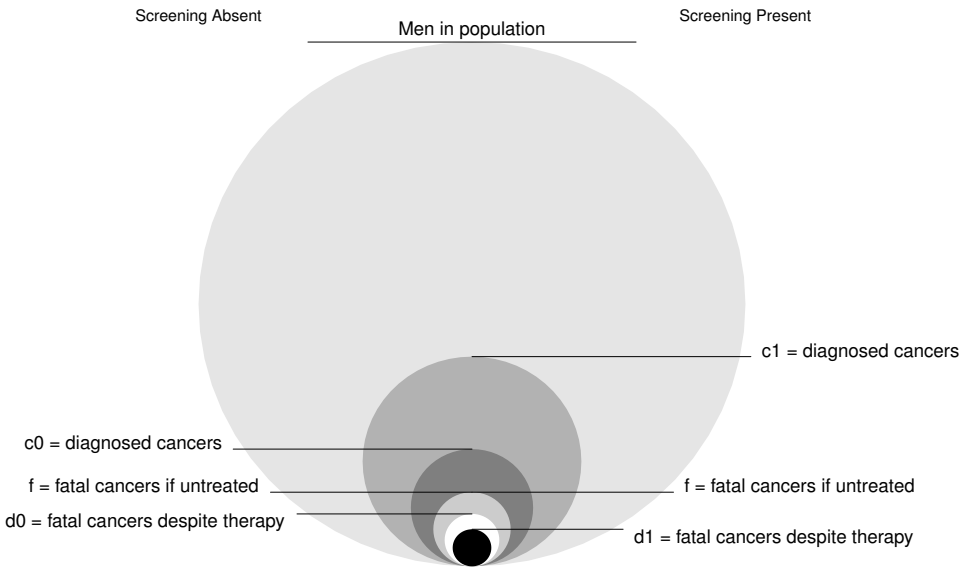


Figure 2–1: Schematic figure illustrating our measure of interest, the proportional reduction in (prostate) cancer mortality $(d_0 - d_1)/d_0$. Modified from Web Figure 1 of Hanley [37].

Despite standard treatment, d_0 and d_1 cancers would eventually prove fatal. These, and *only these*, are the focus of our attention. The difference $d_0 - d_1$, representing the number of deaths averted by early-detection associated early therapy, indicates how effective screening-associated early treatment is. Whereas the d_0 and d_1 individuals who died of these cancers despite treatment can be identified, the $d_0 - d_1$ individuals whose deaths were averted cannot. Presumably some of the d_0 deaths in the absence of screening could have been averted had they been detected and treated earlier. The percentage reduction in cancer mortality due to screening

is thus

$$100 \times \frac{d_0 - d_1}{d_0}.$$

Not all d_0 deaths can be averted, and d_1 cancers will prove fatal despite early detection and early treatment. Some reasons for the d_1 failures might be: low participation in the screening; detected cancers were too advanced to be successfully treated (i.e. aggressive cancers that progressed beyond curability between subsequent rounds of screening, but could have been detected earlier by increasing the frequency of screening); low sensitivity of the screening technique.

A major reason why there is so much controversy around cancer screening is because the amount of overdiagnosis due to screening can be substantial and persons who are overdiagnosed can only be harmed – they often receive invasive treatment that they do not need at the first place. Take the PSA test for prostate cancer as an example: it measures the blood level of PSA (a protein produced by the prostate gland) and men with elevated levels are referred to biopsy. The biopsies inevitably result in detection of some prostate cancers that would never have proven fatal (or even symptomatic) in the absence of PSA screening. Nevertheless, most of them will be treated, often with radical prostatectomy or radiotherapy, which can have serious quality-of-life affecting complications.

Two common cancers with little overdiagnosis (and thus less controversial screening programs associated with them) are the cervical and colorectal cancers. Since the introduction of the Pap smear test in the 1940s, both the number of cervical cancer diagnoses and the death rate from cervical cancer in the US have fallen dramatically. Similarly, the number of both new diagnoses of and deaths from colon cancer have

fallen since 1985 when screening started. Although it is suspected that there could be overdiagnosis and over treatment of precancers of the cervix and colorectal polyps [97, p. 69-71], there is no obvious evidence of over diagnoses of these cancers, which makes screening for them an easier sell.

In the next section, we use a recently updated report on a mammography trial as an example to review and illustrate some of the principles of cancer screening.

2.2 Demystifying cancer screening: using the Canadian mammography trial as an example

Mammography screening for breast cancer has been particularly controversial. There have been eight large randomized trials conducted over the past 50 years since the first one took place in 1963; there have been numerous systematic reviews and meta-analyses of these trials ever since. Yet there remains strong disagreement about the interpretation of these results even when it comes to quantifying the associated benefits or the extent of the overdiagnosis.

Some claim that the benefits in the trials were large for women aged 50 years or older, that is, the reduction in breast cancer mortality in those invited to screening is about 25%, and in women who were actually screened the reduction is about 33% [21]. Some have used the trials to project that a sustained program offering 20 years of screening to women aged 50 to 69, the mortality reduction in breast cancer would be at least 40% [38]. Others, however, argue that the reduction is much more modest – only 10%, and since the rate of overdiagnosis is as high as 50%, they have suggested discontinuing mammography screening [32].

Each time when there was an updated report based on any of these eight trials, a heated debate on the value of mammography broke out all over again. Not surprisingly, the latest BMJ report in February 2014 from the now 25-year-old Canadian National Breast Screening Study (CNBSS) [66] ignited a new round of controversy.

The design of the Canadian study is fairly straightforward – 89,835 women aged 40-59 were randomly assigned to breast examinations by a health professional followed by mammography (five annual screens) or breast examination only during 1980-1985. After 25 years of follow-up, there were a total of 500 and 505 breast cancer deaths in the screening and non-screening arms, respectively. That is, as correctly calculated by the authors, a (25-year breast cancer death) risk ratio of $(500/44925)/(505/44910) \approx 500/505 = 0.99$, or a mortality reduction of 1%. They also calculated that the amount of overdiagnosis due to mammography is 106/484=22%, where 484 is the number of breast cancers diagnosed in the mammography arm and 106 is the number of excess breast cancer cases in the mammography arm.

The negligible, almost nil, reduction in mortality in year 25 is not exactly a new finding, considering that there were never substantial benefits in earlier reports on the same trial. Seven years after the randomization, the cumulative numbers of breast cancer deaths in the 50-59 years old group were 38 and 39 [64], corresponding to a risk ratio of 38/39=0.97, or a reduction of 3%; with a mean of 13 years of follow-up, the numbers were 107 and 105 [65], that is, 2 more breast cancer deaths in the mammography arm.

Just one day after the BMJ report was published, the American College of Radiology and Society of Breast Imaging issued a statement [2], calling it "an incredibly misleading analysis based on the deeply flawed and widely discredited" study. As to why the Canadian study, unlike the other mammography trials, showed little benefit, they re-iterated two reasons: the quality of the mammograms was unacceptably poor and randomization was compromised (in particular, allegedly an excess of women with advanced breast cancers were assigned to the screening arm, which lead to more deaths in that arm). The same complaints were raised many times over the years, mainly by Kopans and Feig [e.g. 48], among others. The co-principal investigator Anthony Miller, on the other hand, responded by claiming that the radiologists are "obviously conflicted" [13], as they read mammograms for a living.

Media coverage did not help. Headlines such as "Vast Study Casts Doubts on Value of Mammograms" in the New York Times [46], followed by reporting that "[o]ne of the largest and most meticulous studies of mammography ever done, involving 90,000 women and lasting a quarter-century, has added powerful new doubts about the value of the screening test for women of any age", would only leave the public even more confused. Aiming at the technical issues, many wrote electronic letters to the BMJ editor in the Rapid Response section, which suggest that understanding of screening studies remain a challenge in the medical community as well as among the public. Whereas the to-screen-or-not-to-screen debate is beyond the scope of this thesis, we aim to introduce some of the key concepts of cancer screening in an accessible way, using selected letters from the Rapid Response as a platform.

Lead time and overdiagnosis

A gynaecologist from Hong Kong writes [27]:

> My sister, age 49, was diagnosed with breast cancer in the UK on the basis of a mammogram. It was 5 cms. When I asked her had she felt it, she said no, and that no-one had ever advised her to check her breasts. My sister was very lucky to be node negative, but after chemo and a local excision she did need a mastectomy. She received excellent care in the NHS. However had she been doing monthly breast self-examination I believe she would have picked up her "lump" when it was below 2 cms – perhaps 6 months - 1 year before – and she would not have needed the mastectomy.

While this sounds entirely plausible, there has been no evidence on benefits of regular self-examination of the breasts [49]. Two other scenarios are also possible: (i) the less-than-2-centimetre lump may not be palpable by self-examination, and (ii) the 5-centimetre lump would not have lead to a fatal cancer in the absence of early detection and therefore the mastectomy would be unnecessary. Since the counterfactual outcome in the anecdote is unobservable, no one knows what would have happened if the mastectomy was not performed.

To illustrate why detecting a smaller tumour earlier is not good enough to prove the value of screening, a story about two identical twins named Hope and Prudence told by Mukherjee [70, p. 293] might help. When offered screening, Hope chooses to be screened and Prudence, suspicious of it, refuses to participate. Some years later, identical forms of cancer develop in them at the same time. Hope is diagnosed in

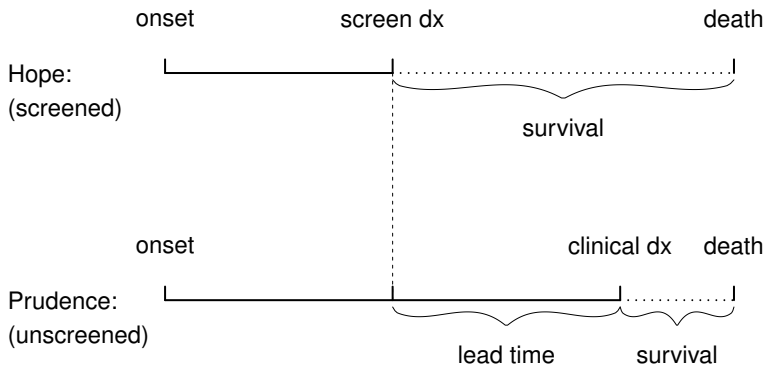


Figure 2–2: Schematic figure showing the disease history of twin sisters Hope and Prudence. Early detection prolongs survival from diagnosis even if death is not delayed.

1995 through early detection, undergoes surgical treatment and chemotherapy, and relapses and then dies in 2000. Prudence is diagnosed when she feels a lump in her breast in 1999; with some marginal benefit from the treatment, she dies at the same time as Hope in 2000. It seems that Hope survives 4 years longer than Prudence, but in fact, both of them die from the same disease at that same time. Figure 2–2 is drawn to illustrate this; it is clear that comparing length of survival from the date of the diagnosis is flawed because early detection pushes the clock of diagnosis earlier in time but does not necessarily delay death. Similarly for the gynaecologist's sister, a smaller tumour may be detected years earlier by either self-examination or a mammogram, but it does not guarantee that she would live longer. This is exactly why randomized trials with mortality as the outcome are needed to establish the benefits of early detection.

A researcher, University of Oxford, UK, says [52]:

The study ignores pre-invasive cancers which have a high survival rate and are often detectable by mammography. It is reasonable to assume that if they had been included in the analysis then the mammography arm of the trial would have demonstrated higher survival rates than those reported. It is not safe to assume that all pre-invasive cancers are indolent (i.e. that they would never go on to harm the patient), especially since the standard model of malignant tumour growth has invasive cancers first developing through a pre-invasive stage.

The study's results contradict the authors' conclusion that mammography is not assisting in saving women's lives. The section of the Results titled "Breast cancer survival" indicates that "The 25 year survival was 70.6% for women with breast cancer detected in the mammography arm and 62.8% for women with cancers diagnosed in the control arm" which the authors demonstrate to be a statistically significant difference. This demonstrates a real benefit to women surviving breast cancer thanks to receiving mammographic screening. Had pre-invasive cancers been included in this study the difference in 25 year survival is liable to have been even larger. Concluding that mammographic screening provides no benefit with respect to saving women's lives based on the analysis presented is unfounded and dangerous.

With a reasonable sensitivity, screening is supposed to detect the cancer earlier, but since this in itself is not an intervention, early detection is not sufficient evidence of any benefits. If survival is counted from diagnosis to death, even without any

treatment, the survival will be longer in the screening arm than that in the non-screening arm, just because screening pushes the diagnosis time earlier. Thus only cancer-specific deaths in the two arms of a randomized study should be counted and compared. Moreover, the denominator in the quoted 70.6% survival probability among the diagnosed individuals in the screening arm includes the overdiagnosed cases, for which the survival probability is, by definition, 100% (since these cancers are non-fatal).

Length-biased sampling

Kopans [47], Professor of Radiology, Harvard Medical School, repeated his criticism that the Canadian trial was compromised since the beginning: "second hand mammography machines" were used; they "failed to fully position the breasts in the machines"; "radiologists had no specific training in mammographic interpretation", and most importantly, the randomization was violated – "women with lumps and even advanced cancers got assigned to the screening arm to be sure they would get a mammogram". Referring to evidence supporting the last allegation, he states that

> It is indisputable that this happened since there was a statistically significant excess of women with advanced breast cancers who were assigned to the screening arm compared to those assigned to the control arm [90]. This guaranteed that there would be more early deaths among the screened women than the control women and this is what occurred in the NBSS. Shifting women from the control arm to the screening arm would increase the cancers in the screening arm and reduce the cancers

in the control arm which would also account for what they claim is "over-diagnosis".

Calling this 'indisputable' seems a strong statement, since at face value, it is not entirely obvious why an excess of advanced cancers early in the screening arm would give evidence supporting failed randomization. Indeed, Gøtzsche [33, p. 57] argues that such an excess does not suggest a failed randomization, "because the screening process increases the likelihood of detecting smaller cancers with positive nodes in the screening arm".

To understand the argument of Prof. Kopans, we indeed need to assume that all of the detected cancers are genuine life-threatening ones, and note that screen-detected cancers are subject to length-biased sampling, that is, a cross-sectional sample from all cancers over time in the preclinical stage oversamples the slow-growing ones, simply because they have been in the preclinical stage longer. Thus, the screen-detected cancers would be expected to be more benign than all cancers occurring in the cohort of women. In particular, screening misses the so-called interval cancers, which are fast growing and move from preclinical to symptomatic stage in between the successive rounds of screening.

Thus, it indeed appears to be unexpected that in the short term the screening arm would show an excess of advanced cancers compared to the control arm, which in turn may be an indication of randomization problems. In the long term, an excess of diagnosed cancers in the screening arm would generally be evidence of overdiagnosis, since eventually the same genuine invasive cancers would manifest both in the absence

and in the presence of screening. However, arguably overdiagnosis is less of an issue when limiting the scope to advanced cancers.

Mortality vs. other outcomes

A Physician, Mount Sinai Roosevelt Division, New York, USA, says [91]:

> Ten year disease-free survival for my patients with mammographically detected cancers is 92 percent compared to 82 percent if the cancer was detected on clinical examination. My results are not exceptional.

This observation in his clinic may not be exceptional, but comparing survival of women detected by mammograms versus those detected by clinical examination is exactly what one should not do in order to prevent the lead-time, length and over-diagnosis biases. The screen-detected cancers are detected earlier (irrespective of whether the subsequent early treatments are effective), exclude the fast growing interval cancers, and include cancers that might not have proven fatal in the absence of early detection.

A Surgeon, Athens University School of Medicine, Greece, also argues that screening should not be evaluated by means of mortality alone [15]:

> Most importantly, a woman who does not die from breast cancer does not mean she does not strive with it; breast cancer remains a very hard and consuming personal, psychological, familial and social adventure for millions of women worldwide. Early detection of a non-palpable breast cancer promptly leads to appropriate management in order to fight the disease at an early stage, and offers optimal care and quality of life.

Early detection unfortunately does not always lead to "appropriate management". When acknowledging that living with breast cancer is not an easy battle, it should also be recognized that receiving unnecessary and often invasive treatment is just as devastating. We usually do not see what would happen if an individual was not treated when diagnosed with cancer, and we certainly cannot observe what would happen to the same individual both in the presence and in the absence of screening. Therefore, the only meaningful comparison and measurable quantity is the difference in mortality between the screened and unscreened groups in a randomized screening trial.

Inferential statistics

An Assistant Professor of Markets, Public Policy, and Law, Boston University School of Management, USA, says [25]:

> Why then do Miller et al. claim that annual mammography screening does not reduce mortality from breast cancer? Because the 95% confidence interval on the effect of screening includes the possibility that it has no effect: the 95% confidence interval on the hazard ratio ranges from 0.88 to 1.12. (A hazard ratio of 1 means no difference in deaths between the two groups.) Nonetheless, the point estimate is that annual screening saves lives, even while the 95% confidence interval indicates substantial uncertainty. [...] While Miller et al. could not reject the hypothesis that annual mammography screening had no effect, they can also not reject the hypothesis that screening saves 135 lives per 100,000 screened, which would justify a screening program substantial costs.

The point estimate of the mortality rate ratio translates to 5 lives saved per 44,910 women screened, or 11 lives saved per 100,000 screened. The confidence interval [0.88, 1.12] is interpreted as ranging from 100,000/(44,910/60.6)=135 lives saved to 135 extra deaths caused per 100,000 screened. While what the commentator says is true, the opposite is also true – the authors could not reject the hypothesis that screening is causing more breast cancer deaths.

Confounding and randomization

Etzioni [26], Statistician, Fred Hutchinson Cancer Research Center, USA, commented on the mortality analyses:

> The first [analysis] looks at the breast cancer death rate restricted to the cases detected during the first five years (the screening period). This is the cumulative death rate in the population but only allows deaths from the cases diagnosed in the first five years. The analysis finds similar death rates on the two arms.
>
> This seems strange because there were there were 666 cases on the mammography arm and 524 cases on the control arm. Thus, if, as the investigators conclude, mammography has no effect, we would actually expect a higher observed cumulative death rate on the mammography arm, unless all of the (142) excess cases are overdiagnosed which is unlikely. Thus, this analysis may bias results against mammography benefit.

When looking at all diagnosed cancers, it is expected that more cases are found in the screening arm, but this should not translate to more deaths in the screening arm compared with the non-screening arm. No matter how effective or ineffective

screening is, it should not produce more cancer deaths. Thus, an observation of excess diagnoses without excess mortality does not really support the earlier assertion that women with abnormalities found in the physical examination were disproportionately placed in the screening arm of the trial. If the randomization in the trial was not compromised, the 142 excess cases, by definition, are overdiagnosed due to screening.

An A&E Physician, Boston University School of Medicine, USA, says [20]:

> I did not see any reference to breast size, shape or 'lumpy, bumpy breast' as possible confounding issues that nullified by these data.

This is a randomized trial, and if the randomization was done properly, all the factors should be balanced between the two arms and therefore there should be no confounding issues.

Finally, Baines [9], Professor emerita, University of Toronto, Canada, defends the validity of the randomization, by saying

> The Canadian study encompassed more than 50 center-years of operation - and Dr. Tabar proposes that one coordinator in a short span of time was able to corrupt the results from 90,000 women by improper randomization of five, ten, even a hundred? of her so-called friends. Absurd. As for the randomization sheets categorized for age by quinquennium, it is hardly surprising that a number of women might first have been entered on sheets designated for women of a different age group. The mistake discovered, the woman would be re-entered on another page. Does Dr. Tabar lives in a world where mistakes are never made - 90,000 correct entries with nary an error?

While almost 90,000 women were randomized, the cumulative numbers of breast cancer deaths were only 500 and 505 in the non-screened and screened groups, respectively, after 25 years of follow-up. In fact there would be a drastic change due to "improper randomization of five, ten, even a hundred" breast cancer deaths – 50 more such deaths in the non-screening arm would result in a mortality reduction of (555-500)/555=10%, and 100 more breast cancer deaths would lead to a (605-500)/605=17% reduction. As also argued by Kopans and Feig [48, p. 758], it does not take much to alter the result, if symptomatic women (presumably some of them with otherwise fatal cancers) were moved from one arm to the other.

In summary, screening trials in cancer are "notoriously difficult to run, and notoriously susceptible to errors" [70, p. 291]; we have here attempted to demystify some basic concepts of cancer screening in an accessible way, using selected responses to the Canadian Study report as examples. In the next section, we review the major guidelines on mammography and meta-analyses that these guidelines were based on, as well as present our careful re-examination of the previous trials.

2.3 Breast cancer mortality reductions due to screening

At the time of writing (February 2014), the U.S. Preventive Services Task Force [95] recommends cytology screening for cervical cancer (every 3-5 years for women age 21 to 65 years), fecal occult blood testing (FOBt) for colorectal cancer (beginning at age 50 and continuing until age 75), annual screening for lung cancer with low-dose computed tomography in heavy smokers aged 55 to 80 years, and recommends against prostate-specific antigen (PSA)-based screening for prostate cancer. However for breast cancer, despite 8 large trials involving 650,000 women having been carried

out in North America and Europe over 5 decades, there is no agreement upon whether mammography screening saves lives, and if so, how many.

The 2002 USPSTF issued a B recommendation for mammography screening every 1 to 2 years for all women older than 40 years. (The grade B recommendation refers to that "there is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial".) This recommendation was based on a meta-analysis by Humphrey et al. [42], which includes all eight randomized screening trials that have been conducted: Health Insurance Plan of Greater New York (a.k.a HIP) [83], Malmö [5], Swedish Two-County [88], Stockholm [29], Gothenburg [12], Edinburgh [4], Canadian [65] and UK AGE trial [69].

In 2009 the USPSTF, further informed by a systematic review [74] and the Cancer Intervention and Surveillance Modeling Network (CISNET) modeling studies [56], recommended biennial mammography screening for all women aged 50 to 74 years and recommended against routine screening of women aged 40 to 49 years. Two years later, a very similar meta-analysis was done for the Canadian Task Force [19], using virtually the same trial data.

Nearly all the meta-analyses mentioned the key features of each trial, such as the screening regimen, management of the control group, compliance with assignment to screening and non-screening groups, as well as the length of follow-up, but few have addressed how these features might affect the mortality reductions. The mortality rate in each arm is commonly calculated simply as the cumulative number of breast cancer deaths divided by the total person-years of follow-up. This is based on the

assumption that deaths in different follow-up years are exchangeable, which is clearly violated according to the first principle in Miettinen and Karp [63, p. 82]:

> The proportional reduction in mortality from the cancer is nothing like a constant over time from the beginning of the screening (for the generally short duration of it) to the end of the follow-up (for an arbitrary duration of it). It thus is logically inadmissible to quantify the reduction by pooling the experience across the entire duration of the follow-up. The proper concern in a trial like this is to address the incidence density of death from the cancer as a function of time since the initiation of the screening. And that function is, of course, different for different durations of the screening.

In Hanley et al. [38], we published a re-examination of five mammography trials among the eight. It was necessary to exclude the Canadian study because the year-specific mortality data are not available from the reports nor obtainable from the authors. The UK AGE trial was also excluded because its participants were much younger than 50 years old, while the Edinburgh study was excluded because of its flawed randomization, evidenced by a substantial difference in the socioeconomic levels between the two arms. The remaining 5 trials differ so greatly in the screening regimens and other features that we did not find it justifiable to meta-analyze them. Instead, we examined the year-by-year pattern of the mortality deficits in each of them separately.

We extracted the year-specific numbers of breast cancer deaths in the screened and unscreened arms from the published articles. (The detailed extraction techniques

are formulated and described in Chapter 7.) From the cumulative numbers of deaths reported in Table 7 in the HIP trial [83] and Table X in the Malmö trial [5], we calculated the yearly numbers of deaths by successive subtractions. The reports of the other three trials contained plots of cumulative numbers of deaths over time (Figure 2 in 88; Figure 2 in 29; Figure 1 in 12). For each of these, we used a graph digitizer to extract the cumulative values, and then converted them into year-specific numbers of deaths, and checked the totals against the total numbers reported in the text. In reports that did not provide sufficiently age-specific data, we used slightly wider or narrower age-at-entry bands. Key features of each trial and the year-specific mortality reductions are summarized in Figure 2–3, a simplified version of Figure 2 in Hanley et al. [38].

Whereas we attempted to identify the nadir achieved in the mortality reduction following the initiation of screening for each trial in Hanley et al. [38], here we merely show how small the year-specific counts of breast cancer deaths are in most of these trials. The methods in Chapter 5 could be straightforwardly extended to combine information across trials, if one wishes to.

2.4 Regarding OSM's editorial

Our manuscript on mammography screening [38] was published by the Canadian Journal of Public Health published in the November/December 2013 issue. In the same issue, Professor Olli S. Miettinen [61] was invited to comment on our piece as an Editorialist. OSM's editorial contained a mix of complimentary and critical remarks. He generously applauds us for respecting two principles of cancer screening, one being that the proportion reduction in cancer mortality are not constant over time and the

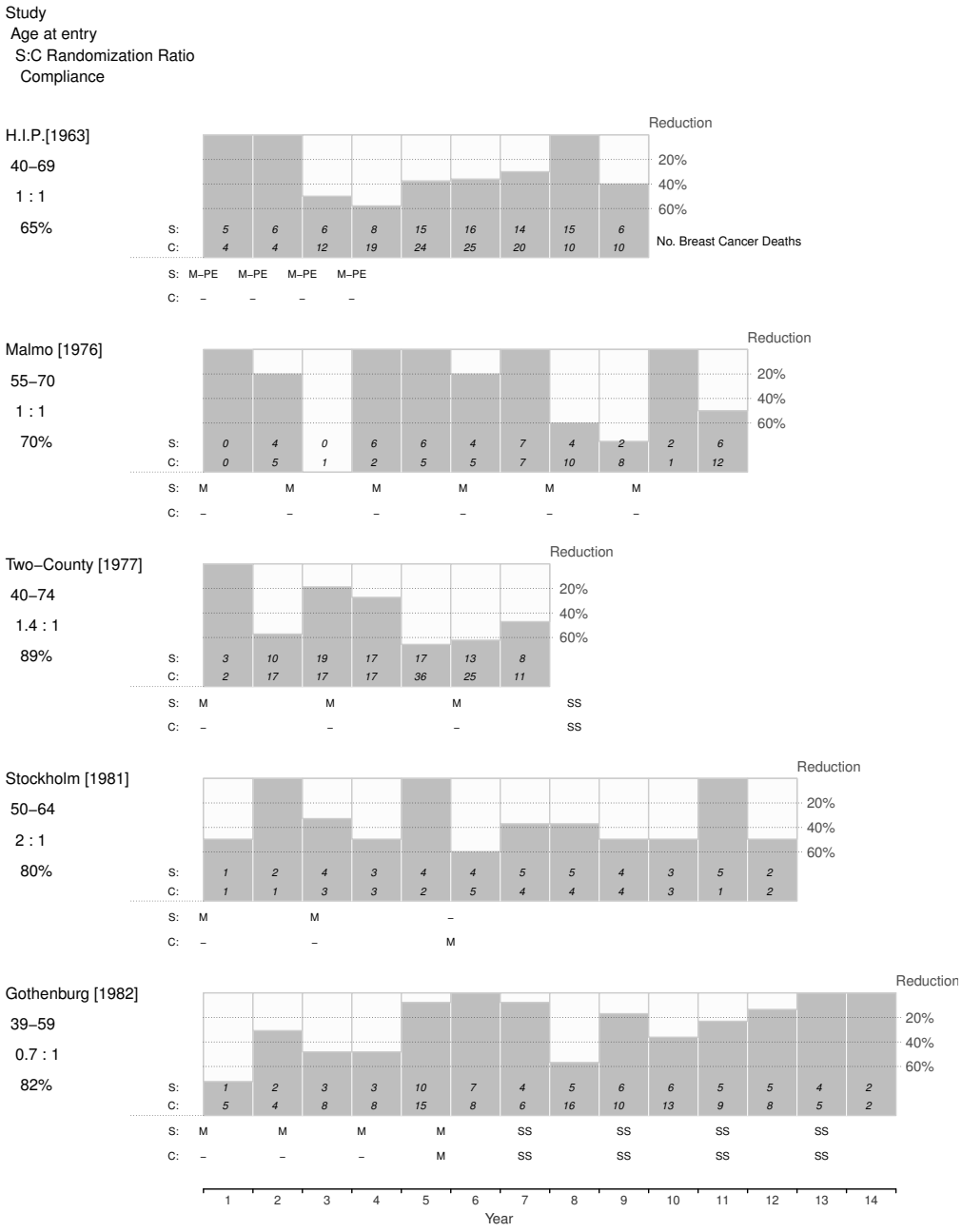


Figure 2–3: The number of rounds of screening, and the approximate timing of each round, together with the yearly numbers of breast cancer deaths in the screening (S) and control (C) arms, and the year-specific mortality reductions.

other being that there is a time lag between the screening's initiation/discontinuation and the manifestation/decline in the magnitude of the mortality reductions. He also appreciates our keen interest in pursuing the asymptotic level of the proportional reduction in mortality from the cancer as the measure of interest.

However there are two items on which OSM seems to disagree with us. First, he argues that our measure, "[t]he proportional reduction in mortality from the cancer in a screening-eligible population", is irrelevant and instead the relevant one should be the "individual benefits from the availability of this service to the women constituting the population at issue". We do not think that there is any contradiction between individual and population level benefits, and OSM himself in this Editorial equates "population-level benefit" with "the sum of the individual benefits". These two are equivalent either when individuals with similar characteristics are considered as exchangeable, or when there is no information on the individual characteristics to distinguish one from another.

Secondly, he argues that our measure is "unrealistic to try to quantify" and instead proposes a "justifiable" measure, in his own words:

> The asymptotic level of the proportional reduction in mortality from the cancer in screening experiments equals something that is critically important to individual women in the population at issue. It equals the proportional reduction in the cancer's rate/probability of incurability, or in its case-fatality rate, attendant to its detection under the screening, when not considering whether the diagnosis is due to the screening or

> to symptoms emerging between two successively scheduled rounds of the screening. [...]
>
> In this individual-centered, clinical-type framework of thought, the population-level benefit – the sum of the individual benefits (cf. above) – from the screening's availability in a given span of calendar time (e.g., the first year of its availability) is, in plain numerical terms (when not accounting for the valuations of the cures), the total number of otherwise incurable cases that, in the population in that period, were cured by screening-afforded early treatments. This is the period-specific number of detections of the cancer consequent to the screening multiplied by three probabilities: the probability of the case being one of a genuine, life-threatening cancer (rather than overdiagnosed as such); the probability of a screen-diagnosed genuine case of the cancer being incurable by treatment delayed to the time when the cancer already would be clinically manifest; and the probability of undelayed treatment upon screen-diagnosis being curative of such an other wise incurable case (i.e., the proportional reduction in incurability addressed above, though adjusted for it to be specific to screen-diagnosed cases).

While OSM's proposed measure (i.e., the probability of benefiting from screening-induced early treatment conditional on having been diagnosed under screening) is perfectly logical, he is addressing a somewhat different estimand from ours. He is concerned with the probability that a women would benefit from early treatment, given that she is already screen-diagnosed. Although directly contrasting benefits of

earlier to later treatment, this measure is not quantifiable in a conventional randomized screening trial and requires an unusual (and perhaps unrealistic, or unethical) design in which screen-diagnosed persons blinded of the result are randomized to either early treatment or delayed treatment when they are later clinically diagnosed.

Our estimand is the probability that a women would benefit from early treatment, given that her cancer is fatal under the usual care. This directly corresponds to the question answered by commonly conducted randomized screening trials. While this is a relative measure, with the denominator being all fatal cancers in the absence of screening, an absolute measure similar to the one characterized by OSM above could be constructed by multiplying the relative measure by the probability of dying due to the cancer in the absence of screening. The absolute measure is also readily estimable as the proportion of cancer deaths in the control arm of a randomized screening trial.

Furthermore, while sometimes the individual-level decision indeed can be about choosing between early versus delayed treatment (e.g. watchful waiting for a detected prostate cancer), most cancers would be treated right away once detected. Thus, a more common individual-level decision is whether to participate in the screening program in the first place. While such a decision requires weighting both the harms and benefits of screening, our proposed measures (the conditional and absolute ones) directly address the benefit side of this decision.

Before ending, OSM contends that we "fail to grasp the true meaning of this measure", and like the Canadian Task Force, we "too do not proceed from tenable, genuinely first principles". In fact, in later chapters, we do derive our measure from

the first principles, including OSM's own 'factor-conditional etiogenetic proportion' [63, p. 48].

CHAPTER 3
Main Source of Inspiration: Understanding Zelen's Work on Screening

Marvin Zelen is a pioneer and leader in the development of statistical models for early-detection programs. His earliest work on screening dates back to 1969 [102], his 1993 paper [101] considers the optimum spacing of screening examinations, and his 1997 paper with Hu [41] served as the basis for the statistical planning of the US National Lung Screening Trial. According to Hu in August 2013, the US National Cancer Institute (NCI) and China NCI were planning, based on the Hu-Zelen model, to launch a randomized screening trial for lung cancer in China in the near future.

In this chapter, we discuss the Hu-Zelen model as a motivation to our work. Although their model construction and its intended purpose are very different from ours, similarly to us, Hu and Zelen [41] took an explicit round-by-round approach to the modeling.

3.1 Background

The arm-specific numbers of cancers detected at and between screening examinations in a randomized screening trial can be used to derive some diagnosis-related measures of the potential benefit of a screening program. Two such quantities are the sensitivity of the examinations, and the (unobservable) length of the sojourn time during which the disease is screen-detectable but asymptomatic, with the premise that if screening is sensitive in detecting the cancers and if the duration of the sojourn time is long, then the earlier detection and earlier treatment of the cancer can

be more successful than the later ones. The statistical estimation of these parameters has been addressed by a number of authors, such as Zelen and Feinleib [102], Albert et al. [3], Walter and Day [96], Day and Walter [22], as well as Shen and Zelen [85, 87].

However, only those cancers that would be fatal in the absence of screening can potentially benefit from being detected and treated early, and thus the diagnosis-related measures are easily confounded by lead time and overdiagnosis, as discussed in Chapter 2. Therefore, a reduction in the cancer-specific mortality in the screening arm compared to the non-screening arm in a randomized screening trial is considered as the definitive evidence of the benefit of screening [96]. Zelen has been a pioneer in the statistical design (e.g. choosing the optimal length of follow-up, and optimal number of and spacing between examinations) of such early-detection trials. With the aim of achieving maximal power of the statistical test for comparing mortality between the screening and non-screening arms, Hu and Zelen [41] developed a model for calculating the cumulative probability of dying of the cancer over a specific follow-up window in the two arms by modelling the entire history of the disease progression.

Built on this, Lee and Zelen [51] developed a more complex model, which takes into account age effect and cancer stage distributions. Together with six other modelling groups within the Cancer Intervention and Surveillance Modeling Network (CISNET) of the US National Cancer Institute, a total of seven models were used to project the impact of screening and treatment on cancer incidence and mortality [16]. All of them involve a very large number of parameters and modelling assumptions, and require multiple data sources (such as trials, registries and surveys) for the

parameter inputs; none of them can produce confidence intervals for the projected mortality impact, and thus the range of the results from these models is taken as a measure of uncertainty [56, 57].

In this chapter, I first show that the Hu-Zelen model can be (i) substantially simplified, (ii) potentially extended to estimate mortality impact using data in randomized screening trials, and (iii) used to project mortality impact of a screening program with a new regimen. Then I discuss the limitations of this model for projecting mortality impact of a screening program and reasons why approaches based on modelling the entire disease progression are not suitable for this purpose. In particular, I argue that this model is oversimplified in parametrizing the screening effect. In the appendix, I provide my programming of the Hu-Zelen model which has been validated by the first author Ping Hu.

3.2 Specification of the Hu-Zelen model

Hu and Zelen [41] formulate a model for calculating the cumulative probability of dying of the cancer before the end of follow-up time τ for an individual randomly assigned to either the control or the screening arm at time $s_0 = 0$. Individuals assigned to the control arm receive standard care with no screening, while those assigned to the screening arm are invited to participate in one or more periodic screening examinations. Let $F_0(\tau)$ and $F_1(\tau)$ be the cumulative probability of dying of the cancer before the end of the follow-up time τ in the control arm and screening arm, respectively. These two probabilities are used to calculate the expected mortality reduction from the cancer due to a specific early detection regimen, in the context of statistical planning of power and sample size of a randomized screening trial. The estimated

risk difference $\hat{F}_0(\tau) - \hat{F}_1(\tau)$ serves as a test statistic for the null hypothesis of no mortality benefit.

The history of the cancer progression is modeled via three state transitions: cancer-free state (S_u) $\rightarrow$ preclinical state (S_p) $\rightarrow$ clinically diagnosed state (S_c). The time interval during which the disease is potentially detectable but not yet diagnosed, is the sojourn time in S_p. It is assumed that the probability density function of the sojourn time $g(t)$ takes an Exponential form, that is, $g(t) = \upsilon \exp(-\upsilon t), \upsilon \geq 0$. Under the assumption of a stable disease model (the probability of transitions both from S_u to S_p and from S_p to S_c are independent of time), the authors showed that the relation between (constant) prevalence ξ, incidence ω, and mean sojourn time $1/\upsilon$ is $\omega = \xi\upsilon$. The average sensitivity of each screening examination is assumed to be ρ.

The same post-diagnosis mortality rate is used for persons clinically diagnosed in the absence of screening and for the interval cancers in the presence of screening. Worth pointing out is that, in order to correct for lead time, the survival time for those diagnosed by one of the scheduled screenings is counted from the *potential* time of clinical diagnosis. This potential time refers to the unobservable time at which the person would have entered the clinically diagnosed disease state in the absence of screening.

For an individual assigned to the control arm, the survival time is measured from the time of clinical diagnosis. Let $g_0^p(t) = \lambda_0^p \exp(-\lambda_0^p t)$, $\lambda_0^p \geq 0$ be the probability density function of survival time for those who are in S_p at randomization, and $g_0(t) = \lambda_0 \exp(-\lambda_0 t)$, $\lambda_0 \geq 0$ for those who are not yet in S_p at randomization.

These two densities are also used for interval cancers for individuals in the screening group who were and who were not in S_p at randomization, respectively.

For an individual assigned to the screening arm, the (now altered) survival time is also measured from the potential time of clinical diagnosis (in order to avoid lead time bias). Let $g_1^p(t) = \lambda_1^p \exp(-\lambda_1^p t)$, $\lambda_1^p \geq 0$ be the probability density function of survival time for those who are in S_p at randomization and detected by a scheduled examination, and $g_1(t) = \lambda_1 \exp(-\lambda_1 t)$, $\lambda_1 \geq 0$ the counterpart for those who are not yet in S_p at randomization but are diagnosed by a scheduled examination. Hu and Zelen [41] show how the risks $F_0(\tau)$ and $F_1(\tau)$ are then obtained through (quite complex) integration of the intensity functions defined in this section; in the following section we present an algebraically simplified version. All the notations used in the Hu-Zelen model are summarized in Table 3–1.

3.3 Simplifying the Hu-Zelen model

In this section, I show how one could algebraically simplify the derivation of the probability of dying due to an interval cancer as specified in the Hu-Zelen model.

Under no screening

For an individual who is assigned to the control arm, as shown in Figure 3–1, there are two mutually exclusive possible paths to the cancer-specific death at time t: (i) she/he is in S_p at time s_0, enters S_c at time u and survives for $t-u$ units, and (ii) she/he is in S_u at s_0, enters S_p at time x ($x > s_0$), has a sojourn time of $u-x$ units in S_p and survives for $t-u$ units. Therefore the probability density function

Table 3–1: A summary table of notations used in the Hu-Zelen model.

Notation	Meaning
S_u	Disease-free or undetectable state.
S_p	Preclinical disease state.
S_c	Clinically diagnosed disease state.
ρ	Sensitivity of the screening test.
ω	Transition rate from S_u to S_p and from S_p to S_c (cancer incidence)
ξ	Prevalence of pre-clinical disease.
s_0	Time of randomization.
τ	End of the follow-up period.
$s_1, \ldots, s_m$	Ordered times of m scheduled screening examinations.
$f_0(t)$	Probability density function (PDF) of dying of the cancer at time t in the control arm.
$f_0^p(t)$	Subset of $f_0(t)$: is in S_p at s_0.
$f_0^u(t)$	Subset of $f_0(t)$: is in S_u at s_0.
$f_1(t)$	PDF of dying of the cancer at time t in the screening arm.
$f_1^T(t)$	Component of $f_1(t)$: screen-diagnosed and treated earlier.
$f_1^I(t)$	Component of $f_1(t)$: an interval cancer.
$f_1^N(t)$	Under the null version of $f_1^T(t)$ when there is no benefit from early treatment.
$F_0(\tau)$	Cumulative probability of dying of the cancer before τ in the control arm.
$F_1(\tau)$	Cumulative probability of dying of the cancer before τ in the screening arm.
$g(t)$	PDF of the sojourn time in S_p.
$g_0^p(t)$	PDF of survival time in the control arm, in S_p at s_0.
$g_0(t)$	PDF of survival time in the control arm, not in S_p at s_0.
$g_1^p(t)$	PDF of survival time in the screening arm, in S_p at s_0.
$g_1(t)$	PDF of survival time in the screening arm, not in S_p at s_0.
υ	Rate parameter of the sojourn time distribution $g(t)$.
λ_0^p	Rate parameter of the survival time distribution $g_0^p(t)$.
λ_0	Rate parameter of the survival time distribution $g_0(t)$.
λ_1^p	Rate parameter of the survival time distribution $g_1^p(t)$.
λ_1	Rate parameter of the survival time distribution $g_1(t)$.

of dying of the cancer at time t in the control arm is $f_0(t) = f_0^p(t) + f_0^u(t)$, where

$$\begin{aligned} f_0^p(t) &= \int_0^t \xi g(u) g_0^p(t-u)\,\mathrm{d}u \\ &= \omega\lambda_0^p \exp(-\lambda_0^p t)\frac{1-\exp\{-(\omega/\xi-\lambda_0^p)t\}}{\omega/\xi-\lambda_0^p}, \end{aligned}$$

and

$$\begin{aligned} f_0^u(t) &= \int_0^t \left\{\int_0^u \omega g(u-x)\,\mathrm{d}x\right\} g_0(t-u)\,\mathrm{d}u \\ &= \omega\lambda_0 \exp(-\lambda_0 t)\left\{\frac{\exp(\lambda_0 t)-1}{\lambda_0} + \frac{\exp(-\omega/\xi-\lambda_0)t-1}{\omega/\xi-\lambda_0}\right\}. \end{aligned}$$

The closed form expressions are simplifications obtained by substituting in the previously specified parametric models; we suppress the intermediate steps. The risk in the control arm is then given by $F_0(t) = \int_0^t f_0(s)\,\mathrm{d}s$.

Under screening

During the interval $[0, \tau]$, a total of m screening examinations are carried out at the ordered time points $s_1 < s_2 < \cdots < s_m$ in the screening arm, with the jth interval denoted by $[s_{j-1}, s_j]$ and its length by $\Delta_j = s_j - s_{j-1}$ for $j = 1, 2, \ldots, m$. For any individual assigned to the screening arm, there are further two mutually exclusive paths to the cancer-specific death at time t: (a) she/he is detected and diagnosed by the rth planned examination at time t_r, $r > 0$, where $t > t_r$, with the corresponding probability density function for survival time denoted by $f_1^T(\tau)$, and (b) she/he is not detected by one of the scheduled screenings (i.e. an interval cancer), with the corresponding density denoted by $f_1^I(t)$. The overall density in the screening arm is then $f_1(t) = f_1^T(\tau) + f_1^I(\tau)$.

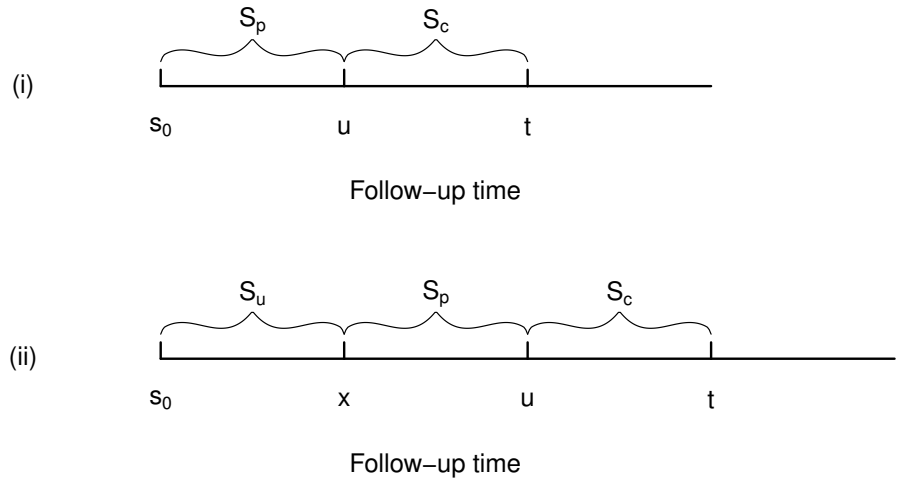


Figure 3–1: Schematic figure showing the two mutually exclusive paths to cancer-specific death in the control arm: at randomization the individual is preclinical in (i) and is disease-free in (ii).

An individual of type (i) in the preclinical state at randomization and following path (a) in the screening arm must have failed to be detected at $r-1$ previous examinations (each with probability $1-\rho$ due to imperfect sensitivity of the screening test), with the corresponding density denoted by $f_{1r}^{T1}(t)$. An individual of type (ii) disease-free at randomization, following path (a), and moving into the preclinical state in the interval $[s_{i-1}, s_i)$, $i \in 1, \ldots, r$, must have failed to be detected at $r-i$ previous examinations, with the corresponding density denoted by $f_{1r}^{T2}(t)$. Thus the density corresponding to path (a) can be written as $f_{1r}^{T}(t) = f_{1r}^{T1}(t) + f_{1r}^{T2}(t)$, where

$$\begin{aligned} f_{1r}^{T1}(t) &= (1-\rho)^{r-1}\rho\xi \int_{s_r}^{t} g(u) g_1(t-u)\,\mathrm{d}u \\ &= \frac{(1-\rho)^{r-1}\rho\xi\omega\lambda_1^p \exp(-\lambda_1^p t)}{\lambda_1\xi - \omega} \left\{\exp(\lambda_1 t - \omega t/\xi) - \exp(\lambda_1 s_r - \omega s_r/\xi)\right\}, \end{aligned}$$

and

$$f_{1r}^{T2}(t) = \sum_{i=1}^{r}(1-\rho)^{r-i}\rho\int_{s_r}^{t}\left\{\int_{s_r}^{s_i}\omega g(\tau-x)\right\}g_1(t-u)\,\mathrm{d}u, (t >= s_r)$$
$$= \sum_{i=1}^{r}\frac{(1-\rho)^{r-i}\rho\xi\omega\lambda_1\exp(-\lambda_1 t)}{\lambda_1\xi-\omega}\{\exp(\omega s_i/\xi)-\exp(\omega t_{i-1}/\xi)\}$$
$$\times\{\exp(\lambda_1 t-\omega t/\xi)-\exp(\lambda_1 s_r-\omega s_r/\xi)\}.$$

Since any individual can be detected in only one screening examination, $f_1^T(t) = \sum_{r=1}^{m} f_{1r}^T(t)$.

For an individual following path (b), Hu and Zelen [41] spent an entire page of algebra deriving the density $f_1^I(t)$ for time of death due to an interval cancer through state transitions. However, we argue that this can be inferred from the quantities that are already specified. Under the null, if there is no difference in the post-diagnosis survival time between the two arms, that is, $g_0(t) = g_1(t)$ and $g_0^p(t) = g_1^p(t)$, then the densities of dying of the cancer in the two arms would be identical: $f_0(t) = f_1(t) = f_1^N(t) + f_1^I(t)$, where $f_1^N(t)$ represents the density of dying of the cancer after being screen-detected but not treated early (or equivalently, not benefiting from the early treatment). In turn, $f_1^N(t)$ is obtained from $f_1^T(t)$ as defined above by replacing $g_1(t)$ with $g_0(t)$ and $g_1^p(t)$ with $g_0^p(t)$.

Therefore $f_1^I(t) = f_0(t) - f_1^N(t) = f_0(t) - \sum_{r=1}^{m} f_{1r}^N(t)$, where

$$
\begin{aligned}
f_{1r}^N(t) &= (1-\rho)^{r-1}\rho\xi \int_{s_r}^{t} g(u) g_0^p(t-u)\,\mathrm{d}u \\
&\quad + \sum_{i=1}^{r}(1-\rho)^{r-i}\rho \int_{s_r}^{t} \left\{ \int_{s_r}^{s_i} \omega g(\tau - x) \right\} g_0(t-u)\,\mathrm{d}u, (t >= s_r) \\
&= \frac{(1-\rho)^{r-1}\rho\xi\omega\lambda_0^p \exp(-\lambda_0^p t)}{\lambda_1^p\xi - \omega} \left\{ \exp(\lambda_0 t - \omega t/\xi) - \exp(\lambda_0 s_r - \omega s_r/\xi) \right\} \\
&\quad + \sum_{i=1}^{r} \frac{(1-\rho)^{r-i}\rho\xi\omega\lambda_0 \exp(-\lambda_0 t)}{\lambda_0\xi - \omega} \{ \exp(\omega s_i/\xi) - \exp(\omega t_{i-1}/\xi) \} \\
&\quad \times \{ \exp(\lambda_0 t - \omega t/\xi) - \exp(\lambda_0^p s_r - \omega s_r/\xi) \}.
\end{aligned}
$$

Finally, having obtained $f_1(t) = f_1^T(t) + f_0(t) - f_1^N(t)$, the risk in the screening arm is given by $F_1(t) = \int_0^t f_1(s)\,\mathrm{d}s$.

3.4 Using the Hu-Zelen model to estimate mortality impact

The Hu-Zelen model has been used for statistical planning of randomized screening trials with fixed parameter inputs, requiring knowledge of the mortality rates for persons diagnosed in the absence and in the presence of screening, in addition to the parameters characterizing cancer incidence, prevalence, sojourn time and screening sensitivity. However, since this is a probability model, in principle it could be used as a likelihood for the purpose of estimating (at least some) of these parameters. Suppose for simplicity that the model can be parametrized in terms of baseline rate $\lambda_0 = \lambda_0^p$ and mortality rate ratio $\theta = \lambda_1/\lambda_0$, where $\lambda_1 = \lambda_1^p$. In this section, we show that the Hu-Zelen model can be used for maximum likelihood-based estimation of θ characterizing the effect of screening-induced early treatments, if the other parameter inputs can be fixed based in external information.

Let N_0 and N_1 denote the numbers of individuals randomized to the control and screening arms, respectively, indexed by $i = 1, 2, \ldots, N_0, N_0 + 1, \ldots, N_0 + N_1$. Further, let t_i denote the observed time of death due to the cancer or censoring (due to the end of the follow-up or death due to another cause), and let e_i denote the observed event type at T_i, taking the value 1 for a cancer-specific death and 0 for censoring.

Then the log-likelihood from the two arms can be written as

$$\log L(\lambda_0, \theta) = \log L_0(\lambda_0) + \log L_1(\lambda_0, \theta),$$

where the log-likelihood contributions from participants in the control arm is

$$\log L_0(\lambda_0) = \sum_{i=1}^{N_0} e_i \log f_0(t_i) + \sum_{i=1}^{N_0} (1 - e_i) \log\{1 - F_0(t_i)\},$$

and from those in the screening arm is

$$\log L_1(\lambda_0, \theta) = \sum_{i=N_0+1}^{N_0+N_1} e_i \log f_1(t_i) + \sum_{i=N_0+1}^{N_0+N_1} (1 - e_i) \log\{1 - F_1(t_i)\}.$$

Here $f_0(t)$, $f_1(t)$, $F_0(t)$ and $F_1(t)$ are specified as in the previous section. However, the parameters ρ, ω and ξ or υ would have to be fixed to a priori specified values, since they are not estimable from the mortality outcome data alone; this limits the usefulness of the Hu-Zelen model in estimation of the mortality impact of screening, and as we argue in the following sections, the modeling assumptions involved might in any case result in overly simplistic characterization of the screening impact.

3.5 Using the Hu-Zelen model to project mortality impact

In this section, we show how to use the Hu and Zelen model to project the mortality reduction patterns if women begin a 20-year breast cancer screening program when they reach age 50. Since the Health Insurance Plan (HIP) study [83] is the only trial which has the required data and parameter inputs publicly available, we use it for the following depiction. In this study, women aged 40-64 years were randomized to attend 4 annual mammography screenings or continue with their usual care without screening.

Shen and Zelen [86] estimated the overall screening sensitivity and mean sojourn time to be $\rho = 70\%$ and $1/\upsilon = 2.5$ years. We take the incidence rate in the absence of screening to be $\omega = 0.0187$, the incidence in the control group reported in the HIP trial. Under the stable-disease assumption, the prevalence at age 50 is calculated to be $\xi = 0.0187 \times 2.5 = 0.47\%$. Hu and Zelen [41] assumed that the median survival times after clinical diagnosis under no screening and screening for those not in S_p at randomization to be $\log(2)/\lambda_0 = 10$ and $\log(2)/\lambda_1 = 17$ years, respectively. For those in S_p at randomization, the median survival times are assumed to be $\log(2)/\lambda_0^p = 11$ and $\log(2)/\lambda_1^p = 20$ years under no screening and screening, respectively.

Figure 3–2 shows the resulting 30-year projection for a program of 20 annual screenings. The early portion of the mortally reduction curve is not realistic, as there is a large reduction immediately after the initiation of screening; not surprisingly the curve stays constant in the middle part where there was sustained screening and gradually tails off after screening was discontinued.

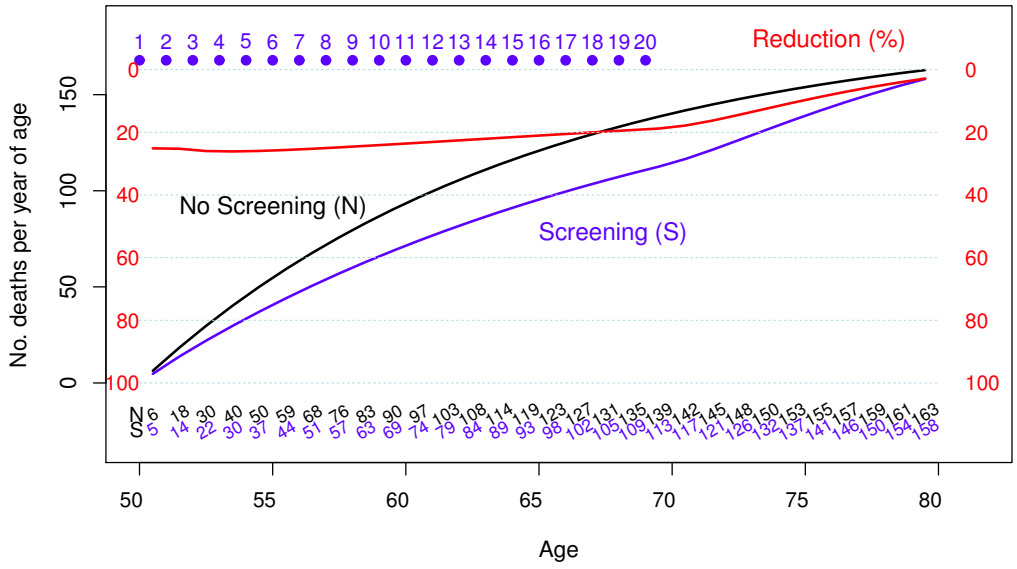


Figure 3–2: Projected yearly numbers of breast cancer deaths, in each of the ages 50 to 80, if 100,000 women did versus did not participate in a 20-year program of annual mammography screening starting when they reach age 50, together with the corresponding percentage reductions.

Figure 3–3 shows the projection for a biennial program; it is a little shallower than the annual one, but the reductions persist for almost the same duration. The comparison between Figures 3–2 and 3–3 matches with the findings of the CISNET models [56], which reported a median of 16.5% and 20.4% reduction in breast cancer deaths with biennial and annual regimen, respectively, for women invited to mammography screening from age 50 until 69.

3.6 Discussion

Given the unobservable nature of the cancer progression, projecting the mortality impact of a screening program by modelling the entire history of the disease has

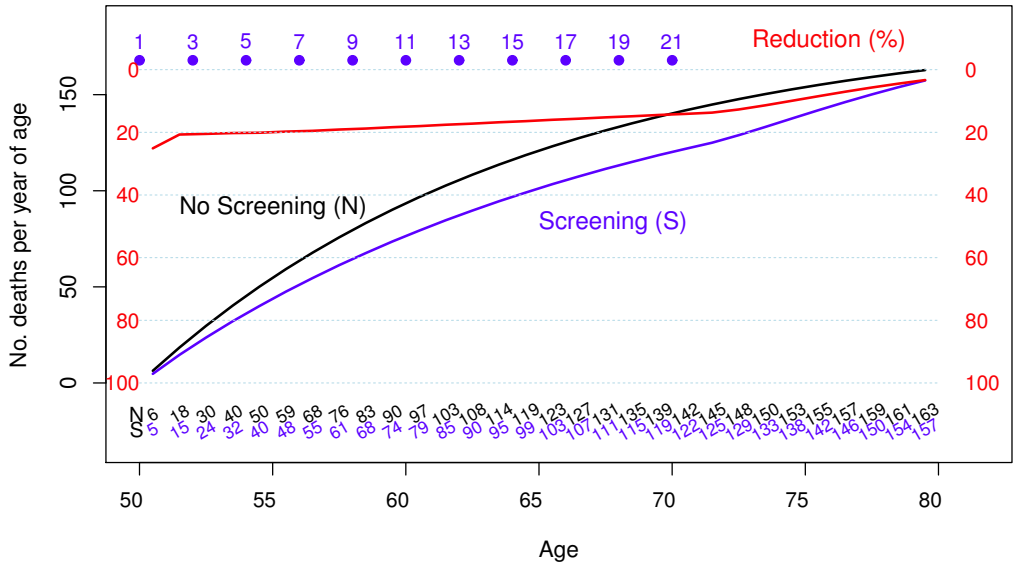


Figure 3–3: Same as in Figure 3–2, but with biennial screening.

a number of disadvantages: (i) a very large number of parameters are involved and they cannot be estimated simultaneously (in fact, only the joint estimation of the sensitivity and sojourn time has been achieved), (ii) it requires a large number of modelling assumptions; (iii) it cannot produce confidence intervals for the projected mortality impact, and (iv) parameter inputs have to be obtained from multiple data sources (such as trials, registries and surveys) but inputs from one source of a specific population in a specific study setting may not be applicable to other contexts.

In addition to the above limitations, there are other reasons why the Hu-Zelen model might not be appropriate for projecting mortality impact of a screening program. The first has to do with the Exponential form assumed for the sojourn time

distribution. As discussed in Liu et al. [53], a fundamental feature of cancer screening is its delayed effect in mortality reduction in asymptomatic individuals. That is, cancer deaths averted by screening combined with therapy would only manifest several years after the onset of screening. The first screening examination combined with therapy detects and eradicates some cancers that otherwise would have proven fatal several years later. The reductions produced by subsequent examinations occur even later. The Exponential model is used for the sojourn time distribution for computational convenience; however, it does not take into account the time lag between the screenings and their induced mortality reductions. That is why the early portion of the projected reduction curve failed to show the delayed effect, as illustrated in Figures 3–2 and 3–3. A possible solution would be to use a distribution which has a mode, such as the Weibull distribution, but this would add more parameters to the already complex model.

Furthermore, it is assumed in the Hu-Zelen model that the sojourn time is independent of the survival time after clinical diagnosis. This assumption is not a realistic one either; instead we would expect a strong correlation between the two: a relatively fast-growing cancer would be aggressive both pre- and post-detection. One way to introduce dependence between the post-diagnosis survival time and sojourn time would be to parametrize the mortality rate $\lambda_0 e^{\alpha Z+\beta(u-x)}$, where $(u-x)$ is the length of the sojourn time and Z indicates whether or not the individual is assigned to the screening arm. This would again add more parameters to the model. Finally, it should be noted that modelling in terms of mortality rates and their ratios does

not characterize cures, but rather delays in the inevitable time of death due to the cancer.

In summary, by studying the Hu-Zelen model, we demonstrate why the prevailing approach to project the impact of screening by modelling the entire cancer progression, such as those used in the CISNET models, has limitations. This motivates us to pursue a completely different approach, that is, focusing on mortality alone and avoid handling sojourn time, the mechanism of screen detection, tumour characteristics at diagnosis and so on. In Chapter 5, we present a novel probability model which specifically takes into account the delay in mortality reductions, and directly addresses what we consider the most relevant question, namely, the probability of being helped by screening-induced early treatment, had the cancer proved fatal in the absence of screening. The mortality impact is estimated using data from randomized screening trials and thus the probabilistic projection is evidence-based. In this minimalist model, we avoid specifying parameters such as sensitivity, prevalence, incidence and sojourn time all together, as well as modelling assumptions associated with each one of them.

CHAPTER 4
Projecting the Yearly Mortality Reductions due to a Cancer Screening Program

Preamble to Manuscript 1. In this manuscript we argue that the hypothesis-testing framework used in trials has limited trialists' attention to summary statistics that do not adequately describe what a sustained screening program might accomplish.

Instead of a single-number rate ratio, we proposed a time-specific rate ratio curve as the measure to address the benefits of a screening program. This bathtub-shaped curve shows when the reductions begin, how big they are, and how long they last. We illustrate how one can compute such curves using an existing model.

In the second manuscript, we develop a novel probability model to directly address the mortality impact of a screening program. But before we do so, it is important to motivate why a new approach is needed.

Projecting the Yearly Mortality Reductions due to a Cancer Screening Program

Zhihui (Amy) Liu, James A Hanley, and Erin C Strumpf

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, Quebec H3A 1A2, Canada.

Paper published in September 2013 in *Journal of Medical Screening*, 20:157–164.

The final published version of this manuscript may be found online at `http://msc.sagepub.com/content/20/3/157` (DOI: 10.1177/0969141313504088).

Abstract

Whether or not to implement a 20-year screening program for a cancer requires weighing the harms and costs against the health benefits (such as the number of cancer deaths averted every year). The evidence of the benefits is often based on a single-number summary, such as the mortality reduction over the entire follow-up time in a single trial, or an average of such one-number measures from a meta-analysis of several trials. We strongly recommend against using the traditional one-number summaries from trials to deduce the yearly mortality reductions expected from a sustained screening program. Instead, we propose using a rate ratio *curve*, and its complement (a mortality reduction curve), to address the mortality impact (timing, magnitude, and duration) of a screening program. This curve is easy to interpret, as it shows when mortality reductions begin, how big they are, and how long they last. We illustrate when and how one could compute such rate ratio curves from screening trials, and use them to compare reduction patterns expected with different screening regimens. We call on trialists to report necessary data to arrive at such projections.

4.1 Introduction

Making a decision on whether or not to implement a 20-year screening program for a cancer requires weighing the harms and costs against the health benefits (such as the number of cancer deaths averted every year). The evidence of the benefits is often based on a single-number summary, such as the mortality reduction over the entire follow-up time in a single trial, or an average of such one-number measures from several trials. We recommend against using such one-number summaries to deduce the yearly mortality reductions expected from a sustained screening program.

As we explain more fully below, we base this recommendation on several reasons, all stemming from the characteristic time-pattern of the mortality reductions produced by any particular screening program, and the affected time-window in question. First, the reductions do not begin in year one, and if/when they do reach a 'constant' level, they do not remain at this level indefinitely. Thus the full pattern (i.e. the timing, magnitude and duration of the reductions) cannot be adequately quantified by one number. In addition, the pattern is specific to the screening regimen (e.g. the number of screens and spacing between them) employed. For example, 20 annual screens might produce yearly reductions that start at year 5, and extend over possibly 25 years; 10 annual rounds would produce similar yearly reductions starting at about the same time, but extending over a shorter span, possibly 15 years. Compared to 10 annual screenings, the yearly reductions produced by 10 biennial screenings are expected to be smaller but over a longer period of time.

In this paper, we address the task of projecting the mortality impact of a screening program. In Section 2, we propose using a rate ratio curve, instead of a single-number summary, to fully describe the expected timing, magnitude and duration of this impact. In Section 3, we identify trials that have had sufficient rounds of screening allowing us to estimate the asymptote of the curve for a program with similar spacing of screens. We also give examples to illustrate how much underestimation is involved in the traditional measure. In Section 4, we show how one could use an existing model (previously used for other purposes), and available trial data to project the program impact, as well as compare reduction patterns produced by regimens with different spacings than those used in trials. Finally, we call on trialists to report necessary data to compute this rate ratio curve.

4.2 The mortality reduction curve, and its shape

4.2.1 The time lag and the affected age window

Consider a cohort of persons who, beginning at age 50, are invited to be screened annually for a cancer until they reach age 69. The mortality impact of the program is the difference in the yearly number of cancer deaths in the absence of screening (when no one is invited to be screened) versus in the presence of screening (when everyone is). We graph this impact in the affected age window in Figure 4–1(a).

The first notable feature is the time lag between when a screening program starts and when the mortality reduction first manifests. Unlike most medical interventions that produce a virtually immediate effect (within hours, days or weeks), cancer screening generates mortality reductions that only become evident several years after the onset of screening [67, 62, 37, 10, 83]. The first screen (say at age 50) detects,

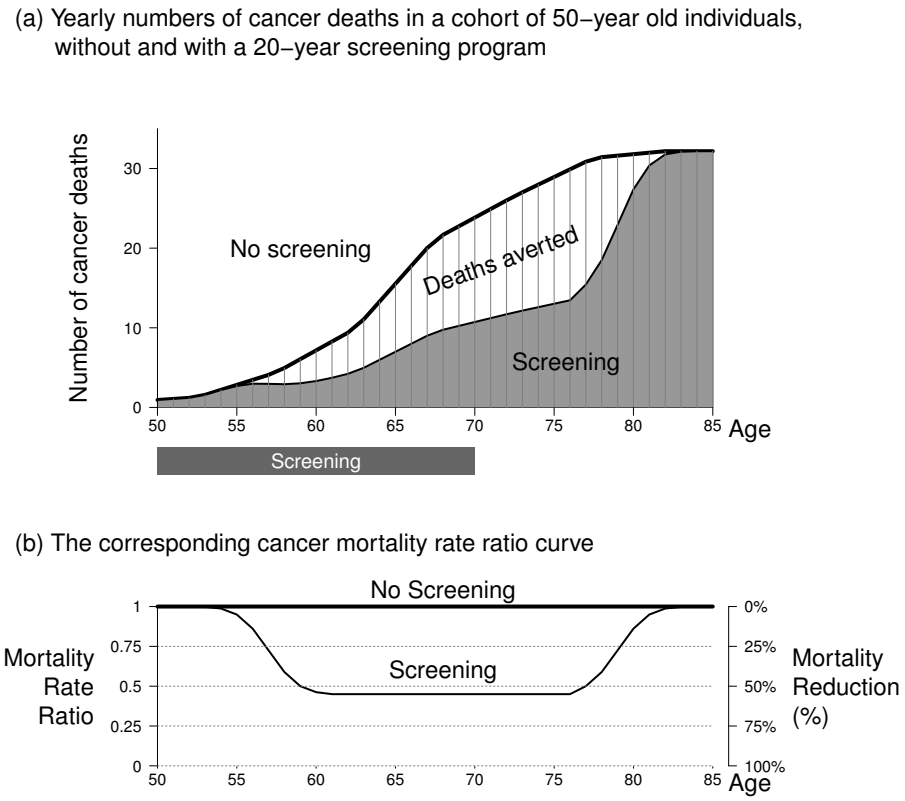


Figure 4–1: Impact of a hypothetical 20-year screening program measured (a) in absolute numbers of cancer-specific deaths averted and (b) as rate ratios and as percentage reductions.

and the resulting earlier therapy eradicates, some cancers that otherwise would have proved fatal several years later (from say 55 to 63). Presumably, the average delay would be longer for cancers of the breast and prostate, and shorter for more aggressive cancers, such as that of the lung. The width of the reduction 'wave' (8 years in our example) reflects the variation in cancer stages at detection and in the rates at which cancers would have progressed otherwise.

Mortality reductions produced by subsequent annual screens (at ages 51, 52, ..., 69) occur even later (from say 56 to 64, 57 to 65, ..., 74 to 82). After the effect of the last screen disappears, cancer mortality rates return gradually (from say 78 to 85) to those in the absence of screening. Thus, the 20 screens affect possibly 35 age-bins in the age-span 50 to 85.

The total number of deaths averted in that span is shown as the white area in Figure 4–1(a). The total number of years gained is the sum of the products of the age-specific number of deaths averted and the age-specific remaining life expectancies. For costing purposes, this total can be averaged over the number averted, invited, or screened.

4.2.2 The mortality rate ratio curve

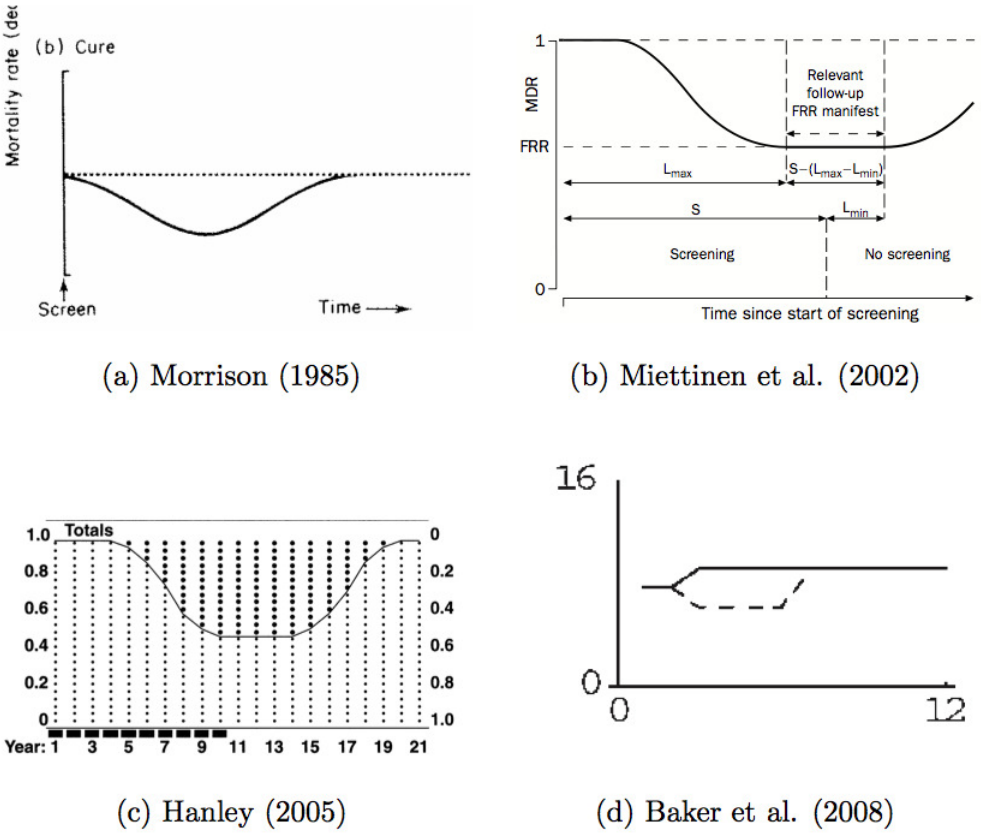
Another way to display the same mortality reductions in Figure 4–1(a) is through a rate ratio curve, as in Figure 4–1(b). The yearly ratio is calculated as the yearly number (or rate) of cancer deaths in the presence of screening divided by the yearly number (or rate) of cancer deaths in the absence of screening. Each yearly ratio can be thought of as the fraction of fatal cancers that could not be helped by screening. Their complements, usually expressed as percentages, represent the yearly mortality reductions.

If the yearly number of fatal cancers remains constant throughout the screening program, the rate ratio curve should exhibit a bathtub shape: it would be close to constant for a large portion of the age-window where the effect of sustained screening is manifest. Little mortality impact is expected in the early portion, i.e. before the deaths averted by the first screen would have otherwise occurred, and again in the

late portion, i.e. long after the deaths averted by the last screen would have otherwise occurred. By describing the timing, magnitude and duration of the yearly reductions over the full time window that would be affected by a screening program, the curve shows when reductions begin, how big they are, and how long they last.

The rate-ratio curve in Figure 4–1(b) is not new: Morrison [67] introduced a schematic version, entitled "changes in the disease-specific mortality rate", to graphically illustrate and emphasize the time lag between the first screen and the beginning and end of the mortality reductions. Early trialists [83] were also keenly aware of the waning effect after the termination of screening. A more comprehensive version, showing what affects the shape, is presented in a theoretical piece by Miettinen et al. [62], and then in an application to mammography with the asymptote as the 'estimand'. Hanley [35] showed how a rate ratio curve could arise as the convolution of the effects of 10 annual rounds of screening, and also studied the asymptote in colon cancer screening; Baker et al. [10] simulated rate ratio curves under screening of large, moderate and little effect. These four versions are shown in Figure 4–2.

Much of the statistical work that has addressed this non-proportional hazards time pattern has focused on statistical tests applied to data from screening trials, and thus on maximizing statistical power [103, 82] dealing with the non-proportionality [81], and selecting the optimal time at which the analysis of trial data should be carried out [41]. The data analysis in each actual trial tested a regimen-specific null hypothesis over some (un-predetermined) follow-up period: "does the amount and spacing of screening used in this trial have a non-zero impact on cancer mortality?"



(a) Morrison (1985)

(b) Miettinen et al. (2002)

(c) Hanley (2005)

(d) Baker et al. (2008)

Figure 4–2: Hypothetical rate-ratio curves, as depicted in textbooks and other publications. (a), (b) and (d) invoke the bathtub shape, while (c) derives it from the convolution of the separate effects of 10 annual rounds of screening.

There has been much less focus on deducing the impact of a sustained screening program.

4.3 Distinction between nadir in a trial and asymptote in a program

4.3.1 Trial nadir and program asymptote

Our focus is on identifying the *asymptote* of the rate ratio curve, since it represents the sustained reduction that could be expected from a screening program. In

the following, we describe how one can - but only in some instances - estimate the program asymptote from trial data.

Figure 4–3 shows the distinctive patterns produced by a trial of 3 annual screenings versus by a program of 20 annual screenings. If each round of screening reduces mortality over 5 future years, then three rounds would produce 3 waves of such reductions. The affected time window spans over a total of 7 years, with a maximum reduction of 35% in year 6. In contrast, a program of 20 screenings would produce 20 such waves, affecting many more years, with a sustained reduction of 46% for 16 years, much deeper and longer than the width and the maximum depth of the reductions seen in a trial. As is seen by comparing panels (a) and (b), the nadir seen in a trial usually underestimates the asymptote in a program. However, even if one wished to just measure the nadir carefully by, for example, smoothing [38] to avoid overestimation resulting from the yearly statistical fluctuations, they would find that few trials have provided yearly data that might allow them to do so. Instead, the universal practice is to report an averaged reduction, computed over the entire follow-up time of the trial. Since this average includes the almost-zero reductions outside the affected time window, it is even smaller than the nadir, and thus an even greater underestimate of the program asymptote of interest.

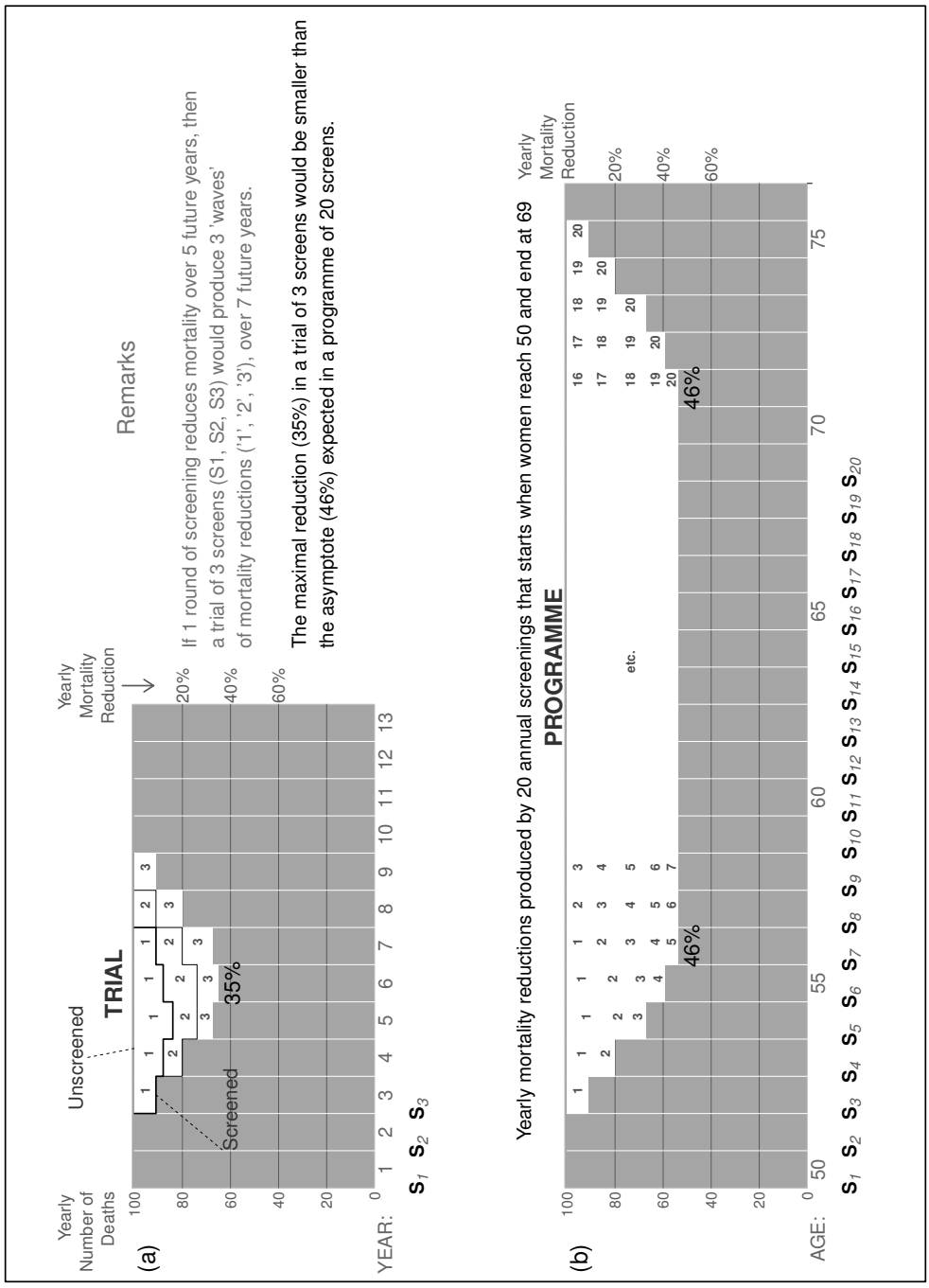


Figure 4–3: Schematic figure showing the mortality patterns of a trial and of a program (see full caption on the next page).

Figure 4–3 caption: The 35% maximal mortality reduction produced by a (hypothetical) trial of 3 annual screenings (a) does not necessarily reach the 46% asymptote produced by a program of 20 annual screenings (b), particularly if the impact of each round is spread over more than 3 years. Shown in (a) is a hypothetical *trial* of 3 annual rounds of cancer screening (S_1, S_2, S_3) compared with no screening. The depth of the white rectangle in each year represents the percentage mortality reduction, relative to an unscreened group, for the year shown on the horizontal axis. Annual mortality reductions produced by screening only begin to be expressed, in year three (when the first effect of S_1 is discernible); they are greater in years 4 and 5, reaching a maximum of 35% in year 6 (when the combined effect of S_1, S_2 and S_3, denoted by '1' , '2' and '3 respectively, is maximal); in year 7 the combined effects begin to wear off, and the mortality in the screening arm begins to revert to that in the non-screening arm; in year 9, the last effect of S_3 is discernible. Thus the maximum reduction is 35% and it would have been greater than if screening had not been discontinued at year three. By contrast the *average* effect of screening over the 13 years of observation (the metric used by task forces) would be 12%. Shown in (b) is a hypothetical screening *program* with annual screening beginning at age 50 and continuing until age 69, compared with no screening. Again, the depth of the white rectangle represents the percentage mortality reduction for the age shown on the horizontal axis. The mortality reduction reaches 46% at age 56 and is maintained at that level for many age-bins – until three years after the last screen when it starts to increase again.

The report of the National Lung Screening Trial (NLST) [73] presented in Table 4–1, illustrates the difference between evidence based on a few screenings which produce *some* reductions in lung cancer mortality over a short time-window, and the level of data needed to project what would occur if 50-year-old people were offered regular screenings until they reached age 69. The deficit of 88 deaths in part (a) of the table is clearly statistically significant, and expectedly shows that 3 CT screenings would reduce lung cancer mortality by some non-zero amount. But the pattern of the yearly deficits in part (b) is incomplete and puzzling. If the 42% deficit in year 6 were to be followed by two similarly large deficits in years 7 and 8, then it would suggest that a screening program could achieve an asymptote twice the size of the reported 20% reduction. If instead the deficit in year 6 were to be followed by diminishingly small deficits of the sizes seen in years 1-5, it would suggest that the deficit in year 6 was merely a statistical aberration, and that the asymptote in a program would be much smaller than the reported 20%.

The additional numbers of cancer deaths in years 7 and 8 were unknown at the time of the report, because the causes of the deaths that occurred in these latter years had not all been adjudicated by the time the overall mortality reduction became statistically significant. This is a striking example of the distinction between getting a statistical significant result with just 3 screens, and providing evidence on what a screening program (of possibly many more screens) would achieve.

The importance of using time-specific rates to pursue the asymptote of the curve was also highlighted in a recent review of screening trials in colon and prostate cancer. Whereas the overall reduction in the largest colon trial been reported to be 20%, the

re-analysis, which took account of the timing of, and interruptions in, screening, found that an uninterrupted program would yield reductions with an asymptote of 40% [35]. In screening trials for prostate cancer, where the time lag between screening and when the mortality deficits manifest are even longer, the deficits produced by the first screen would not be expected for at least six years; however the majority of the follow-up has only extended to about year 11 in the European Randomized Study of Screening for Prostate Cancer (ERSPC) [80]. A re-analysis [36] showed that the reductions only began in year 7, and reached an asymptote of approximately 50% by year 12. One commentator [50] put it well: "perhaps a better summary of the European trial result is not the 20% overall reduction in prostate cancer mortality, but the combination of no reduction in the first seven or so years and a reduction of about 50% after 10 years".

Several task forces have examined screening *programs* for breast, lung, colon and prostate cancers. Although their stated purpose was to estimate what a sustained program would do, all of the meta-analyses they used merely averaged the *overall* reductions seen in different trials. Thus they all greatly underestimated the asymptotes that would characterize the programs they considered [37].

A few authors have explicitly dealt with the delay, either by using the hazard ratio from a certain time point onwards [14], or (in those trials with a sufficiently long duration of screening), by 'letting the data speak for themselves' as to when the asymptote begins [62, 80].

4.3.2 An alternative metric

An alternative approach, that indirectly addresses the asymptote and directly acknowledges the time-pattern of the reductions produced by a limited number of rounds of screening, is to examine the mortality impact *only in cancers diagnosed during the screening period*. This avoids the dilution, which Baker et al. [10] refers to as "post screening noise", described above: cancers that arise long after the screening is discontinued could not have been affected by the screening carried out in the trial. In one version [89] of this alternative approach, where the cumulative incidence of cancers deaths – in those diagnosed in this screening period – in the two study arms are compared, it is assumed that there is no over-diagnosis in the screening arm. The other version [99] avoids having to make this assumption by using the number of cancers that were diagnosed in the non-screening arm during the screening period. The efficacy of the 3 rounds of CT screening is then determined by calculating the 'deficit' of (442-354 =) 88 cancer deaths, and expressing this 88 as a percentage, not of 442, but of the number that *could possibly have been helped* by screening (the 88 who were, and the xxx whose cancers, despite being diagnosed in the screening period in the screening arm, proved fatal nevertheless). Unfortunately, as of the time of writing, this number xxx is not known. Were the data to derive it reported, one could use it as a rough proxy for the asymptote of interest.

Since the approaches described above do not allow one to make projections for a program that uses a *different* spacing of screening examinations that was used in a

Table 4–1: Numbers of lung cancer deaths in the NLST report.

(a) What was reported in NEJM (August 4, 2011)

Follow-up Year:	1	2	3	4	5	6	7	8	ALL
Screens	↑	↑	↑						
X-ray Arm:									442
CT Arm:									354
Reduction:									20%

(b) Year-specific data extracted from graph in that report

X-ray Arm:	37	68	82	95	84	73	4	?	
CT Arm:	31	57	67	84	72	42	3	?	
Reduction:	16%	16%	18%	12%	14%	42%	?	?	

trial, we now describe some (necessarily-model-based) ones that do. This round-by-round approach also allows one to deal with trials whose nadir may not have reached the asymptote.

4.4 Projecting the reduction patterns that would be produced by different regimens than those used in trials

4.4.1 Approaches

Since a trial usually does not contain sufficient rounds of screening, the nadir observed in it would underestimate the asymptote expected in a sustained program with the same spacing of screenings. Thus, modelling assumptions are required to extrapolate from a trial of say 3 annual screens to a program with say 20 annual screens. The *'round by round approach'* we have described in Figure 3 can also be immediately applied to programs with different durations and spacings (e.g. 20 annual screens versus 10 biennial screens).

Several projections of the mortality reductions due to cancer screening have been based on extensive modelling of the natural histories of cancers and how their progress is altered by earlier detection and therapy. Many of these efforts [56, 57, 39] have also quantified the associated costs and use very sophisticates simulation modelling to examine the impact of prevention, screening, and treatment on cancer incidence and mortality at the population level. These approached usually require a very large number of parameter inputs, obtained from diverse data sources (such as trials, registries and surveys).

We first illustrate a round-by-round approach, using the model proposed by Hu and Zelen [41]. Previously, it has mostly been used for planning early-detection trials, including the recent NLST, where the yearly numbers were aggregated for the power calculation for the interim and ultimate statistical tests performed during and at the end of the trial. We use it here to generate and display the rate ratio curve proposed in Section 4–2, to show the projected timing, magnitude and duration of the yearly reductions in a *program* (the yearly numbers that the software aggregates for power calculations do not appear to have been previously used for this purpose). The Hu-Zelen model the mortality in each year under the screening and no-screening scenarios via a total of seven parameters (see Fig 4–4) quantifying the sensitivity of the screening test, the natural (and altered) course of cancer from initiation to normal clinical diagnosis and post clinical diagnosis under the no-screening and screening scenarios.

4.4.2 Illustration

Since sufficient information to fit new parameter values has not yet been extracted from the completed NLST, we will use some modifications of the input values [73] used to plan the trial. Rather than use the FORTRAN software the trial statisticians used to implement the Hu-Zelen integrals, we re-programmed them in R. The only modifications we made were to two of the input parameters, to better represent how the cancer deaths are averted. In the planning, the authors assumed the 'average' CT sensitivity would be 85%, and that those whose cancers were detected by screening would have their (counterfactual) post-clinical-diagnosis survival altered from an exponential distribution with a median 1.53 or 1.74 years to one where the median was 2.42 or 2.21 years: (the planning calculations assumed that all would eventually die of their cancer; moreover, there was no possibility of a 'cure', unless by a 'cure' one means that one dies of another cause). Instead, in light of the very rapid progression of many lung cancers, and the possibility of over-diagnosis, we assumed that the 'real' sensitivity was much less, and that the possibility of cure (rather than a very short extension of a few months of life) was confined to subgroup of screen-detected cancers; the remainder, even if detected by screening, would continue to have virtually the same mortality rates as their counterparts who were not screened. Thus, we set the 'sensitivity' at 25% rather than 85%, and the median survival of 30 years ('cure') for those whose otherwise fatal cancers were found at a curable stage.

Figure 4–4 (top) shows the resulting 35-year projection for a program of 20 annual screenings. With the exception of the slightly unrealistic (but numerically

inconsequential) pattern at the front end (see below), the rate ratio curve, and its complement the reduction curve, resemble the anticipated bathtub-shape presented in Figure 4–1. The curve stays close constant for the middle part where there was sustained screening, and it gradually tails off after screening was stopped. The 'excess' deaths after years 25 are a consequence of the assumed exponential survival model in which cancer deaths are merely delayed, not averted - in keeping with the corresponding pattern shown in version (b) of the Figure in Morrison's textbook.

Figure 4–4 (bottom) shows the projection for a biennial program; it is a little shallower than the annual one, but the reductions persist for almost the same duration. The oscillations in the 'round by round' waves are more prominent than in (a), and reflect the local effects of variations in the progression rates of different cancers together with the intra-individual variability in their stages at each examination time. The considerably smaller morality reductions than in (a) emphasize the fact that two year screening intervals allow many more lung cancers to progress to the incurable stage in the interim.

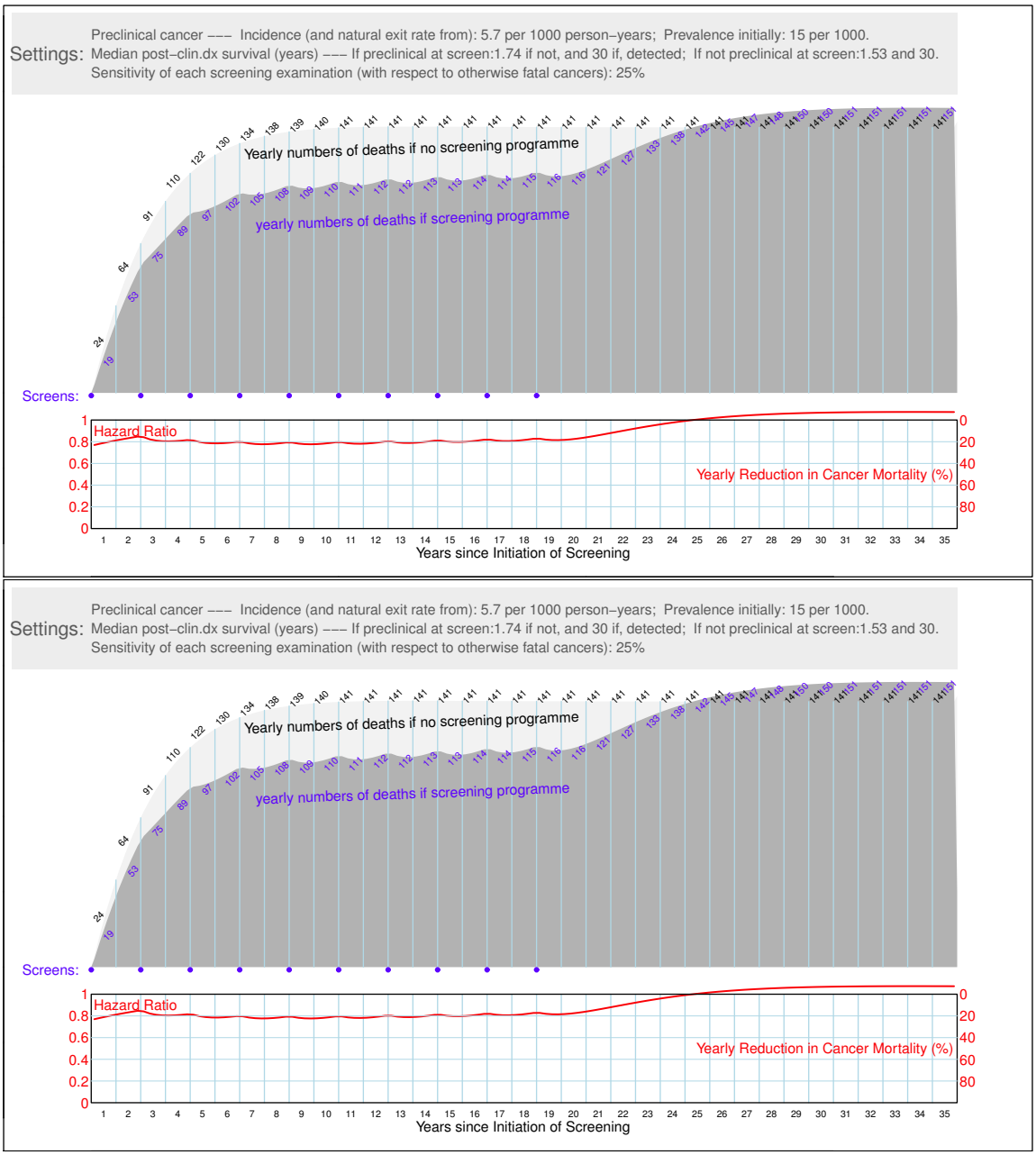


Figure 4–4: Illustration of the mortality projections due to annual and biennial screenings using the Hu-Zelen model (see full caption on the next page).

Figure 4–4 caption: A 35-year projection of lung cancer mortality reductions for a program of (top) 20 annual and (bottom) 10 biennial screenings, based on the same Hu-Zelen model used to plan the NLST trial but with the 7 indicated input parameters (see text regarding the sensitivity and survival inputs), together with the associated (almost-bathtub shaped) rate ratio curves. The comparison is between screening with low-dose CT screening and Chest X-Ray (shown to be virtually ineffective in the PLCO trial). The 'excess' deaths after years 25 are a consequence of the exponential survival assumption in the Hu-Zelen model, in which cancer deaths are merely postponed, not averted – similar to the pattern shown in Figure 2-5(a) in Morrison's textbook. Newer program projections will be made once we have extracted the parameter values from the NLST data.

Possible reasons why the early portion of the projected curve does not show the anticipated time lag more clearly may include (i) the numbers of cancer-specific deaths are expected to be very small in the first few years, which lead to large uncertainty in the early portion of the rate ratio curve; (ii) the exponential form, assumed for the sojourn time distribution, does not take into account the time lag between screenings and their induced mortality reductions, (iii) the assumption of independence between an individual's sojourn time and their post-clinical diagnosis survival time: we would expect a strong correlation, that is, a relatively fast-growing cancer would be aggressive both pre- and post-detection; and (iv) the mortality rates do not explicitly accommodate cures from cancer nor deaths from other causes.

In order to deal with these front-end and back-end issues, considerably more refinements would need to be incorporated into the model, such as stage-specific sensitivities, transition rates, and survival distributions, as well as age-specific competing risks. While Zelen and colleagues, and other CISNET investigators, have indeed incorporated such refinements, they now face the reality of having to deal with the over-diagnosis that accompanies the newer screening tools, and the added model complexity and uncertainty. Instead, we are currently exploring a minimalist model that focuses only on the mortality reductions.

4.5 Summary

Unlike therapeutic trials in patients, cancer screening trials in asymptomatic persons generate mortality reductions that can only manifest several years after the onset of screening. The often reported single-number cumulative mortality reduction, in either a trial or a meta-analysis of trials, is of limited use in projecting the

timing, duration and magnitude of the mortality reductions that would be expected from a sustained screening program, of longer duration and possibly with a different screening regimen.

Instead, we propose using a rate ratio curve, and its complement, the mortality reduction curve, to address the mortality impact (timing, magnitude, and duration) of a screening program. This curve is easy to interpret, as it shows when reductions begin, how big they are, and how long they last. We illustrate, using an existing model, how one could compute such rate ratio curves, and quantitatively compare the impact of different screening regimens over the appropriate time-window.

Our message is two-fold: we (1) recommend against using one-number summaries to deduce the yearly mortality reductions expected from a sustained screening program, and (2) call on trialists to report necessary time-specific mortality data to allow the appropriate computation of rate ratio curves that allow the mortality impacts of different screening programs to be compared over the appropriate time horizon.

Acknowledgments

This work was supported by the Canadian Institutes for Health Research.

CHAPTER 5
A Conditional Approach to Measure Mortality Reductions due to Cancer Screening

Preamble to Manuscript 2. Miettinen and Karp [63, p. 36] make a distinction between *etiogenetic causality* and *interventive causality*, knowing about the former being characterized by its retrospective nature:

> Ad-hoc knowing about etiogenetic causality – etiognosis, that is (cf. above) – is tantamount to having a causal explanation of an existent outcome (level of a morbidity, or presence of an illness); it thus inherently is retrospective from the vantage of an existent outcome.

This very much agrees with the work of a detective, as noted by Mukherjee [70, p.9], who quotes Sherlock Holmes, in Sir Arthur Conan Doyle's *A Study in Scarlet*:

> In solving a problem of this sort, the grand thing is to be able to reason backwards. That is a very useful accomplishment, and a very easy one, but people do not practice it much.

Keeping this in mind, in this manuscript we develop a novel conditional probability model for measuring the mortality impact of cancer screening, directly addressing the probability of averting cancer death through the introduction of screening, given that the cancer would have proven fatal in the absence of it. We formulate the corresponding causal estimand in terms of potential outcome random variables, specify the identifying assumptions required to estimate it based on observed mortality data,

and propose an estimation method based on conditional likelihood. The model is aimed at extracting the time-specific mortality impact of a single round of screening; these functions can then be compounded to construct the mortality impact of a particular screening regimen. We apply the method in two case studies, where the model is fitted to existing trial data, and used for projecting the mortality impact of a sustained screening program.

A Conditional Approach to Measure Mortality Reductions due to Cancer Screening

Zhihui (Amy) Liu, James A Hanley, Olli Saarela, and Nandini Dendukuri

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, Quebec H3A 1A2, Canada.

Revision invited in September 2014.

Abstract

Evidence of benefits produced by cancer screening is commonly reported as a mortality reduction calculated over the entire follow-up period of a randomized screening trial. However, such a single-number statistic is of limited use in projecting the mortality impact expected from a sustained screening program. We develop a novel probability model to project the mortality impact, by parametrizing the conditional probability of being helped by a single round of screening, given that the cancer would have proven fatal otherwise. This represents a major shift in focus in quantifying the impacts of screening, enabling extracting more relevant statistical evidence from existing trial data. We illustrate our approach using data from screening trials in lung and colorectal cancers.

5.1 Introduction

The decision on whether to implement a long-term cancer screening program in a population requires weighing the harms and costs against the health benefits, such as the number of cancer deaths averted every year. The evidence of the benefits is often based on a single-number summary, such as the cancer-specific mortality reduction over the entire follow-up period in a randomized screening trial, or an average of such one-number measures from a meta-analysis of trials. However, such a single-number statistic is of limited use in projecting the mortality reductions that would be produced by a screening program, of longer duration and possibly with a different screening regimen [53]. In particular, this is because the mortality reduction is not constant over time after the initiation of the screening, as has been recognized and discussed for instance by Morrison [68, p. 36], Miettinen et al. [62], Hanley [35, 36, 37], Miettinen and Karp [63, p. 81], and most recently by us in Hanley et al. [38].

Our objective is to project the mortality impact produced by a sustained screening program based on trial data, a task that requires modelling of the round-specific impacts. Microsimulations previously used for this purpose [e.g. 11, 100, 56] are based on modelling of the entire disease history (e.g. from disease free to pre-clinical disease state to clinically diagnosed). Such models will typically involve a large number of parameter inputs that are not obtainable from a single data source, as well as many generally unverifiable assumptions. In contrast, the conditional approach we propose eliminates parameters characterizing prevalence, incidence, sensitivity, state transitions, or sojourn time, and produces evidence-based, probabilistic projections.

We specify our estimand of interest as the conditional probability of being helped by screening (through earlier treatment) given that the cancer would have proven fatal in the absence of screening. This estimand is equivalent to the 'factor-conditional etiogenetic proportion' of cancer deaths due to lack of screening associated early treatments [63, p. 82]. We show that this conditional probability has a direct interpretation as the proportional reduction in cancer mortality, and that it can be decomposed into a function of round-specific reductions. We suggest a parametric form for the round-specific reduction, based on which we then formulate a likelihood function.

The remainder of the paper is organized as follows. The estimand and the assumptions necessary to identify it are specified in Section 2. In Section 3, we formulate a parametric model to characterize the round-specific impact and the resulting likelihood expressions for individual-level and aggregated data. In Section 4, we fit our model to data from screening trials in lung and colorectal cancer and illustrate the resulting projections. The paper concludes with a discussion in Section 5.

5.2 Specifying the estimand

5.2.1 Notation

In a randomized screening trial, subjects asymptomatic of cancer are randomly assigned to either a screening or non-screening arm at time $s_0 = 0$, and all are followed up for death due to the cancer or another cause, or until the end of follow-up at time τ, whichever comes first. During the interval $[0, \tau]$ a total of m screening examinations are carried out at the ordered time points $s_1 < s_2 < \cdots < s_m$ in

the screening arm, with the jth interval denoted by $[s_{j-1}, s_j)$ and its length by $\Delta_j = s_j - s_{j-1}$ for $j = 1, 2, \ldots, m$.

We define a screening indicator Z_i taking the value 1 if individual i is assigned to the screening arm, with $Z_i = 0$ otherwise. Let T_i denote the observed time of the event (i.e. death due to the cancer, death due to another cause, or Type I censoring due to the end of the follow-up period at τ). We take this to be $T_i = Z_i T_{1i} + (1 - Z_i) T_{0i}$, where T_{1i} and T_{0i} denote the potential/counterfactual event times under screening and in the absence of it, respectively. (This corresponds to assuming either 'stable unit treatment value', [e.g. 7], or 'consistency', [e.g. 17].) Similarly, let E_i denote the observed event type, taking the value of 1 for cancer-specific death, 2 for death due to another cause and 0 for censoring. This is given by $E_i = Z_i E_{1i} + (1 - Z_i) E_{0i}$, where E_{1i} and E_{0i} are indicator variables for the potential/counterfactual event types under screening and in the absence of it. The unobservable gained survival time for individual i is $G_i \equiv T_{1i} - T_{0i}$.

5.2.2 Object of inference

We take the estimand to be the probability that a cancer-specific death in the absence of screening was indeed 'caused' by the absence of screening associated early treatments (cf. the 'factor-conditional etiogenetic proportion' of Miettinen & Karp [63, p. 48]). This probability in turn is equivalent to the probability of being helped by screening, had it been available. This is specified as the conditional probability

$$H(t) \equiv P(T_{1i} > t \mid T_{0i} = t, E_{0i} = 1) \tag{5.1}$$

of surviving beyond time t under screening, given a cancer death at time t in the absence of it. Since an individual's cancer can be detected, and subsequently successfully treated, as a result of only one screening examination, we introduce a random variable $S_i \in \{s_1, s_2, \ldots, s_m, \infty\}$ to represent the time of being detected, and subsequently successfully treated, with $S_i = \infty$ taken to mean that the cancer was not detected in any of the scheduled screenings. Furthermore, since only the screening examinations before the time of death $T_{0i} = t$ can potentially be helpful, we take $m(t) \equiv \max\{j \in \{1, 2, \ldots, m\} : s_j < t\}$ to denote the last screening examination before t. Now we can express (5.1) as

$$\begin{aligned} H(t) &= \sum_{j=1}^{m(t)} P(S_i = s_j \mid T_{0i} = t, E_{0i} = 1) \\ &= \sum_{j=1}^{m(t)} P(S_i = s_j \mid T_{0i} = t, E_{0i} = 1, S_i \geq s_j) P(S_i \geq s_j \mid T_{0i} = t, E_{0i} = 1) \end{aligned} \quad (5.2)$$

The first term inside the sum (5.2) is the probability of being helped as a result of the jth screening given that the previous screenings at times $s_1, s_2, \ldots, s_{j-1}$ failed to detect the cancer. Since only new or previously undetected cancers can be detected in the jth screening, we take the probability

$$Q_j(t) \equiv P(S_i = s_j \mid T_{0i} = t, E_{0i} = 1, S_i \geq s_j) \quad (5.3)$$

as our measure to quantify the mortality impact of a single round of screening; modelling of the round-specific impact is needed to project the mortality impact of a sustained screening program. The probability (5.2) is fully specified in terms of

(5.3), $j = 1, \ldots, m$, as

$$H(t) = \sum_{j=1}^{m(t)} Q_j(t) \prod_{k=1}^{j-1} \{1 - Q_k(t)\} = 1 - \prod_{j=1}^{m(t)} \{1 - Q_j(t)\}, \tag{5.4}$$

which follows from the failure probability function for a discrete failure time random variable [e.g. 44, p. 9]. The representation (5.4) in turn enables likelihood construction through parametrization of the functions $Q_j(t)$ (Section 5.3).

5.2.3 Identifying assumptions

Since (5.1) is expressed in terms of unobservable quantities, further assumptions are needed to identify it based on observed data. One possible approach would be to assume an accelerated failure time model, such as $T_{1i} = T_{0i} e^{g(T_{0i})}$ for the potential outcomes [e.g. 40]. In this case

$$\begin{aligned} P(T_{1i} > t \mid T_{0i} = t, E_{0i} = 1) &= P(T_{0i}(e^{g(t)} - 1) > 0 \mid T_{0i} = t, E_{0i} = 1) \\ &= P(g(t) > 0 \mid T_{0i} = t, E_{0i} = 1), \end{aligned}$$

which equals one whenever the acceleration/deceleration function g is positive, and zero otherwise. This suggests that direct modelling of the gained survival time G_i is unhelpful in addressing the probability of being helped by screening. Instead, we pursue modelling in terms of cause-specific sub-density functions $f_k(t) \equiv P(T_{ki} \in \mathrm{d}t, E_{ki} = 1)/\mathrm{d}t$ [cf. 44, p. 252], $k = 0, 1$, for individual i dying of the cancer at time t in the absence and presence of screening, respectively. (We use $\mathrm{d}t$ to denote both an infinitesimally small interval around t and the infinitesimal length of this interval.)

In order to estimate (5.1), four identifiability assumptions we make in this paper are (i) monotonicity $T_{0i} \leq T_{1i}$; (ii) strongly ignorable assignment, that is,

$\{(T_{1i}, E_{1i}), (T_{0i}, E_{0i})\} \perp\!\!\!\perp Z_i$ and $0 < P(Z_i = 1) < 1$ [cf. 79, p. 43]; (iii) curative early treatments, in the sense that

$$P(T_{1i} > t \mid T_{0i} = t, E_{0i} = 1) = P(T_{1i} > t, E_{1i} \neq 1 \mid T_{0i} = t, E_{0i} = 1); \tag{5.5}$$

and (iv) screening specificity, that is, $E_{0i} = 2 \Rightarrow T_{1i} = T_{0i}, E_{1i} = 2$.

Assumption (i) states the potential time of death of any cause for an individual in the screening arm is at least as long as that in the non-screening arm [cf. 7], that is, screening cannot shorten anyone's life. Assumption (ii) is satisfied automatically due to randomized allocation in the trial. Assumption (iii) states that the screening-associated early treatments cure the cancer, in the sense of delaying the cause-specific death beyond a death due to a competing cause (or censoring). Assumption (iv) states that the screening technique is specific in the sense that it does not lead to early detection and treatment of conditions other than the site-specific cancer of interest.

Nine different types of event histories, possible under the assumptions (i)-(ii), are illustrated in Figure 5–1. Subject 1 would die of another cause which could not have been prevented by screening. The death of subject 2 due to another cause was delayed due to screening. This is possible, although unlikely, if the screening can also lead to detection of other conditions than the site-specific cancer of interest. The same applies to histories for subjects 3 and 4. Subject would 5 be alive at the end of the follow-up time, and the time of death due to the cancer for subject 7 would be the same with and without screening; thus during the follow-up neither of them could have benefited from screening. Subjects 6, 8 and 9 would die of the

cancer in the absence of screening, but in the presence of screening they would die of another cause, or die later due to the cancer, or be censored at the end of the follow-up, respectively; thus, they could benefit from early-detection of the cancer and consequent therapy. Introducing assumption (iii) rules out the event histories of type 5. As we will demonstrate in Section 5.2.4, this is required for identification of (5.1), since delayed cancer deaths in the screening arm cannot be distinguished from the non-delayed ones based on the observed data. Similarly, we need to use assumption (iv) to rule out histories of type 3, since these would show as excess mortality in the screening arm.

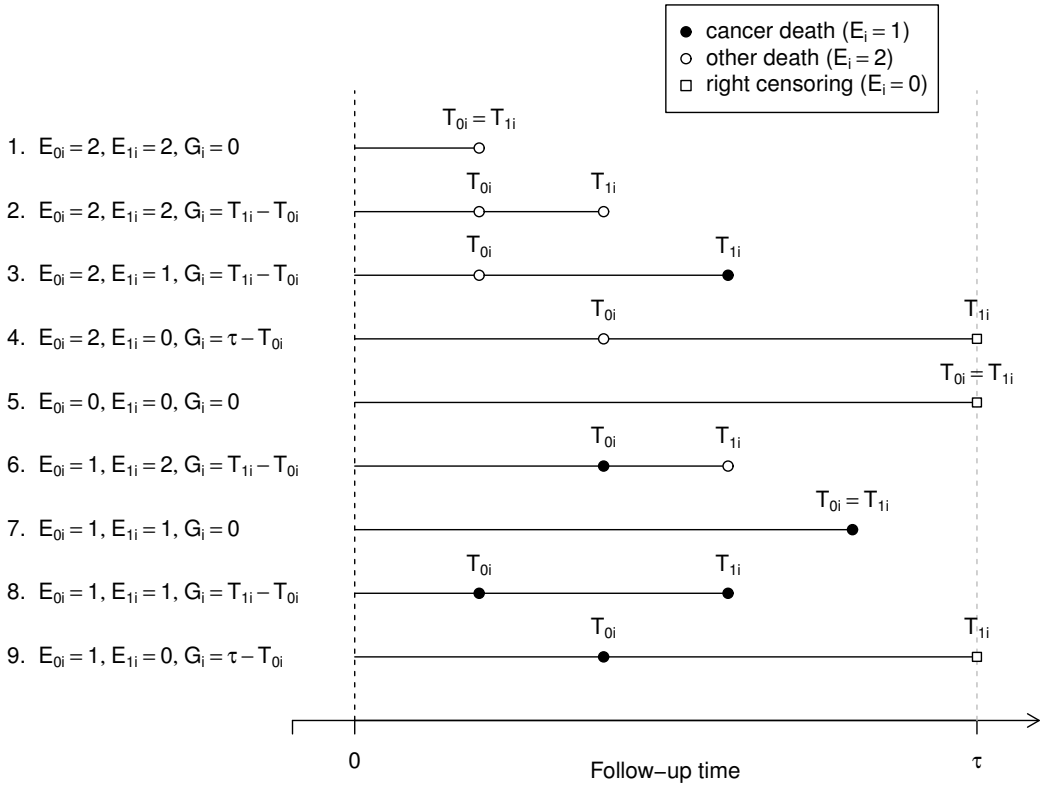


Figure 5–1: Illustration of 9 different possible event histories, one row per individual.

Further, we note that under continuous time no two cause-specific counting processes can jump simultaneously [1, p. 55], unless they are in fact the same process (which would occur if there is no screening effect). In the present setting this means that $T_{0i} = T_{1i} \Rightarrow E_{0i} = E_{1i}$, ruling out event histories of the type $E_{0i} \neq E_{1i}, G_i = 0$ not present in Figure 5–1.

5.2.4 Equivalence between the probability of being helped and mortality reduction

We show that under the assumptions stated in Section 5.2.3, probability (5.1) of being helped by screening is equivalent to time-specific reduction in cancer mortality, a quantity that can be estimated based on the trial data. We may express (5.1) as

$$
\begin{aligned}
& P(T_{1i} > t \mid T_{0i} = t, E_{0i} = 1) \\
& = 1 - P(T_{1i} \leq t \mid T_{0i} = t, E_{0i} = 1) \\
& = 1 - \sum_{k=0}^{2} P(T_{1i} \in \mathrm{d}t, E_{1i} = k \mid T_{0i} = t, E_{0i} = 1) \\
& = 1 - P(T_{1i} \in \mathrm{d}t, E_{1i} = 1 \mid T_{0i} = t, E_{0i} = 1) \\
& = 1 - \frac{P(T_{0i} \in \mathrm{d}t, E_{0i} = 1 \mid T_{1i} = t, E_{1i} = 1) P(T_{1i} \in \mathrm{d}t, E_{1i} = 1)}{P(T_{0i} \in \mathrm{d}t, E_{0i} = 1)} \\
& = 1 - \frac{f_1(t)}{f_0(t)}.
\end{aligned}
$$

The second equality is due to the monotonicity assumption (i). The third equality follows from the continuous time model for the counting processes. The fifth equality is due to $P(T_{0i} \in \mathrm{d}t, E_{0i} = 1 \mid T_{1i} = t, E_{1i} = 1) = 1$, which follows from assumptions (i), (iii) and (iv).

While the estimand is specified in terms of potential outcome variables, the ignorability assumption (ii) enables its estimation using the observed outcomes in the two trial arms, because

$$1 - \frac{f_1(t)}{f_0(t)} = 1 - \frac{P(T_i \in \mathrm{d}t, E_i = 1 \mid Z_i = 1)}{P(T_i \in \mathrm{d}t, E_i = 1 \mid Z_i = 0)} = 1 - \frac{f(t \mid Z_i = 1)}{f(t \mid Z_i = 0)},$$

where $f(t \mid Z_i) \equiv P(T_i \in \mathrm{d}t, E_i = 1 \mid Z_i)/\mathrm{d}t$.

5.2.5 Relationship to cumulative mortality reduction

Our estimand, the probability of being helped, which equals the time-specific mortality reduction, has a natural connection to the cumulative mortality reduction, a measure commonly used to quantify the impact of screening in randomized trials (see Section 5.4). With the same assumptions as stated in Section 5.2.3, we can express the probability of surviving beyond the potential time of death in the absence of screening, had the cancer proven fatal without screening *before* time t, as

$$\begin{aligned}
&P(T_{1i} > T_{0i} \mid T_{0i} \leq t, E_{0i} = 1) \\
&= 1 - P(T_{1i} \leq T_{0i} \mid T_{0i} \leq t, E_{0i} = 1) \\
&= 1 - \frac{\int_{v \in [0,t]} P(T_{0i} \in \mathrm{d}v, E_{0i} = 1 \mid T_{1i} = v, E_{1i} = 1) P(T_{1i} \in \mathrm{d}v, E_{1i} = 1)}{\int_{v \in [0,t]} P(T_{0i} \in \mathrm{d}v, E_{0i} = 1)} \\
&= 1 - \frac{\int_0^t f_1(v)\,\mathrm{d}v}{\int_0^t f_0(v)\,\mathrm{d}v} \equiv 1 - \frac{F_1(t)}{F_0(t)}.
\end{aligned}$$

In the context of planning a trial, Hu and Zelen [41, p. 823] use the risk difference $F_0(\tau) - F_1(\tau)$ at the end of the follow-up period as the measure of the impact of the planned screening regimen used in the trial. As demonstrated here, under the assumptions of Section 5.2.3, the proportional risk difference, $1 - F_1(\tau)/F_0(\tau)$, is

equivalent to the probability of being helped by screening given a cancer death during the follow-up window $[0, \tau]$ in the absence of screening.

5.3 Methods

In this section, we present a parametric model characterizing the effect of a single round of screening. The mortality reduction at time t can then be obtained as a compound of the impacts of each screen that persons have received up to t, as shown in Section 5.2.2.

5.3.1 Model formulation

As emphasized by several authors referred to in the Introduction, a quintessential feature of cancer screening is the non-constancy of its impact over time. According to Miettinen [61], it is a fundamental truism that the mortality reduction "cannot be constant over successive intervals of time after the screening's initiation; that it is initially nil, then increases and later declines, and ultimately totally vanishes". Unlike most medical interventions that produce a virtually immediate effect, a cancer that would prove fatal within months from now is not likely to be cured by screening today, while a cancer that is cured today due to early detection would otherwise have proven fatal several years from now.

'Memoryless' property

To start with, we assume that the probabilities of being helped by each round of screening as functions of time are shifted versions of each other, that is,

$$Q_1(t) = Q_2(t + \Delta_1) = \cdots = Q_m\left(t + \sum_{k=1}^{m-1} \Delta_k\right). \tag{5.6}$$

For simplicity, take the screenings to be equally spaced so that $\Delta_1 = \Delta_2 = \ldots = \Delta_{m-1} = \Delta$, and take the first two screens as an example. Now

$$
\begin{aligned}
& Q_1(t) = Q_2(t+\Delta) \\
\Leftrightarrow \quad & P(S_i = s_1 \mid T_{0i} = t, E_{0i} = 1, S_i \geq s_1) \\
& = P(S_i = s_1 + \Delta \mid T_{0i} = t + \Delta, E_{0i} = 1, S_i \geq s_1 + \Delta) \\
\Leftrightarrow \quad & 1 - P(S_i > s_1 \mid T_{0i} = t, E_{0i} = 1, S_i \geq s_1) \\
& = 1 - P(S_i > s_1 + \Delta \mid T_{0i} = t + \Delta, E_{0i} = 1, S_i \geq s_1 + \Delta) \\
\Leftrightarrow \quad & P(S_i \geq s_1 + \Delta \mid T_{0i} = t, E_{0i} = 1, S_i \geq s_1) \\
& = P(S_i \geq s_1 + \Delta + \Delta \mid T_{0i} = t + \Delta, E_{0i} = 1, S_i \geq s_1 + \Delta).
\end{aligned}
$$

Thus, modelling assumption (5.6) can be interpreted as a 'memoryless' property for the random variable S_i. This is plausible, since the length $T_{0i} - S_i$ of the interval from time of screen detection to the potential time of death without screening is kept constant. Further, the above probabilities are conditional on not being detected in the previous screening examinations, and if the sensitivity of the screening test and participation rates are high, the cancers to be detected at any s_j are mainly 'new' ones, having progressed to the detectable state in the interval $[s_{j-1}, s_j)$. While this applies to the repeat screenings, the first, or 'prevalence' screening might involve a different stage distribution of cancers; we address this question briefly in Section 5.5.

Examples of possible parametrizations

The effect of one round of screening could be characterized in terms of maximal reduction (γ), the time lag between the time of screening and the maximal reduction

(location parameter μ), and the spread of the reductions over time (scale parameter σ). Using these three parameters, a possible formulation for the time-specific reduction due to one screen is

$$Q_j(t;\gamma,\mu,\sigma) \equiv \gamma \exp\left\{-\left(\frac{t-(\mu+\sum_{l=1}^{j}\Delta_l)}{\sigma}\right)^2\right\}, \tag{5.7}$$

where $0 \le t < \infty$, $0 \le \gamma \le 1$, $\mu > 0, \sigma > 0$. Function (5.7) characterizes how deep, how far into the future, and how wide the mortality reductions produced by a single screen are.

A possible limitation of formulation (5.7) is that it does not enforce the restriction $\lim_{t\to s_j^+} Q_j(t) = 0$ (if the disease has already progressed so far that death would have resulted immediately after the detection, any subsequent therapy comes too late to help the patient). An alternative non-symmetric formulation that satisfies this restriction could be

$$\begin{aligned} Q_j(t;\gamma,\alpha,\beta) &\equiv \gamma \frac{f(t-\sum_{l=1}^{j}\Delta_l;\alpha,\beta)}{f((\alpha-1)\beta;\alpha,\beta)} \\ &= \gamma \left\{\frac{t-\sum_{l=1}^{j}\Delta_l}{(\alpha-1)\beta}\right\}^{\alpha-1} \exp\left\{(\alpha-1)-\frac{t-\sum_{l=1}^{j}\Delta_l}{\beta}\right\}, \end{aligned} \tag{5.8}$$

where $0 \le t < \infty$, $0 \le \gamma \le 1$, $\alpha > 1$, $\beta > 0$. Here $f(t;\alpha,\beta)$ is the probability density function of a gamma distribution with the mode $t = (\alpha-1)\beta$. By scaling down the density function by its maximum value, we are restricting the time-specific mortality reductions to be between zero and one. Possible shapes with various parameter inputs for the symmetric and non-symmetric formulations can be found in Figure 5–2.

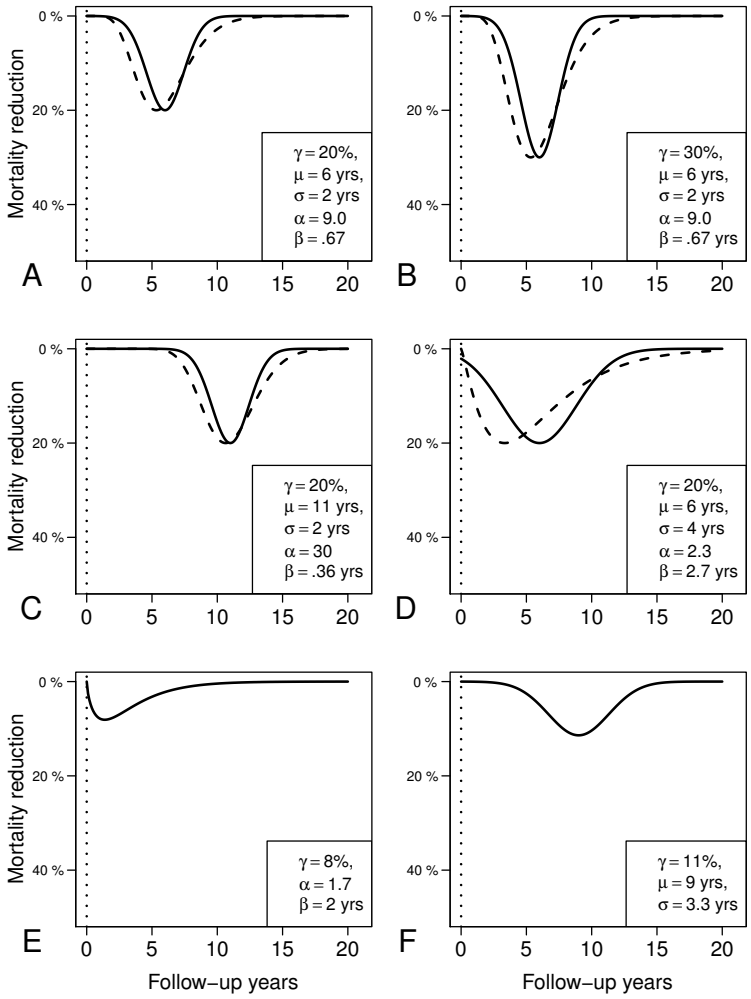


Figure 5–2: Impact of a single round of screening at time $s_1 = 0$, with different patterns determined by different parameter inputs. Solid and dashed lines correspond to Equations (5.7) and (5.8), respectively. Panels E and F correspond to the fitted reduction patterns in the examples of Sections 5.4.1 and 5.4.2, respectively.

Taking θ to be the collection of the three parameters, appropriately transformed, the compound reduction $H(t;\theta)$ resulting from $m(t)$ screens before time t can now be obtained by substituting $Q_j(t;\theta)$ into equation (5.4).

5.3.2 Likelihood formulation

Individual-level data

Since we are interested in modelling the mortality reduction rather than absolute mortality, we adopt a conditional approach. The conditional likelihood contribution of individual i is given by the distribution of the screening allocation indicator Z_i given that there was a cancer death at t, that is, $Z_i \mid (T_i = t, E_i = 1) \sim$ Bernoulli$\{\pi(t)\}$, resulting in likelihood contributions of the form $\pi(t)^{Z_i}(1-\pi(t))^{1-Z_i}$ for each cancer death. Here, with equal allocation $P(Z_i = 1) = P(Z_i = 0) = 0.5$ between the two arms,

$$\begin{aligned}
&\pi(t) \equiv P(Z_i = 1 \mid T_i = t, E_i = 1) \\
&= \frac{P(T_i \in \mathrm{d}t, E_i = 1 \mid Z_i = 1)P(Z_i = 1)}{P(T_i \in \mathrm{d}t, E_i = 1 \mid Z_i = 0)P(Z_i = 0) + P(T_i \in \mathrm{d}t, E_i = 1 \mid Z_i = 1)P(Z_i = 1)} \\
&= \frac{f(t \mid Z_i = 1)/f(t \mid Z_i = 0)}{1 + f(t \mid Z_i = 1)/f(t \mid Z_i = 0)} \\
&= \frac{1 - H(t;\theta)}{1 + 1 - H(t;\theta)}. \qquad (5.9)
\end{aligned}$$

Aggregated data

If the individual-level mortality data are not reported or accessible, our model can be fitted to aggregated (e.g. yearly) numbers of deaths in each arm, extractable from the cumulative mortality curves in the published trial reports [54]. Let $D_{0j} = \sum_i \mathbf{1}_{\{E_i=1, t_{j-1}<T_i<t_j, Z_i=0\}}$ and $D_{1j} = \sum_i \mathbf{1}_{\{E_i=1, t_{j-1}<T_i<t_j, Z_i=1\}}$ denote the numbers of cancer-specific deaths during the interval $[t_{j-1}, t_j)$, $j = 1, 2, \ldots, J$, in the non-screening and screening arm, respectively. Thus the distribution of D_{1j} conditional on the total deaths during interval j is $D_{1j} \mid (D_{0j} + D_{1j} = d_j) \sim$ Binomial(d_j, π_j),

where

$$\begin{aligned}\pi_j &= \frac{N_1[F(t_j \mid Z_i = 1) - F(t_{j-1} \mid Z_i = 1)]}{N_0[F(t_j \mid Z_i = 0) - F(t_{j-1} \mid Z_i = 0)] + N_1[F(t_j \mid Z_i = 1) - F(t_{j-1} \mid Z_i = 1)]} \\ &= \frac{[F(t_j \mid Z_i = 1) - F(t_{j-1} \mid Z_i = 1)]/[F(t_j \mid Z_i = 0) - F(t_{j-1} \mid Z_i = 0)]}{1 + [F(t_j \mid Z_i = 1) - F(t_{j-1} \mid Z_i = 1)]/[F(t_j \mid Z_i = 0) - F(t_{j-1} \mid Z_i = 0)]} \\ &\approx \frac{1 - \int_{t_{j-1}}^{t_j} H(t;\theta)\frac{1}{t_j - t_{j-1}}\,\mathrm{d}t}{1 + 1 - \int_{t_{j-1}}^{t_j} H(t;\theta)\frac{1}{t_j - t_{j-1}}\,\mathrm{d}t}, \end{aligned} \tag{5.10}$$

where $N_1 = N_0$ are the numbers of individuals randomized to screening and control arms, respectively. Notably, $\lim_{t_j \to t_{j-1}} \pi_j = \pi(t_{j-1})$, reducing to the individual-level formulation in (5.9). The resulting log-likelihood function is the sum of contributions from the entire duration of the follow-up time, given by $l(\theta) \equiv \sum_{j=1}^{J}\{D_{1j}\log(\pi_j) + D_{0j}\log(1-\pi_j)\}$.

5.3.3 Estimation

The likelihood functions for individual-level or aggregated data in Sections 5.3.2 and 5.3.2 can be maximized with respect to parameters specifying the mortality reduction function $H(t;\theta)$ using standard numerical optimization methods. Since all the parameters in (5.7) or (5.8) are positive, re-parametrizations should be used when applying a normal approximation to the likelihood in order to obtain standard errors for the parameter estimates. However, rather than the individual parameters, our main interest is in obtaining measures of uncertainty for the mortality projections. Since the projections are based on a probability model fitted using maximum likelihood, time-specific confidence bands may be constructed straightforwardly by sampling parameter estimate values from the approximate large-sample sampling

distribution $N(\hat{\theta}, i(\hat{\theta})^{-1})$, where $i(\hat{\theta})$ is the observed information matrix at the maximum likelihood point, and calculating the projection curve at each value. The 2.5% and 97.5% sample quantiles at each time point can be easily obtained based on 10,000 random draws.

5.3.4 Generalizations

In this subsection, we extend our model to accommodate unequal allocation of person-time between the screening and non-screening arms, less than full compliance, and multiple screening arms within a trial.

If the randomization ratio between the screening arm and the non-screening arm is $N_1/N_0 \equiv \phi : 1$ instead of 1:1, such as in the Swedish two-county trial [88], as well as two other mammography screening trials in Stockholm [29] and Gothenburg [12], then equation (5.10) becomes

$$\pi_j = \frac{\phi\{1 - \int_{t_{j-1}}^{t_j} H(t;\theta)\frac{1}{t_j - t_{j-1}}\,\mathrm{d}t\}}{1 + \phi\left\{1 - \int_{t_{j-1}}^{t_j} H(t;\theta)\frac{1}{t_j - t_{j-1}}\,\mathrm{d}t\right\}}.$$

Sometimes multiple screening arms are employed within the same trial, such as the Minnesota colorectal cancer study [84] in which participants were randomly assigned to be screened annually, biennially, or not at all. To accommodate this, let D_{kj} denote the number of cancer-specific deaths in arm k, where $k = 0, \ldots, K$, during the jth interval. Given the total number of deaths $d_j = \sum_{k=0}^{K} D_{kj}$, the split into the K study arms is distributed as

$$D_{0j}, \ldots, D_{Kj} \mid \left(\sum_{k=0}^{K} D_{kj} = d_j\right) \sim \text{Multinomial}(d_j, \pi_{0j}, \ldots, \pi_{Kj}),$$

resulting in a log-likelihood function $l(\theta) \equiv \sum_{j=1}^{J} \sum_{k=0}^{K} D_{kj} \log(\pi_{kj})$, where $\pi_{0j} = 1 - \sum_{k=1}^{K} \pi_{kj}$.

While our estimand (5.1) should be interpreted as an intention-to-treat type effect, with the potential outcome (T_{1i}, E_{1i}) corresponding to being randomized to the screening arm of the trial, as opposed to actually undergoing screening as scheduled, in the projection task it might be appropriate to upscale or downscale the mortality impact of the screening program by the expected participation rate. In addition, a relevant quantity for decision making at the individual level would be the mortality impact conditional on compliance. Assuming that the compliance in the screening round j of the trial, denoted as $C_{ij} = 1$, is completely at random in the sense that $P(C_{ij} = 1 \mid T_{0i} = t, E_{0i} = 1, S_i \geq s_j) = P(C_{ij} = 1) \equiv c_{Tj}$, the complier probability of being helped in this round is simply

$$P(S_i = s_j \mid T_{0i} = t, E_{0i} = 1, S_i \geq s_j, C_{ij} = 1) \equiv Q_j^*(t) = \frac{1}{c_{Tj}} Q_j(t), \tag{5.11}$$

since $P(S_i = s_j, C_{ij} = 0 \mid T_{0i} = t, E_{0i} = 1, S_i \geq s_j) = 0$. Differential compliance between the successive rounds of screening in the trial may now be accounted for by using the relationhip (5.11) in fitting the likelihoods (5.9) or (5.10), by replacing $Q_j(t)$ in equation (5.4) by $c_{Tj} Q_j^*(t)$, with the parameter estimates then representing complier effects under completely random non-compliance. Now if the expected compliance in the screening program round j is c_{Pj}, the mortality impact of this round can be projected simply as $c_{Pj} Q_j^*(t)$, with the compound impact given by formula (5.4). We demonstrate this approach in the example of Section 5.4.2. A full

treatment of possibly non-random noncompliance in our modelling framework is a topic for a further paper.

5.4 Examples

5.4.1 The US National Lung Screening Trial

We illustrate our methods using data from the US National Lung Screening Trial [72], which compared lung cancer mortality among 53,454 heavy smokers randomized to either low-dose CT scans or chest X-rays. The screening regimen in the trial comprised three annual rounds, the first one soon after randomization, with a reported 20% cumulative mortality reduction in the CT arm after 7 years of follow-up. We, on the other hand, are interested in the mortality reductions that would be produced by a sustained screening program targeted to such high-risk individuals.

A very parsimonious model still producing a reasonable reduction pattern for a single round of screening can be obtained by fixing $\beta = 2$ in (5.8), giving a two-parameter model based on the χ^2-kernel. The fitted reduction curve due to one round of screening is shown in Figure 5–2E. Since the individual-level data from the trial were provided to us by the National Cancer Institute, we could fit this model to both the exact times of death (Equation 5.9) and the yearly and half-yearly aggregated numbers (Equation 5.10). The fitted curves due to three screenings are presented in Figure 5–3A, which suggests that the aggregated numbers are near-sufficient statistics for the mortality reduction: the curves fitted to aggregated data are almost identical to the individual-level fit. The maximum mortality reduction produced by the three rounds of screening is around 20%, which fades after the screening was discontinued. However, the *projected* reduction pattern in Figure 5–3B based on 10 rounds of

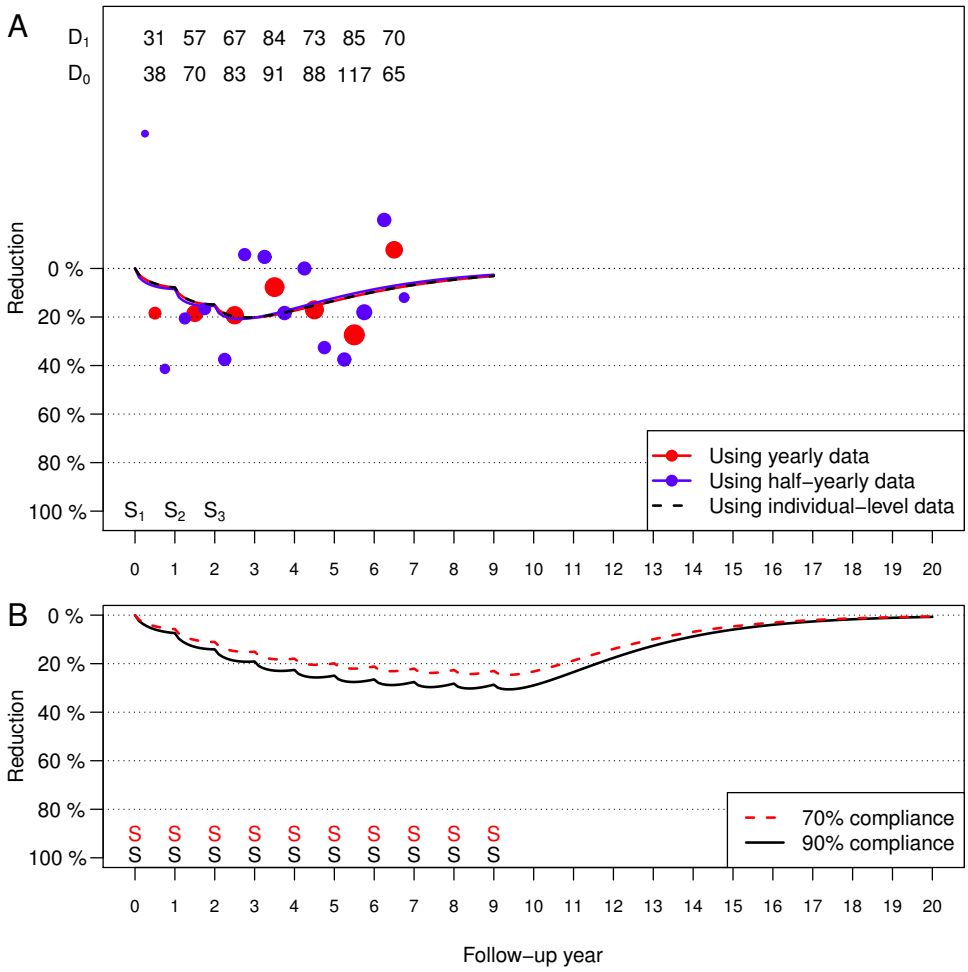


Figure 5–3: Panel A: Empirical and fitted mortality reductions based on individual-level, as well as aggregated yearly and half-yearly, data from the National Lung Screening Trial trial. The size of each dot is proportional to the information contribution of the empirical year-specific mortality ratio. Panel B: Projection of time-specific lung cancer mortality reductions that would be generated by 10 years of annual CT (versus chest X-ray) screening.

annual screening and 90% compliance demonstrates that the mortality reductions would plateau at a nadir of around 30%, should the screening be continued long enough.

5.4.2 The Minnesota Colorectal Cancer Screening Study

Shaukat et al. [84] reported that the mortality from colorectal cancer in the screening arm with 11 annual and 6 biennial fecal occult-blood (FOB) tests is 32% and 22% lower than that in the non-screening arm, respectively. The study involved 46,551 participants equally allocated to the three arms and followed up for 30 years. These mortality reductions were achieved despite a 4-year funding-related hiatus in screening, and averaging over the entire 30-year follow-up. Presumably, the reductions would have been larger without such an interruption.

To study this, we extracted the yearly numbers of deaths from the published figure of cumulative colorectal cancer mortality, and present the observed and fitted mortality reductions in Figure 5–4A. The fitted model was specified using the parametrization (5.7), and the pattern of reduction due to one round of screening is shown in Figure 5–2F. While not obvious in the cumulative mortality curves [Figure 1 of 84], our fitted ones exhibit a W shape, showing the lagged responses to the two phases of screening: after a delay of some years, a nadir of around 50% reduction for annual and 30% for biennial schedule were reached before reverting to what they would be in the absence of screening; this pattern is repeated when screening was resumed.

Figure 5–4B shows the projected reductions due to 16 years of continuous (annual and biennial) screening. The time patterns generated by these two regimens are similar in that benefits start to emerge 6 years after the initiation of screening, continue to manifest through years 7-12, reach the nadir in year 13 and continue onwards. However, the projected sustained reduction is close to 60% for annual

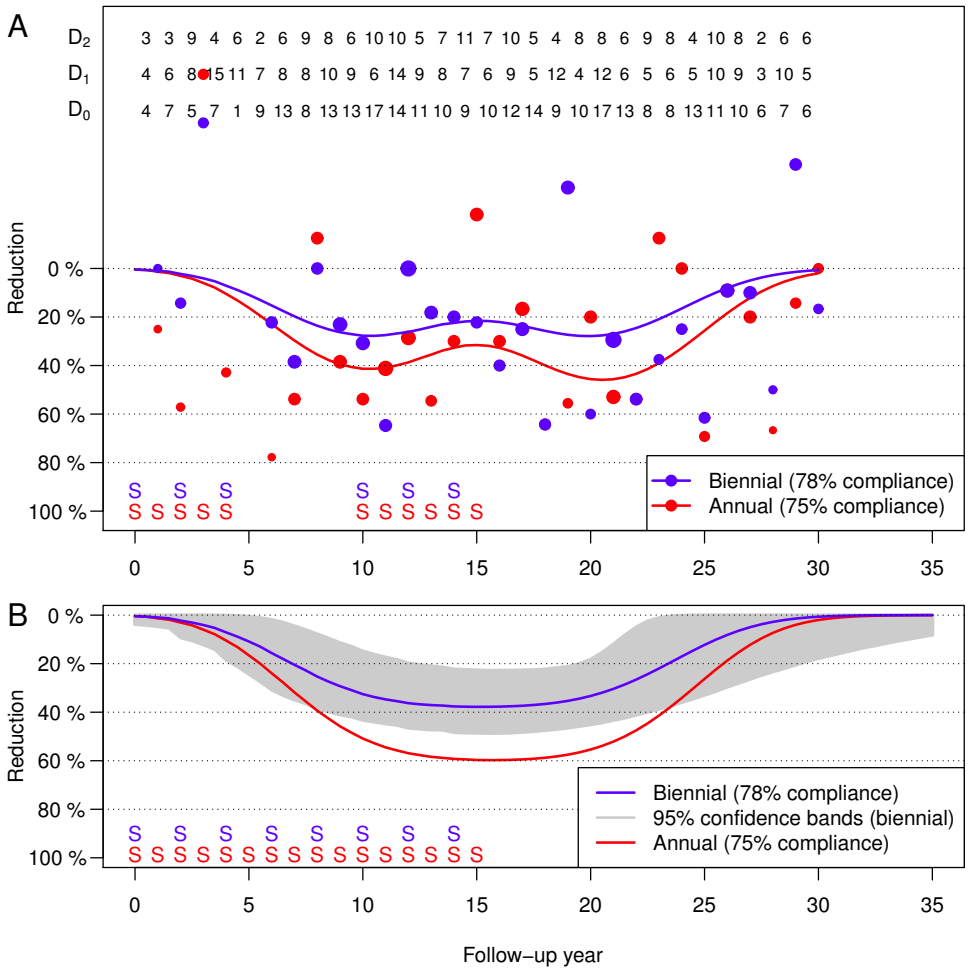


Figure 5–4: Panel A: Empirical and fitted mortality reductions based on the yearly numbers of colorectal cancer deaths in the two screening arms of the Minnesota Colorectal Cancer Screening Study, with the 4-year hiatus. The size of each dot is proportional to the information contribution of the empirical year-specific mortality ratio. Because the hiatus was in calendar-time rather than follow-up time, and entries were staggered, the timing of the screens, each denoted by an S, is only approximate. Panel B: Projection of yearly mortality reductions in colorectal cancer that would be generated by 15 years of uninterrupted annual and biennial fecal occult blood screening. The grey area represents time-specific 95% confidence bands under the biennial screening regimen.

screening and 40% for biennial in the time-window affected, assuming the same compliance rates, 75% and 78% (annual and biennial, respectively, [55]), as in the trial. The 95% time-specific confidence bands in Figure 5–4B are obtained, as outlined in Section 5.3.3, for the biennial regimen based on 10,000 random draws.

5.5 Discussion

Although we did not make distinctions between the impact pattern of the first round of screening and that of the subsequent ones, more parameters could easily be added for modelling the effect of the first, or prevalence, screen, provided that there are sufficient data to enable estimation of the added parameters. For instance, one option would be to model the maximal reduction γ as a (decreasing) function of time, for instance $\text{logit}\{\gamma(t)\} = \gamma_1 + \gamma_2/t$. Another option, motivated by the FOB testing for colorectal cancer, would be to employ six parameters to characterize two modes, corresponding to immediate and remote mortality impact of removing colorectal cancers and polyps, respectively. However, our experience is that the parametric models should be fairly simple to ensure identifiability of the estimation problem; this is not the case if similar reduction patterns can be produced by multiple parameter combinations.

Instead of explicitly modelling sensitivity of the screening examinations or the effectiveness of the subsequent treatment, we concentrate on modelling the probability of being helped by screening associated early treatments; this is because the former two are included in the latter. For instance, a given probability of being helped resulting from a high sensitivity of detecting the cancer combined with ineffective treatment is not distinguishable from one resulting from a low sensitivity

combined with an effective treatment. In particular, estimating sensitivity of the screening would be problematic, since the true positives are those screen-detected cancers which would eventually have proven to be fatal in the absence of screening, making the true disease status inherently unobservable. Our conditional approach circumvents the overdiagnosis problem by focusing on cancer deaths instead of cancer diagnoses.

To summarize, our conditional approach addresses the mortality impact directly, by parametrizing the time-specific conditional probability of being helped by screening given that the cancer would have proven fatal otherwise, which under the assumptions stated in Section 5.2.3 is equivalent to the time-specific mortality reduction, a quantity estimable from trial data. By fitting our model to data from lung and colorectal cancer screening trials, we illustrated how the parameter estimates can be used to project and compare reduction curves that could be produced by long-term screening programs. Our methods can provide policy makers and funders more relevant evidence on how effective cancer screening programs are and could be.

Acknowledgement

This research was funded by the Canadian Institutes for Health Research [grant number 115204]. The authors thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

CHAPTER 6
More on the National Lung Screening Trial

Preamble to Manuscript 3. In this manuscript we discuss in depth the US National Lung Screening Trial, and reanalyze the data, with the particular focus on the effect of the length of the follow-up period, and time-specific, rather than cumulative, mortality reductions. We also compare inferences from aggregate mortality data to inferences from individual-level mortality data and observe that the individual-level data do not add much information.

More on the National Lung Screening Trial

Zhihui (Amy) Liu, James A Hanley, and Olli Saarela

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, Quebec H3A 1A2, Canada.

Unsubmitted manuscript.

6.1 Background

The National Lung Screening Trial (NLST) in the United States, launched in 2002, set out to test whether the earlier treatments prompted by three annual screenings with low-dose helical CT (the study arm), compared to three annual screens of standard chest X-ray (the control arm), could reduce mortality from lung cancer in heavy smokers aged 55-74 years. This is the largest and most expensive randomized trial that the National Cancer Institute (NCI) has ever conducted; more than 53,000 persons across 33 US centres were enrolled into the study and followed up for an average of 6.5 years, with a cost of over 250 million dollars.

In August 2011, nine years after the NLST had been launched, the NCI published the initial findings in the New England Journal of Medicine (NEJM). The main result was that there was an approximately 20% reduction in lung cancer mortality in the CT arm compared to the X-ray arm over about 6.5 years of follow-up [72].

The aim of the NLST was to establish whether three CT screenings would be statistically significantly different from three X-ray screenings. The size of the trial and the length of the follow-up were decided based on the Hu-Zelen model [41], to result in 90% power to detect a 21% reduction in lung cancer mortality [73]. However, such a criterion applied as a stopping rule could prevent quantifying the full effect of early detection and early treatment. While many are pleased with the statistically significant result, we suspected that more information could be revealed by examining the time-specific data, rather than focusing on this one-number reduction cumulated over the available follow-up period.

At the time of the NEJM report, readily available were only the total numbers of lung cancer deaths in the two arms up to January 15th, 2009 (i.e. the official cutoff for mortality analysis) and the two cumulative mortality curves (their Figure 1B). Like what we did with the mammography screening trials [38], by digitizing the cumulative curves, we extracted the yearly numbers of lung cancer deaths from the figure in the NEJM report. These have been published in Table 1 of Liu et al. [53].

The time pattern of these mortality data was puzzling and incomplete. We could not confidently interpret the mysterious 42% mortality reduction in year 6 – for the first 5 years the annual reductions were all around 15%; the numbers of deaths in year 7 were too small (single-digit) to be of use, because the majority of the deaths in year 7 had not been adjudicated. If the 42% deficit in year 6 were to be followed by two similarly large deficits in years 7 and 8, then it would suggest that a screening program could achieve an asymptote twice the size of the 20% reduction reported. If, instead the deficit in year 6 were to be followed by two small deficits of the sizes seen in the first 5 years, it would suggest that the deficit in year 6 was merely a statistical aberration, and that the asymptote would be much smaller than 20%. These two hypothetical scenarios are presented in Figures 6–2 and 6–3, which should be contrasted to Figure 6–1.

6.2 New data available

It is striking how little progress has been made in the past century towards data-sharing. Sir Francis Galton wrote in 1901 [30],

> I hope moreover that some means may be found, through its efforts, of forming a manuscript library of original data. Experience has shown the

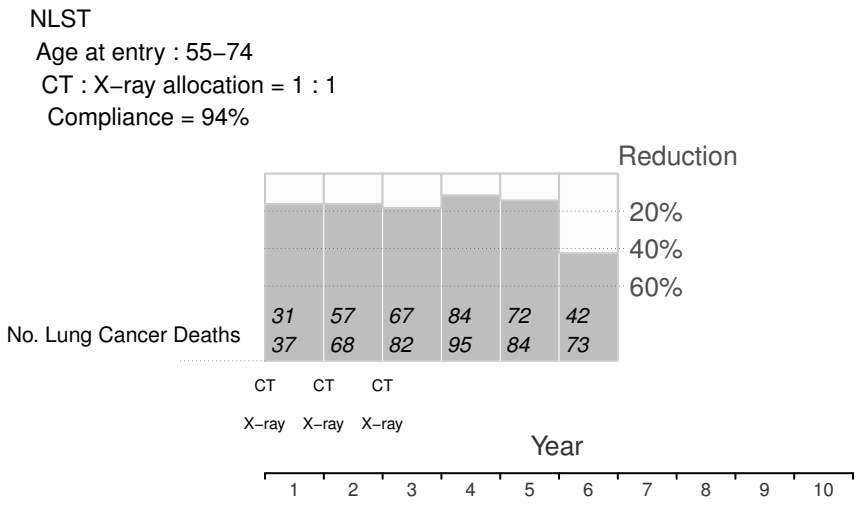


Figure 6–1: NLST yearly numbers of lung cancer deaths, extracted from published NEJM report.

> advantage of occasionally rediscussing statistical conclusions, by starting from the same documents as their author. I have begun to think that no one ought to publish biometric results, without lodging a well arranged and well bound manuscript copy of all his data, in some place, where it should be accessible under reasonable restrictions, to those who desire to verify his work.

In screening trials, the cumulative measures often hide the reduction patterns over time, and because of this, we have been on a 'campaign' calling on trialists to report (at least) the yearly numbers of cancer-specific deaths (as opposed to just a cumulative mortality reduction over some arbitrary follow-up window), if disclosing the individual-level data is not possible [53]. From our experience, the aggregated numbers are in fact 'near-sufficient' statistics. Moreover, even if yearly counts are not

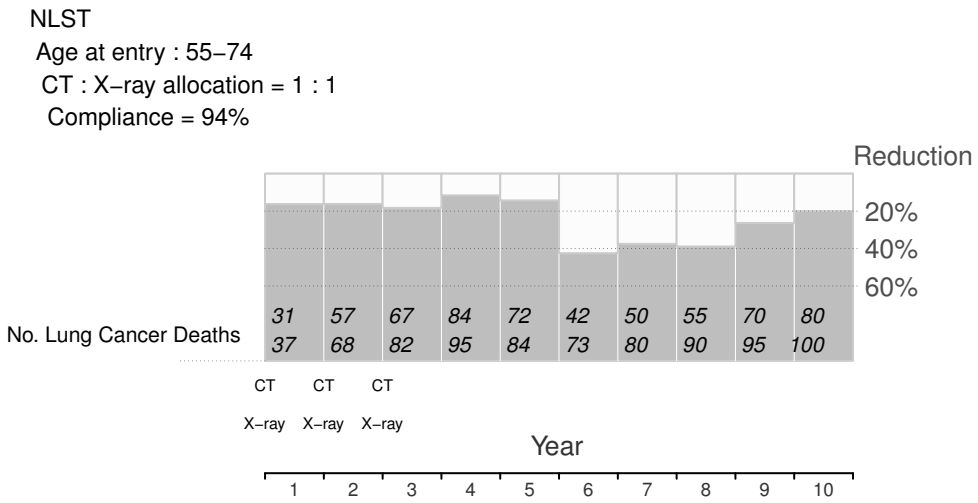


Figure 6–2: NLST yearly numbers of lung cancer deaths, with relatively large hypothetical reductions in years 7-10.

obtainable, we proposed to combine information from multiple trials by adding up the trial-specific log-likelihoods to obtain an overall log-likelihood for more accurate parameter estimates, which avoids sharing neither the individual-level data or the aggregated data. This idea was presented by Hanley at the Statistical Society of Canada Annual Meeting in 2012.

Having said this, the NCI generously reached out and made their individual-level data available to qualified investigators in early 2013. We immediately took advantage of the offer. Information on all of the 53,452 randomized persons (26,722 in the CT arm and 26,730 in the X-ray arm) was recorded in the `patient` file. Using the following variates: number of days from randomization to the end of follow-up (`fup_days`), number of days from randomization to death (`death_days`) and cause

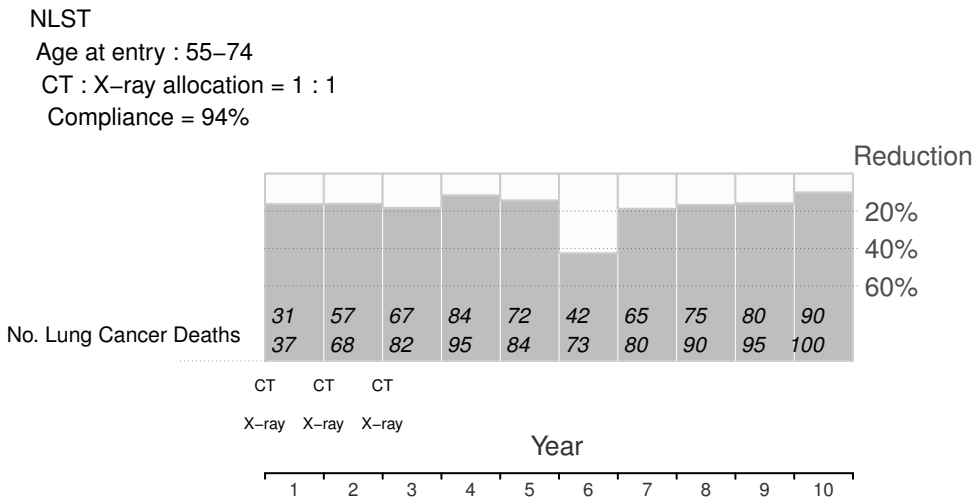


Figure 6–3: NLST yearly numbers of lung cancer deaths, with relatively small hypothetical reductions in years 7-10.

of death (`finaldeathLC==1` for death from lung cancer), we can make a population-time plot to illustrate the study base and some key features of the trial. Figure 6–4 shows that (i) the randomization ratio was 1:1, and the amount of population time was similar between the two arms, (ii) the median length of follow-up was about 6.5 years and most people were alive by the end of the follow-up; and (iii) although these smokers (compared to the general population) may have an elevated risk of dying from lung cancer, in absolute terms, lung cancer mortality was still quite low in both arms – there were a total of 1,019 deaths from lung cancer over the entire follow-up, 467 in the CT arms and 552 in the CXR arm. Thus, the empirical 6.5-year risk ratio of $(467/26722)/(552/26730) = 0.846$ and the mortality rate ratio of $(467/171{,}412)/(552/170{,}355) = 0.841$ are very close. The cumulative mortality reduction from lung cancer over 6.5 years is $1 - 0.846 = 15.4\%$.

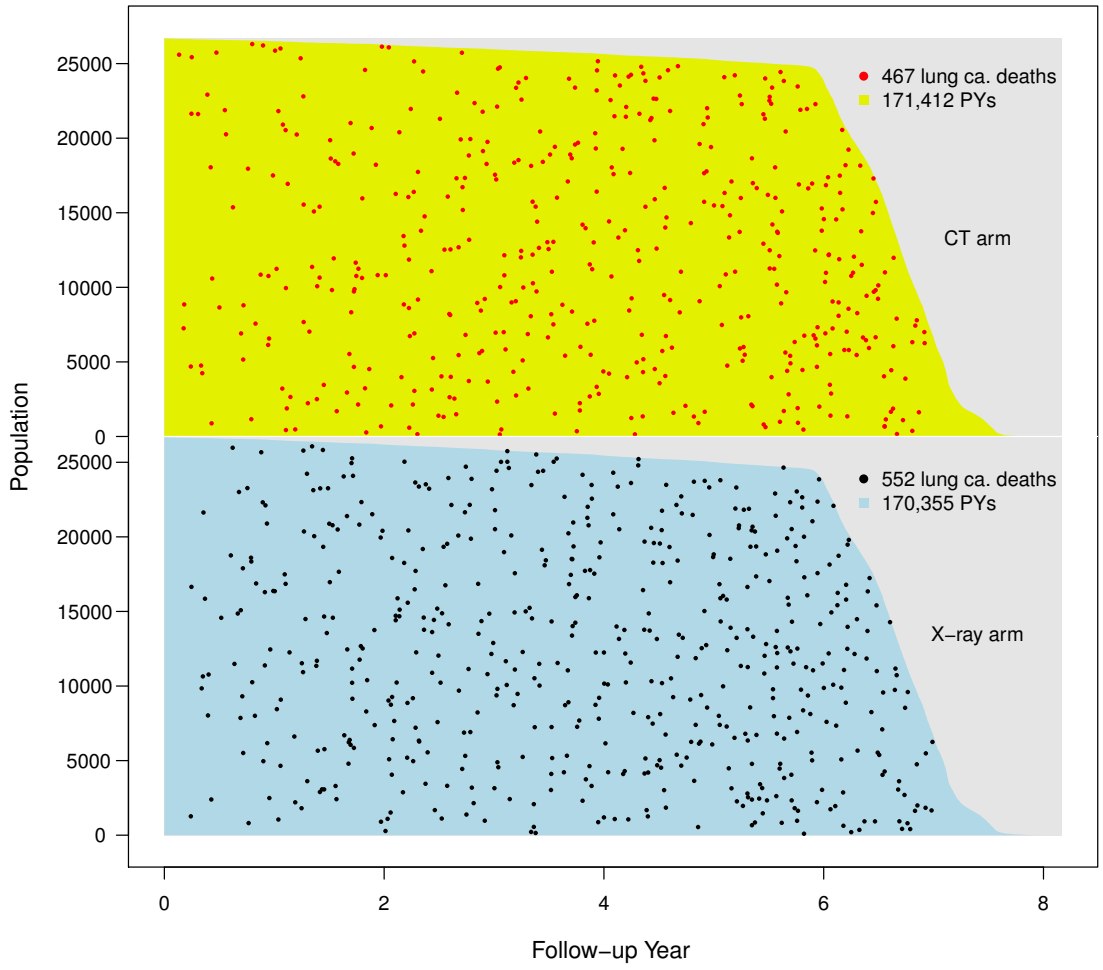


Figure 6–4: NLST number at risk for the two arms, along with lung cancer deaths, using the individual-level data provided by the NCI.

Checking the yearly numbers that we extracted in Table 6–1(a) against those calculated from individual-level data in Table 6–1(b) was one of our first tasks, by including lung cancer deaths before the cutoff date only. They were almost identical, only differing by a few deaths. Next we *included* lung cancer deaths also after the

Table 6–1: Yearly numbers of lung cancer deaths in the NLST. Part (a) was based on our extraction from the NEJM report, (b) and (c) are based on the individual-level NLST data; in (b) only deaths that occurred before the cut-off (i.e. January 15th, 2009) were included, and in (c) all deaths occurred before and after the cutoff date were included.

(a) Year-specific data extracted from figure in NEJM report

Follow-up Year:	1	2	3	4	5	6	7	Total
Screens	↑	↑	↑					
X-ray Arm:	37	68	82	95	84	73	4	442
CT Arm:	31	57	67	84	72	42	3	354
Reduction:	16%	16%	18%	12%	14%	**42%**	25%	20%

(b) Year-specific data including deaths before the cutoff only

X-ray Arm:	38	70	83	91	88	74	4	448
CT Arm:	31	57	67	84	72	45	3	359
Reduction:	18%	19%	19%	8%	18%	39%	25%	20%

(c) Year-specific data including deaths before and after the cutoff

X-ray Arm:	38	70	83	91	89	116	65	552
CT Arm:	31	57	67	84	73	85	70	467
Reduction:	18%	19%	19%	8%	18%	27%	-8%	15%

cutoff in Table 6–1(c), which confirms our earlier suspicion that the counts were incomplete starting in year 6. Now with the additional mortality data in year 7, the reduction in year 6 turns out to be less dramatic – 27% instead of 42%. Although not all deaths in year 7 had been adjudicated, if a similar fraction of deaths were adjudicated between the two arms, then it would suggest that the signal had been fading away by year 7.

The description of all the variables can be found in `participant.dictionary.d091412.rtf`, while the ones we use in this chapter are included in Table 6–2.

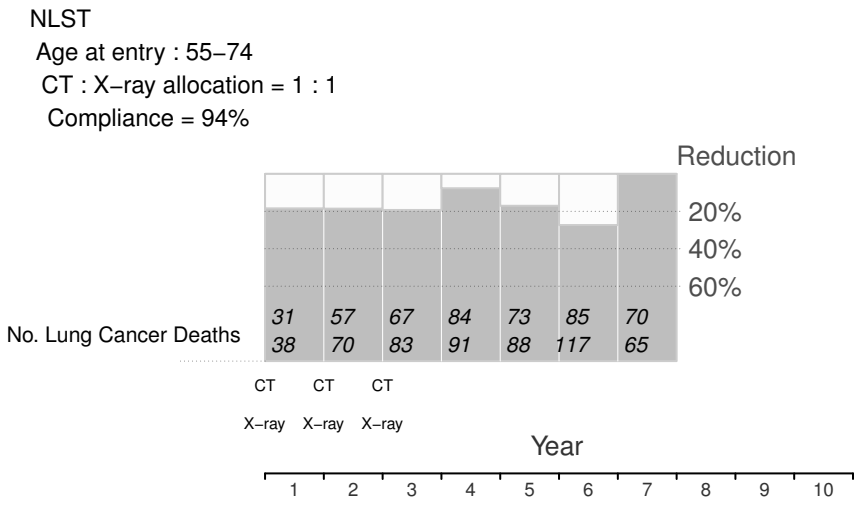


Figure 6–5: NLST yearly numbers of lung cancer deaths, corresponding to table 6–1(c).

Table 6–2: These are the only variables needed for our model fitting, with descriptions provided by the NCI participant dictionary.

`pid`:	patient ID, one per patient
`rndgroup`:	a binary variable indicating to which arm the participant was randomized, 1=CT and 2=X-ray.
`age`:	age at randomization, in years.
`death_days`:	days from randomization to death.
`finaldeathLC`:	the authoritative variable indicating whether lung cancer was the cause of death.
`deathcutoff`:	a binary variable indicating whether deaths occurred before the cutoff for the official final analysis of lung cancer mortality.

6.3 Methods

The trial involved only 3 annual CT screenings, taking place at randomization, the beginning of year one and the beginning of year two. By year seven, the impact on mortality had started to fade. If a screening program were to be implemented in a population in Canada or the US, it would aim to screen regularly for a longer period, say 10 years. In this section, we (i) describe a two-parameter formulation to characterize the impact of one round of screening, (ii) show how much more information is gained by using individual-level data instead of the aggregated (e.g. yearly) ones, (iii) use the parameter estimates to project how large the expected impact would be with 10 rounds of annual CT screening, and (iv) illustrate how to access screening benefits for subgroups.

We observed that using the three-parameter formulation of Chapter 5 in the lung cancer context made the estimates for the location and scale parameters highly correlated, indicating that a two-parameter model would be sufficient. Compared with the colorectal cancer application in Chapter 5, we presume that this is because lung cancer progresses and kills faster, reducing the need for a separate location parameter characterizing the delay. Thus, to avoid identifiability concerns, we considered parsimonious two-parameter models to parametrize the impact of a given round of screening j. One such possibility would be the shape of a half-ellipse, that is, the function

$$Q_j(t;\alpha,\beta) \equiv \alpha\sqrt{1-\frac{(t-\sum_{l=1}^{j}\Delta_l-\beta)^2}{\beta^2}},$$

where α represents the maximum reduction due to the jth round, and β the time to reach the maximum reduction since that round. However, these kinds of non-smooth

curves ($Q_j = 0$ when $t > 2b$) lead to non-smooth log-likelihoods, and consequently, maximization via a Newton-Raphson type of algorithm is not possible.

A smoother 2-parameter formulation based on the χ^2-kernel could be

$$Q_j(t;\gamma,\nu) \equiv \gamma \left\{ \frac{t - \sum_{l=1}^{j} \Delta_l}{\nu - 2} \right\}^{\nu/2-1} \exp\left\{ \frac{\nu - t - \sum_{l=1}^{j} \Delta_l - 2}{2} \right\},$$

$0 \leq t < \infty$, where γ is the maximum reduction due to the jth round and $\nu > 2$ the average time to reach the maximum reduction since that round. This is equivalent to a formulation using the three-parameter gamma kernel (Equation 5.8) by fixing the scale parameter to be 2.

As shown by Figure 5–3A in the previous chapter, knowing the exact times of death does not add more information than knowing just the counts of deaths every year or every 6 months. No matter whether we use the aggregated data (yearly or half-yearly counts of lung cancer deaths) or the individual-level data, the fitted reduction curves are almost exactly the same. The estimated γ (%) and ν (years) and their standard errors are presented in Table 6–3.

Table 6–3: The two parameter estimates and their standard errors from our fitted model based on a χ^2 kernel. The results are very similar no matter which format of data were used: yearly, every 6 months, or individual-level.

Format of data	$\hat{\gamma}$ (SE) %	$\hat{\nu}$ (SE) years
Yearly	8.6 (4.4)	3.38 (1.81)
Half-yearly	9.0 (5.8)	3.12 (2.15)
Individual-level	8.5 (4.5)	3.38 (1.88)

Figure 5–3B gives us a view of what would happen with 10 years of annual screenings. The compliance in the trial was over 90%, thus if a similarly high level of compliance were maintained in the program, then the maximum reduction would

be around 30%, which doubles the 15% reduction achieved with three rounds of screening in the trial.

One could easily study whether a younger or older age group would benefit more from early detection, by choosing data on those aged, say, 65 years or younger at randomization. The fitted curve and the corresponding projection based on 10 rounds of annual screening are shown in Figure 6–6. Our choice of the age group is rather arbitrary, but this serves as an illustration for other subgroup analyses, such as splitting by gender, ethnicity group, medical history and so on.

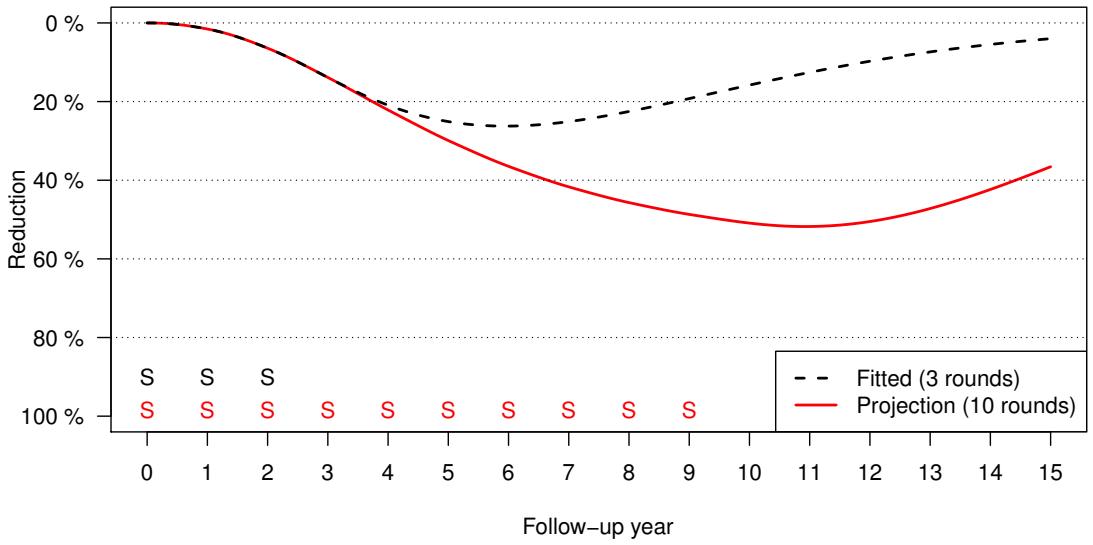


Figure 6–6: Fitted reduction curve (dotted, black) based on the NLST data for persons aged below 65 at onset of screening and projected curve based on 10 rounds of annual screenings.

6.4 Discussion

We had not used the individual-level NLST data when we initially submitted the manuscript of Chapter 5 to a statistical journal, and an associate editor argued that "Good science surely would mean retrieving the individual patient records rather than trying to draw conclusions from summary reports". Leaving aside that we are (in the business of) reporting statistical evidence, rather than 'drawing conclusions' (which is not even possible in the presence of incomplete information), now that we have obtained the individual-level data, we could demonstrate that it does not add more information by using individual-level data instead of just the aggregated (e.g. yearly or half-yearly) ones, which was not surprising to us.

The most common types of lung cancers are non-small cell lung cancer (less aggressive) and small cell lung cancer (more aggressive). With information on histopathology available, one may be interested in knowing which type would benefit more from an early detection program and by how much. We chose not to do so, because histology is an outcome variable rather than a baseline characteristic, and it is not revealed until after at least one round of screening. It is not clear how the information concerning cell type can help the implementation of a screening program.

The empirical 6.5-year risk of dying from lung cancer for a heavy smoker screened with standard chest X-rays is $552/26{,}730 = 2.07\%$ (i.e. 21 in 1000), and is $467/26{,}722 = 1.75\%$ (i.e. 18 in 1000) with 3 annual CT screenings. The corresponding risk difference between the two is 0.32% (3 in 1000), and therefore approximately $1/0.32\% = 313$ people would need to undergo 3 rounds of CT screening to avert one lung cancer death over 6.5 years compared with standard X-rays. (As the previous analysis

indicated, all the mortality benefits had already manifested within the follow-up period of the trial, so this calculation would not be affected by extending the follow-up period.) Say that on average the cost for a low-dose CT screening is $300. Then a total of $\$300 \times 313 \times 3 = \$281,700$ would be needed to avert one lung cancer death, without having included any costs associated with the diagnosis following a positive screening and treatments after the diagnosis. Assume further that those who would benefit from CT screening gained on average 10 years of survival. This back-of-the-envelope calculation results in a cost of 3 annual rounds of CT screening alone of $\$281{,}700/10 = \$28{,}170$ per life year gained, without having included costs for associated diagnosis and treatments. A similar calculation was carried out by Miettinen [59] in 2000, before the NLST was initiated; his assumed inputs resulted in a cost of $10,000 per life-year saved, which was noted to be "well within the range of practice-acceptability". If, however, the survival gain in our calculation was only 5 years on average, the cost per life-year gained would be $56,340, which may exceed acceptable cost-effectiveness. Continued CT screening after 3 years would likely further increase the relative cost since the costs directly add up for each additional round of screening but mortality reductions plateau (as shown in Figure 5–3B).

CHAPTER 7
Recovering the Raw Data Behind a Non-parametric Survival Curve

Preamble to Manuscript 4. We wish that everyone would be as generous as Francis Galton was 113 years ago. In the inaugural editorial in Biometrika, he suggested a 'manuscript library' for exchange of raw data; unfortunately few today do.

When the raw data are not available from authors, several methods/tools have been proposed for extracting data from survival curves using digitizing software. Instead of using a digitizer to read in the coordinates from a raster image, we propose directly reading in the lines of the PostScript file of a vector image.

We have provided R code for importing PostScript files. To demonstrate the practicality, we include several worked examples in the manuscript and many more on the accompanying website. Our methods are more relevant today, as virtually all articles are available online and many graphs are vector images thus can be converted to a PostScript file. Compared with previous approaches, one advantage of ours is that there is no need to repeat the digitization process, that is, the extraction is completely replicable, and the information more precise.

We hope that readers will find the software examples and our website useful. And, for the skeptics, we provide some numerical examples, and a formal error analysis.

Recovering the Raw Data Behind a Non-parametric Survival Curve

Zhihui (Amy) Liu, Benjamin Rich, and James A Hanley

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, Quebec H3A 1A2, Canada.

Revision invited in September 2014.

Abstract

Background: Researchers often wish to carry out additional calculations or analyses using the survival data from one or more studies of other authors. When it is not possible to obtain the raw data directly, reconstruction techniques provide a valuable alternative. Several authors have proposed methods/tools for extracting data from such curves using digitizing software. Instead of using a digitizer to read in the coordinates from a raster image, we propose directly reading in the lines of the PostScript file of a vector image.

Methods: Using examples, and a formal error analysis, we illustrate the extent to which, with what accuracy and precision, and in what circumstances, this information can be recovered from the various electronic formats in which such curves are published. We focus on the additional precision, and elimination of observer variation, achieved by using vector-based formats rendered by PostScript, rather than the lower-resolution image-based formats that have been analyzed up to now. We provide some R code to process these.

Results: If the raster based images are available, one can reliably recover much of the original information that seems to be "hidden" beneath published survival curves. If the original images can be obtained as a PostScript file, the data recovered from a PostScript file can then be either input into these tools, or processed directly. We found that the PostScript used by Stata discloses considerably more of the data hidden behind survival curves than that generated by other statistical packages.

Conclusions: When it is not possible to obtain the raw data from the authors, reconstruction techniques are a valuable alternative. Compared with previous approaches, one advantage of ours is that there is no observer variation thus no need to repeat the digitization process, that is, the extraction is completely replicable.

7.1 Background

Researchers often wish to carry out additional calculations or analyses using the survival data from studies of other authors. Since it is not always possible to obtain the raw data directly from the authors, one is forced to make do with the information that can be recovered from the articles. The researchers differ in their reasons for obtaining such data, and in the number of studies involved. Thus, to motivate this paper, we first briefly recount three of our own experiences. They focus on randomized trials of cancer screening, where the mortality deficits produced by cancer screening are delayed. Thus, a sequence of time-specific hazard ratios (i.e., a rate ratio 'curve') that accommodates this delay is more appropriate than the single-number hazard ratio typically reported by trialists. However, our methodology is applicable to any situation where published data are in the form of cumulative incidence curves, or survival curves, of a step function form.

Hanley [35] re-examined the effect of annual/biennial fecal occult blood screening on the incidence of colorectal cancer in a trial that offered screening for the first 6 years, and, after an unplanned hiatus resulting from a lack of funding, for years 10 to 16. For each arm, he calculated incidence rates in *each* of the 18 years (the original article reported a single overall incidence). To do so, he reconstructed the yearly numbers of cases from (i) the overall numbers of cases in each arm reported in a table, (ii) the curves of cumulative incidence versus year for each of the three arms, and (iii) the arm-specific numbers at risk at years $0, 2, \ldots, 18$ given at the bottom of the figure.

In order to obtain the year-specific mortality rates in five trials, Hanley et al. [38] extracted the yearly numbers of breast cancer deaths in the screening and control arms from the published articles. For two trials, we calculated the yearly numbers of deaths directly from the cumulative numbers of deaths reported in tables. For the other three, we had to back-calculate year-specific numbers of deaths from plots of cumulative numbers of deaths over time.

Hanley [36] re-analyzed the data from the European Randomized Study of Screening for Prostate Cancer (ERSPC), a 7-country and 17-year randomized trial [80]. The re-analysis required the numbers of prostate cancer deaths, and man-years, for each follow-up year in each arm. In the first submitted version, Hanley reconstructed them using the two Nelson-Aalen plots in a figure of the article [80], the numbers of men randomized, and the numbers of men at risk at 5, 7 and 10 years shown at the bottom of the figure. He also contacted the principal investigator, who apologized that it was not possible to provide the exact numbers, but did state that the reconstructed ones were "very close." The first version of the submitted manuscript did not mention this personal communication, and so, quite appropriately, in his review, the Editor wrote:

> But we would need more detail on the methodology of how the estimates for each year were derived. We also need to understand the extent to which the results depend on reading numbers off a fairly small figure taking up about 1/6 of one page in a journal. Is the author confident that the resulting errors are acceptably small?

How confident can one be? Some guidance on data-reconstruction can be found in the meta-analysis literature, since the summaries are not always reported in the way meta-analysts would wish, and since simplifying assumptions, such as a constant hazard ratio, may be inappropriate [34, 75, 43]. Duchateau et al. [23] expressed caution, pointing out that the number of events should not be estimated from the Kaplan-Meier curves for meta-analytic purposes unless virtually no patients are lost to follow-up or censored and there are still many patients at risk in the two groups at the time at which the number of events is to be determined. Other authors have shown that in some circumstances, and by making some assumptions, it is possible to extract additional information. Among the earliest to do so were Parmar et al. [76], who described how to estimate the log of the hazard ratio, and its variance, from the survival curves themselves, rather than from numbers and summaries reported in the text. Although focusing on assessing the accuracies of different techniques for combining published survival curves, Williamson et al. [98] are one of the first to mention using digitized images, obtained by "scanning the survival curves and imported them into the CorelDRAW! 3.0 graphics package." They, and several others since then, have focused on the many practical challenges: Williamson et al. [98] illustrated how information on the numbers at risk may be used to improve the estimation; Tudur et al. [94] reviewed the practicality and value of previously proposed methods; Tierney et al. [93] provided a spreadsheet to estimate hazard ratios and associated statistics from published summary statistics or data extracted from Kaplan-Meier curves. The `grImport` package is intended to add extracted images to R plots, but in the "Scraping data from images" section, Murrell [71]

extracts data from a survival curve and shows that the resulting curve matches the original.

Most recently, Guyot et al. [34] provide a method (and R code) to "derive from the published Kaplan Meier survival curves a close approximation to the original individual patient time-to-event data from which they were generated." They did so "using an algorithm that maps from digitized curves back to KM data by finding numerical solutions to the inverted KM equations, using where available information on number of events and numbers at risk." They assessed the reproducibility and accuracy of several statistics based on reconstructed KM data by comparing published statistics with statistics based on repeated reconstructions by multiple observers.

Increasingly, the figures in electronic publications are *vector*-based and rendered by PostScript, rather than *image*-based. Thus, in this note, we take advantage of this much higher resolution to eliminate the variation introduced by human digitizers, and achieve greater precision and accuracy. The much greater precision can also be used to gain greater detail as to numbers at risk at various time points, and the approach can handle survival curves containing hundreds of steps.

Using worked examples and a formal error analysis, we illustrate the extent to which, and with what accuracy and precision, and in what circumstances, the original information can be recovered from the vector-based and image-based formats in which such curves are published. We describe an R function we use to extract the relevant PostScript data used to draw lines, and to convert the PostScript co-ordinates to co-ordinates in the time-survival $\{t, S(t)\}$ space. If users wish, these can then be used as input to the R software provided by Guyot et al. [34], or the

spreadsheet provided by Tierney et al. [93], or further processed directly by the user. Our own applications have been in estimating yearly event rates using aggregated person time and event counts, rather than in reconstructing individual-level data, but what we describe can be applied to both.

In some instances, it is possible to obtain even more than was visible to the human eye, or a digitizer, and we describe a Stata-specific data-disclosure practice that helped in that respect. Before doing so, we first briefly review the general principles that one can use to derive as much information as possible from a non-parametric survival curve.

7.2 Methods

7.2.1 Principles

To start with, we will assume that the Kaplan-Meier or Nelson-Aalen curve values can be measured with sufficient accuracy and precision (we will relax this requirement in later sections). In such cases, first principles – and some deductions – generally allow one to recover not only (i) the distinct 'event' time t that defines each risk set [we denote the ordered distinct event times by $t_1, t_2, \ldots, t_k$], but also for each risk set (ii) the number at risk n, and (iii) the number of events d. Then, by successive subtractions, one can calculate (iv) the number of observations censored between successive risk sets c. Unless the exact times of censored observations are indicated on the graph, the recovered data can be compressed into the sequence

$$\{n_0,\ c_0,\ t_1,\ n_1,\ d_1,\ c_1,\ t_2,\ n_2,\ d_2,\ \ldots\ \}.$$

If the exact censoring times are indicated on the graph, then in principle, the entire dataset can be reconstructed; otherwise the best that one can do is to use interpolation, together with the description of the recruitment period and closing dates of the study, to impute the locations of the censored observations within the various time intervals. Most authors have spaced them uniformly within these intervals.

To review the principles and illustrate the reasoning, we begin with an small example, using a widely used illustrative dataset. Figure 7–1 (a) shows the Kaplan-Meier estimate of the survivor function for patients with Acute Myelogenous Leukemia (AML) in the 'maintained' group, available in the `survival` package [92] in R. The question at the time was whether the standard course of chemotherapy should be maintained for additional cycles for these patients. To start with, we ask the reader to ignore the additional information we show on each panel, and to limit their attention to the curve, with its steps and censoring marks.

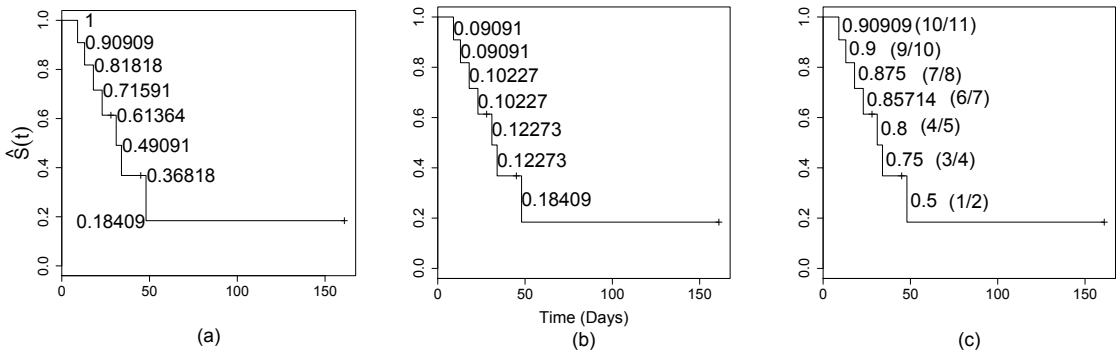


Figure 7–1: Kaplan-Meier estimate of the survivor function, showing the heights and ratios of heights (a) Kaplan-Meier estimate of the survivor function for patients with AML in the maintained group, showing the heights $S(t_j)$; (b) Same K-M curve showing the jumps $J(t_j)$; (c) Same K-M curve showing the ratios of heights $S(t_j)/S(t_{j-1})$. The curve shown in each panel was fitted and drawn using the `survival` package in R.

Let $S(t_j)$ denote the survival probability, or the 'height' of the survival curve, at time t_j, and define the 'jump' $J(t_j)$ as $S(t_{j-1}) - S(t_j)$. We usually would know it, but suppose we do not even know n_0, the number of subjects at time $t_0 = 0$. Without any other information except the step-function values and the times of the steps, how much of the raw information can one recover from such a graph, if the S's are known with sufficient accuracy? (By sufficient accuracy, we mean that the true value can be reliably deduced to be n_j, and not $n_j - 1$ or $n_j + 1$.)

A quick inspection of Figure 7–1 (a) shows that there are 7 jumps and 3 censoring marks, so n_0 is at least 10. Even *without* censoring marks, the differences in the size of the jumps indicate *some* censoring – if there were none, all jumps would be either of equal size $(1/n_0)$, or multiples of this, i.e., m/n_0 if $m > 1$ events in a risk set. As shown in Figure 7–1 (b), $J(t_3) > J(t_2)$, while $J(t_5) > J(t_4)$, and $J(t_7) > J(t_6)$; in addition, since the last observation is censored, we can infer that there must be at least 4 censored values in total.

One way to understand why (single-event) jumps located further to the right can *only be larger* than those that precede them is via Efron's re-distribution-to-the-right algorithm [24]: initially a probability mass of $1/n_0$ is placed at each observation time. Proceeding from left to right, as a censored time is encountered, its mass is redistributed in equal portions to all observations on its right. This procedure of sweeping out the censored observations is repeated until all of their associated masses have been redistributed.

In Figure 7–1 (b), the first two jumps $J(t_1)$ and $J(t_2)$ are of equal size of 0.09091, or 1/11, suggesting that there may have been initially 11 persons at risk (Of course,

without having further information, it could also have been 22 or 33, but subsequent values of the curve will effectively rule these out). The fact that the 3rd jump is bigger establishes that there must be a censored observation at or after t_2 and before t_3. But since (unlike the other censored observations that fall strictly between events times) it is not denoted by a tick mark on the graph, the censoring must, by convention, have occurred *immediately after* the event(s) at t_2, but due to the discreteness of the data, have been recorded as a "t_2+". Thus, while censoring marks may give more precise locations of the censored observations, statistical packages do not necessarily display *all* of them, and so one should not rely on identifying all of them just from the tick marks.

Following Efron's algorithm, $J(t_3)$, of size 0.10227 can be seen to be the sum of the original mass of 1/11 (0.09091) and (1/8)th of the same-size mass associated with the censored "t_2+' observation that was redistributed among the 8 who were at risk just after t_2, i.e., $J(t_3) = J(t_2) + 1/8 \times J(t_2)$. However, the arithmetic and the multiple possible 'legacies' and configurations become complicated, if there are multiple events at the same observed time, or if more than one observation in an interval is censored. Thus, as the expressions for absolute sizes of the jumps start to become complicated, how else might we determine the numbers at risk – and the numbers of events – at the time of each successive jump?

We found it easiest to first assume that each $d_j = 1$, then derive the corresponding n_j, then use any anomalies in the pattern of successive n_j's to revise d_j to a larger integer, and scale the corresponding n_j down accordingly. One way to go from d_j to n_j is to exploit the 'product of conditional survival probabilities' structure of

the K-M estimator: reverse the sequence of products that are used as the estimator, and divide the $\hat{S}(t_j)$ by $\hat{S}(t_{j-1})$. The resulting ratio is $1 - d(t_j)/n(t_j)$, where $d(t_j)$ denotes the number of events at time t_j and $n(t_j)$ is the number at risk at time t_j. If we can establish what $d(t_j)$ is, then we get the *simple expression* for n_j :

$$n(t_j) = \frac{d(t_j)}{1 - \hat{S}(t_j)/\hat{S}(t_{j-1})},\ j = 1, 2, \ldots. \tag{7.1}$$

Indeed, as shown in Figure 7–1 (c), we can infer by using this expression that the numbers at risk at $\{t_1, \ldots, t_7\}$ are $\{n_1, \ldots, n_7\} = \{11, 10, 8, 7, 5, 4, 2\}$.

The initial numbers – which are usually reported in publications – and the sequence of 'fitted' or 'inferred' numbers at risk, can be used to establish with virtual certainty *the number of events at each distinct event-time – the d_j's.* if there indeed is a single event at each distinct event-time, then the inferred numbers at risk will – apart from the (usually small) measurement errors – form a monotonically decreasing sequence. Systematic departures from monotonicity are immediately evident: if there were in fact 2 events at a distinct event-time, the 'fitted' number at risk, n_j, will be 1/2 of what it should be, and will stand out distinctly from its singleton-based neighbours; if there were 3 events, the 'fitted' number at risk will be 1/3 of its neighbours, and so on. We will illustrate this later when discussing the example in Figure 7–2 (right). From the $\{s_1, \ldots, s_7\}$ thus established, and the $\{n_1, \ldots, n_7\}$ we can then by subtraction deduce that in our example $\{c_1, \ldots, c_7\} = \{0, 1, 0, 1, 0, 1, 1\}$.

If the time-spacings between the adjacent t's are relatively short, or if the numbers at risk at specific time-points (e.g,. yearly or monthly) are indicated on the graph, then by further interpolation of the sequence of numbers at risk, the total

amounts of person time for each time-interval of interest can be established with minimal error. Survival plots typically have a width : height aspect ratio larger than 1. Thus, the relative errors will tend to be smaller on the 'time' than on the 'person' dimension of the person-time denominator inputs to the calculated event rates.

The above formula referred to the Kaplan-Meier curve. If instead of the survival curve, the graph shows the Nelson-Aalen estimator of the *cumulative hazard rate function*, given by $H(t_j) = \sum_{t_i \leq t_j}[d(t_i)/n(t_i)]$, then the expression for $n(t_j)$ is

$$n(t_j) = \frac{d(t_j)}{\hat{H}(t_j) - \hat{H}(t_{j-1})},\ j = 1, 2, \ldots. \tag{7.2}$$

It is not always obvious from the label the vertical axis whether an increasing "Nelson-Aalen" curve refers to this sequence of H's, i.e., integrated hazards, or to the cumulative incidence, or risk, i.e., $CI_j = R_j = 1 - \exp[-H_j]$. If indeed it is the latter, i.e., the complement of S, then the formula for n_j becomes

$$n(t_j) = \frac{d(t_j)}{\log[\hat{S}(t_{j-1})/\hat{S}(t_j)]}. \tag{7.3}$$

Until now, we have assumed that the vertical and horizontal co-ordinates of the vertices can be measured with 'sufficient' accuracy. We now turn to what can be achieved using the actual K-M and N-A curves that can be extracted from bitmap images and vector-based graphics in publications.

7.2.2 Practicalities

Just a decade or two ago, it was still common, but time-consuming, to use of the 'pencil and ruler' approach to "read off survival probabilities" [76] from a

(possibly enlarged) hardcopy graph. This practice could involve substantial measurement error, especially when the print was small or the resolution was poor. Today, since most graphs can be either accessed electronically or converted into such a format, the labour intensive work can be reduced, with improved precision and accuracy. In our website `www.biostat.mcgill.ca/hanley/software/DataRecovery` we have collected together a number of graphs found in electronically published articles. Those images are typically of two types, what the Adobe Acrobat documentation refers to as 'raster images' and 'vector objects'.

Raster images

A raster image, or bitmap, consists of pixels (the smallest addressable screen elements in a display device) arranged in a two-dimensional grid. Each pixel, represented by a dot or square, has its own coordinates and color. When one zooms in more and more, the image becomes grainier and the individual dots that make up the lines and symbols on the graph become more evident.

In a black-and-white or grayscale image, white is typically represented by the value 1, black by a 0, and grey by an intermediate value; color images use a more elaborate coding scheme involving multiple channels, such as RGB or CMYK. Just as in digital photography, the larger the numbers of pixels the more faithful the representation of the original values. For an example from prostate cancer screening (a topic to be discussed further below), see Figures 2 and 3 in the article by Andriole et al. [6]. For convenience, we have included this article and its images on our website.

Raster images can be stored in a number of file formats; the most common are .jpeg, .png, .tiff and .gif. They can be generated in a number of ways, such as (i)

scanning the hardcopy and storing it as a raster image, (ii) (if it is in a page of an electronic document) zooming in on the area containing the graph and taking a screenshot, or (iii) (if it is already embedded in a PDF file) using the 'export images' feature in Adobe Acrobat.

The desired points on the graph can be extracted from the image file in one of two ways. The more technical way is to use a programming language such as Basic, C++, or SAS to read the color values into a 2-D array, identify from the colors of the dots the pixel locations of key landmarks (such as the axes intersect, and the furthest apart vertical and horizontal tick marks), and finally determine which sequences of pixel locations contain the dots that make up the curves of interest. Whereas the `ReadImages` package [58] makes it easy to read the array into R, the programming to process the array is still a considerable challenge, particularly for the portions where curves overlap.

The easier way is to use a graph digitizer, a computer program which (i) imports and displays the selected image on the screen and (ii) allows the user to identify horizontal and vertical landmarks by way of the cursor, and to click on as many locations on the graph as are desired, then converts and stores the corresponding (x, y) values. A number of graph digitizers (such as *GraphClick*, *Engauge Digitizer* and *Plot Digitizer*) are available for free on the Web. Guyot et al. [34] report that the software *DigitizeIt* (`http://www.digitizeit.de/`) performed well. Because digitizations of raster images have been covered in detail by Guyot et al. [34], we will not give examples, but merely contrast their accuracy with those of vector images in the theoretical error analysis below.

Vector images

A vector based figure or graph consists of geometrical primitives or elements such as points and lines; it can be identified by the fact that it can be enlarged indefinitely without loss of quality (see examples on our website). Two endpoints of a line are represented by two (x, y) pairs, and a dot by a line of zero length. The 'Post' in PostScript – the most common language for producing them – refers to the principle of device-independence: the elements are rendered in real time from the stored co-ordinates of the elements, regardless of the local hardware on which the software is used. This portability principle underlies the Portable Document Format (PDF), developed by Adobe; PDF files are based on the PostScript language.

The contents of a PDF document are typically stored as a binary file, but both the Adobe Acrobat Pro application, and the Preview application provided in Mac OS, can export a PDF document (or the page of it that contains the graph of interest) as a PostScript file, which contains the commands. Such files tend to be large and contain much technical information, but it is easy (although tedious) to identify the commands that produce the axes, tick marks and the sequence of line segments or dots that make up the K-M and N-A curves.

In PostScript, locations on a page are measured in printer points (72 points per inch) from the upper left corner of the page. Thus, a 2 inch (144 point) x-axis, extending from $t = 0$ and $t = 5$, and physically from 1 to 3 inches from the left side of the page, and located 5 inches (360 points) below the top of the page would be specified by the line segment (72, 360) $\leftrightarrow$ (216, 360). Suppose that the ends of the 1.5 inch (108 points) high y-axis correspond to $S = 0$ and $S = 1$, respectively.

Then from these PostScript co-ordinates, we can determine that the line segment (144, 300) $\leftrightarrow$ (146.88, 300) is a horizontal portion of the step function taking the value $S = (360 - 300)/108 = 0.555$ in the interval $t = (144 - 72)/(144/5) = 2.5$ to $t = (146.88 - 72)/(144/5) = 2.6$, and that the segment (146.88, 300) $\leftrightarrow$ (146.88, 303) is a vertical jump at $t = 2.6$, of length $\Delta S = 3/108 = 0.028$ from $S = 0.555$ to $S = 0.583$.

Surprisingly, some publications include a mix of formats. Indeed, in the publication used as the source of Figure 1 of [34], the axes in the original NEJM figure had been rendered as vectors in PostScript, but the two curves are superimposed as an image. The composite was analyzed as an image by Guyot et al. [34]. By contrast, the other figure in that NEJM publication was rendered entirely in PostScript, albeit with some very complex paths to form the line segments (see website).

7.3 Results

7.3.1 Example presented in full here

Figure 7–2 refers to a study by Pearson and colleagues [31]. 14,264 patients with nonvalvular atrial fibrillation but high risk for stroke were randomly assigned to receive either warfarin or rivaroxaban. The investigators sought to determine whether rivaroxaban was non-inferior to warfarin for the primary end point of stroke or systemic embolism. The published cumulative event rates are shown in the left panel of Figure 7–2. We processed this image by applying our R function to the PostScript file (see website). The right panel in Figure 7–2 shows the highly accurate estimates of the $\{n_j\}$ provided by PostScript data alone. The numbers were derived by applying equation (7.1) to the $S(t_j)$ estimates derived from the PostScript commands.

The numbers at risk at days 0, 120, 840, were reported at the bottom of the figure in the article. Clearly, even if they had not not provided, they could have been very accurately estimated just from the successive $S(t_j)$ estimates alone (the slight lack of monotonicity in series (a) in Figure 7–2 reflects rounding errors in the PostScript co-ordinates.) Moreover, the successive $S(t_j)$ estimates provide accuracy estimated of the numbers at risk at not just at this limited number of time points, but at all time points at which there was at least 1 event. This also shows how a $d_j = 1$ can be reliably distinguished from a $d_j = 2$ or $d_j = 3$, simply by inspection.

In many other graphs like this one, that contain upwards of a hundred steps forming a smooth pattern, we have also been able to obtain quite accurate estimates of the numbers at risk, and thus the numbers of events in time intervals, by smoothing the digitized curves. We consider a few, and refer the reader to our website for more details, and for the R code for this and other examples.

7.3.2 Further examples, elaborated on website

(1) *Colistin for the Treatment of Ventilator-Associated Pneumonia [45].* This report is interesting for two reasons: the fact that despite including this descriptor in the title, it is not a case-control study; and the contradictory information in the Kaplan-Meier curve. The correspondence pointed out that the K-M curves seemed to be based only on those who died, but the authors deflected the criticism by noting, correctly, that "when two or more events can coexist at a specific time, so the drop can be twice as large or more." We leave it to the interested reader to use the JPG files one can export from the pdf file to determine if – as seems to the naked eye – 6 of the jumps in the combination arm in Figure 7–2 are of size 1/8th each, and 1

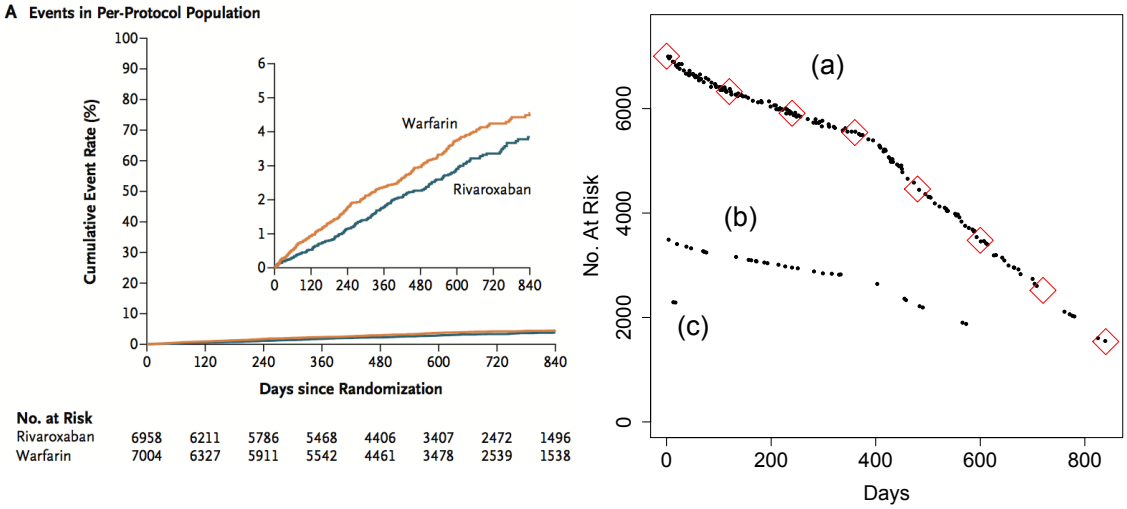


Figure 7–2: (left) Cumulative events rates in atrial fibrillation patients who received warfarin or rivaroxaban. (right) The vertical location of each dot represents the estimated number at risk in the warfarin arm in the risk set in question (horizontal location). The numbers were derived by applying equation (7.1) to the $S(t_j)$ estimates derived from the PostScript commands used to render the vector image. The diamonds represent numbers at risk at days $0, 120, \ldots, 840$, reported at the bottom of the figure in the article. Clearly, even if they had not been provided, they could have been very accurately estimated just from the successive $S(t_j)$ estimates alone. The slight lack of monotonicity in series (a) reflects rounding errors in the PostScript co-ordinates. Each n_j in series (b) is based on the (clearly false) assumption that the corresponding $d_j = 1$; at these distinct failure times, clearly, $d_j = 2$, so each n_j is twice that shown. Likewise the n_j's in series (c) are based on assuming $d_j = 1$, when, again clearly, $d_j = 3$, and the n_j should be three times that shown.

is of size 2/8, at variance with the 11 deaths reported in Table 1, and only possible if all of the 43 - 8 observations were censored before the very first death at day 7 or 8. In this small example, the answers from a digitizer would probably be sufficiently accurate to determine that indeed, the curves seem to be based only on those who died.

(2) *Marriage risk of cancer research fellows [28].* The Lancet recently attempted to match the whimsical nature of the articles in the Christmas Edition of the BMJ, by publishing a 'marriage-free survival' curve in an article. The article began "Research fellows aiming to obtain a PhD or MD/PhD degree face many hazards at work, including exposure to toxic substances and harassment by reviewers of their papers" and lamented the fact that "However, few data exist on the sociocultural risk factors encountered at work – eg, their risk of marriage." The data and the curve provide a useful teaching example, small enough to be worked by hand, and to have students figure out when and how many 'individuals with a bachelor status were censored at the time of analysis." As can be seen in the correspondence on the Website, the authors gladly shared the 13 observations with JH, so that teachers can be spared having to reverse-engineer them in order to check that their students did so correctly.

(3) Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated C-Reactive Protein (JUPITER) [78]. The report of this study has prompted some concerns about how the number needed to treat was calculated, using 5-year risks that were based on survival curves that ended at year 4.5, but that, because of small numbers of events, were quite erratic in years 4 and 5. The projections also raised the issue of whether (as in our screening example) reductions in event rates are immediate, or delayed, and how long they persist after statins are discontinued. The authors did not answer our request that they share just the half-yearly numbers of deaths: we wished to use them, along with the half-yearly numbers of at risk that were included in the Figures, to calculate time-specific hazards and hazard ratios. Fortunately, even though even though the placebo and Rosuvastatin curves

were displayed in a rectangle less than 60 printer's points, or 5/6ths of an inch, tall and just over 1 inch wide, it was possible to use the PostScript commands to quite accurately determine where along the 4.5 year time axis the unique death times were located, and how many there were at each time point.

In order to motivate the formal error analysis, presented in the 'Precision' section, we consider the time-specific numbers of persons at risk in the Rosuvastatin arm: the Figure reported 8.9, 8.4, 3.9, 1.4 and 0.5 thousand persons at risk at the beginning of year 1, 2, 3, 4 and 5. We can use these to test how well they (and the numbers at the times of the events) could been estimated by applying equation (7.1) to the points on the curve; the curve had a vertical range of just 30 printer's points, with the co-ordinates in the PostScript file recorded in increments of 0.001 of a point. Since the *cumulative incidence* curve has a vertical range of 0 to 0.04, it has an effective resolution of $0.04/30,000$. However, since equation (7.1) uses successive ratios of the *complement* of the cumulative incidence curve, the individually estimated numbers at risk have less precision, particularly at the beginning of the curve, where they are larger, and the jumps smaller. Using the R code supplied on the website, one can see that over year 1, the successive ratios lead to estimated numbers at risk with a sawtooth appearance, alternating between approximately 10.3 and approximately 6.9 thousand. However, this noise is easily removed by smoothing, and the mid-year estimate of approximately 8.6 thousand is quite accurate. Moreover, despite the noise at the individual event-times, the times at which there were 2 or more events stand out clearly.

In the Figure in the report of the ERSPC trial [80], the cumulative incidence curve has a vertical range of 0.01 in the probability scale, and 80 on the printer's points scale. Upwards of 100,000 men were at risk in each arm in the early follow-up years. Thus, the resolution of 0.001 printer's points used to render the Figure limits the absolute resolution of the estimates of the numbers at risk, but still ensures small relative errors in these. We return now to how we were able to extract even more precise information from that Figure, and to the circumstances that led us to discover the value of vector-based data-extraction.

7.3.3 An unexpected data-disclosure bonus

Originally, to extract the ERSPC [80] data, JH used Acrobat Reader to zoom in on the Figure so that it filled the screen. He pasted a screenshot of this into the GraphClick software to digitize the two curves. From these, and interpolated numbers at risk for years 1-4, 6, 8, 9 and imputed numbers at risk for years 11 and 12, he was able to compute estimated yearly numbers of deaths and man-years at risk.

In his subsequent pursuit for greater precision, and to answer the Editor, he noticed (and readers of this note will notice) that when the Figure in the ERSPC report is enlarged in Acrobat Reader, the re-drawing takes a surprisingly long time. Even though the total sample size was 162,000 men, there were only 540 deaths, and so, allowing for some multiplicities, there should be even fewer than that many steps in the two step functions. Curiosity prompted him to convert the PDF file to PostScript, and examine how the steps were drawn. To his surprise (and the disbelief of the study epidemiologist who has told him that the curves had been computed

and drawn using Stata but that it was impossible from what was in the Figure to go back from them to what he had requested), the PostScript file contained the exact coordinates of each of 89,308 and 72,837 line segments or dots, one per man! This explained why the curves took so long to be re-rendered by Adobe Reader, and the page to be printed. The horizontal and vertical coordinates of each of these segments/dots thus provided the exact numbers of men being followed at each point in follow-up time, and thus at the exact times of the vertical steps in the curves (corresponding to prostate cancer deaths). The number of prostate cancer deaths at each time point was obtained by multiplying the size of the step by the number being followed at that time. The numbers were then aggregated by year and study arm to produce the counts listed in Figure 1b in the published re-analysis [36].

To illustrate *just how much* data are disclosed by the way Stata makes the curves, we present side by side in Figure 7–3 the original NEJM figure on the left, together with on the right the 'numbers of men at risk' curves that we were able to recreate using the data contained in the PostScript file 'behind' the Figure on the left. The unusual shape of each 'numbers at risk' curve – which we derived from the PostScript data behind the published Figure – is explained by the recruitment method. In the methods section of the NEJM article, we read that, in the Finnish portion of the study,

> men were recruited at the ages of 55, 59, 63, and 67 years. (...) the size of the screening group was fixed at 32,000 subjects. Because the whole birth cohort underwent randomization, this led to a ratio, for the

screening group to the control group, of approximately 1:1.5. (...) Follow-up for mortality analyses began at randomization [January 1 in each of 1996, 1997, 1998 and 1999] and ended at death, emigration, or a uniform censoring date (December 31, 2006).

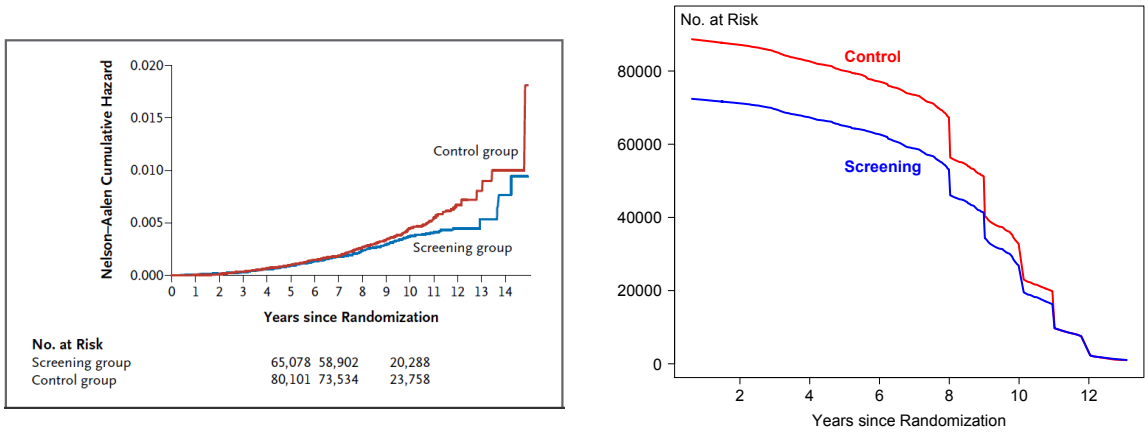


Figure 7–3: (left) Screenshot of the Nelson-Aalen curves in the original NEJM report of the ERSPC and (right) numbers at risk at each time point after randomization, derived from the PostScript file. The large numbers censored exactly at the end of follow-up years 8, 9, 10 and 11 are because the men in the Finnish portion of the trial were randomized on January 1st, 1996, 1997, 1998 and 1999, and were still alive on December 31, 2006. The shallower slope of the curve in years 1-8 is due to deaths, while the steeper slope of the curve in years 9-13 reflects the staggered entries, beginning in different years in the 7 different countries.

The 160,00 data points in the Kaplan-Meier curves in the ERPSC report were produced by an early version of Stata. To test whether the latest version continues to draw each censored observation as an invisible dot on the curve, we used Stata version 12 to construct a Kaplan-Meier curve based on the same AML data we used in Figure 1, and to save it as a PDF file. We then used Adobe Acrobat to export it to a PostScript file, and extracted the line segments (the `.pdf`, `.ps` and `.R` files are

provided on the Website). They reveal that the Stata curve was drawn using 20 line segments – 1 for each of the 7 vertical steps, 1 for each of the 6 horizontal lines for the intervals that do not contain a censored observation, 2 each for the 2 horizontal lines, 2 for the 2 intervals that contain 1 censored observation each, and 3 for the 3 censored observations that do not coincide with a vertical step.

7.3.4 Distortions produced by further processing

Interestingly, in the ERSPC Figure, while the numbers and sizes of the jumps do make sense, the numbers at risk, derived by simply counting how many observations (each one plotted as a dot) exceed the time point in question, do not agree perfectly with those would have obtained from the successive survival ratios described in the "Principles" section above. We traced this discrepancy to the fact that, even when just one death is involved, the jumps implied by the PostScript data are not entirely monotonic, suggesting either some rounding at the time they were generated in Stata, or some post-Stata processing by other graphics software. Given the very large numbers at risk, and thus the very close agreement between the two, the fact that they are Nelson-Aalen rather than Kaplan-Meier curves does not explain the discrepancies. In fact, as can be seen by studying the examples on our website, this post-processing seems to be common, and sometimes results in quite elaborate ways to draw what appear to the eye as simple step functions. In the exemestane for breast cancer study (see website), it took almost 2500 line segments to produce two step functions based on a total of 43 events!

7.3.5 Precision

How precise are the data extracted from raster and vector images? One can assess this question at a number of levels, beginning with the precision of the $\hat{S}$ (or $1 - \hat{S}$) measurements themselves. Consider a typical 300 dots per inch (dpi) *raster* image in which the full (0, 1) S-axis is 1.6 inches, or 480 pixels, high. This gives a resolution of $\Delta S \approx 0.002$. [A 'downwards' curve that ends at say $S = 0.9$, but on a plot that uses the full (0,1) scale, squanders considerable precision: it makes more sense to plot the 'upwards' function, $1 - S$, up as far as 0.1, making the $1 - S$ values accurate to within ± 0.0005.]

Consider instead a *vector* image containing the same curve, on the same 1.6 inch ($= 72 \times 1.6 = 115.2$ points) vertical scale. Because the co-ordinates given in the PostScript file exported by Adobe Acrobat are recorded to 3 decimal places, the resolution is $\Delta S = 1/(115.2 \times 1000) \approx 0.00001$, or 200 times that of the raster image.

While both of these resolutions give adequately precise measures of $\hat{S}$, and allow one to determine how many events are involved in each jump, they may not give such precise measures of the number at risk at each jump, since it is measured as the *reciprocal* of $1 - \hat{S}(t_j)/\hat{S}(t_{j-1})$. As an empirical assessment of the precision of the derived measurements, Figure 7–2 shows the estimated numbers from a raster image and a vector image, along with – as a validity check – the reported numbers at risk at the end of each time interval. They match very well with those given in the articles.

The accuracy can also be quantified using a theoretical error analysis. Consider two adjacent values on the same cumulative incidence curve, where the vertical axis

goes from 0% to 5%, reported (after some rounding) to be y_0 and y_5 points respectively above some landmark; suppose that without rounding they would be Y_0 and Y_5 points above. Denote the vertical locations (similarly rounded) of the two adjacent points on the graph as y' and y'', with $y'' > y'$, corresponding to unrounded values of Y' and Y''. Then the estimates of the number at risk is

$$\hat{n}(t_j) = \frac{20(y_5 - y_0) - (y' - y_0)}{y'' - y'}.$$

In the Appendix, we provide the variance of this derived quantity, assuming that the errors (e's) contained in the four y's are equal and independent of each other. In practice the PostScript points are rounded to 3 decimal places; thus the true location Y associated with a reported location of $y = 563.384$ points lies between 563.3835 and 563.3845 points. If errors are uniform over this 0.001 range such that $\sigma_e \approx 0.001/\sqrt{12} = 0.0003$ points, then the coefficient of variation (CV) is

$$CV[\hat{n}(t_j)] = 100\% \times 2.8 \times 0.0003 = 0.084\%.$$

Similarly, if points are rounded to 2 decimal places, then the corresponding CV is 0.84%. [35]

7.3.6 Software and further examples

The most time-consuming task in extracting the relevant co-ordinates from a PostScript file is visually searching through the file to find the commands that draw lines or dots, and skip the large number of irrelevant commands. We did find that the R package `grImport` imports PostScript images. Its main focus is adding the extracted images to R graphical plots, but the author's webpage gives a reference

[71] where he describes extracting data from a survival curve and shows that the resulting curve matches the original. The package requires Ghostscript and does not handle the PostScript output produced by more recent versions of Adobe Acrobat. Thus, we wrote our own R function. It does not use intermediate software, but extracts the same graphics 'paths' as `grImport` does. The examples in our website show how to use the extracted co-ordinates to easily identify the key tick marks and other landmarks. These are needed to transform the extracted data from mere co-ordinates on a page that is say $8.5\times72 = 612$ points wide by $11\times72 = 792$ points high into the relevant co-ordinates with respect to the $S(t)$ and t axes of interest for further analysis. The R function, and several examples in which we apply it, can be found at `www.biostat.mcgill.ca/hanley/software/DataRecovery`.

7.4 Discussion

The availability of raster based images, and the practical tools provided by authors such as Tierney et al. [93], and Guyot et al. [34] are particularly valuable in recovering the raw data. As they, and now we in our examples here and the ones on our website, have shown, one can reliability recover much of the original information that seems to be "hidden" [23] beneath published survival curves.

A digitizer provides more accurate and precise measures of the jumps or ratios. However the screen itself has limited resolution, and much greater resolution is possible if the original images can be obtained as a PostScript file. The data recovered from a PostScript file can then either be input into these tools, or processed directly.

As we document in our website, some PostScript files contain more information that one would need to draw simple step functions. Thus, in some instances, end-users ('data-scrapers') may have to do some further processing, or select just parts of the overly-elaborate paths used to create lines. We have found that some of the graphics files that authors submit with their manuscripts must have been touched or redrawn by the publishers.

We found many grainy images in some of the best journals, and wonder if authors do not fully appreciate the principle of portability and device-independence of vector based graphics. We urge authors to submit PDF rather than raster images.

The Postscript used by Stata discloses considerably more of the data hidden behind survival curves than that generated by other statistical packages.

7.5 Conclusions

When it is not possible to obtain the raw data from the authors, reconstruction techniques are a valuable alternative. Compared with previous approaches, which use manual digitization of raster images, our method takes advantage of the much greater precision of vector-based images rendered via PostScript. Our extraction is replicable, and eliminates the observer variation that accompanies the digitization process.

Acknowledgements

We thank the Natural Sciences and Engineering Research Council of Canada, Le Fonds Québécois de la recherche sur la nature et les technologies, and the Canadian Institutes of Health for their support.

7.6 Appendix: Error Analysis

If we take two adjacent points on the same cumulative incidence curve and the y axis goes from 0% to 5%, then the estimate of the ratio is $[20(c-d)-(a-d)]/[20(c-d)-(b-d)]$ and thus

$$\hat{n}(t_j) = \frac{20(c-d)-(b-d)}{a-b} = \frac{\mu_1 + e_1}{\mu_2 + e_2},$$

where a and b are the heights of two points on the curve, and c and d are the values corresponding to 5% and 0%, μ_1 and μ_2 are the error-free numerator and denominator, i.e. before any loss of data, and e_1 and e_2 are the errors associated with them, i.e. the observed data with rounding.

Assuming all four error variances are equal to σ_e^2 and independent of each other, then

$$Var[\hat{n}(t_j)] = \frac{\mu_1^2}{\mu_2^2}\Big[\frac{V_1}{\mu_1^2} + \frac{V_2}{\mu_2^2} - 2\frac{C_{1,2}}{\mu_1\mu_2}\Big],$$

where $V_1 = Var[20(c-d)-(b-d)] = Var[20c-19d-b] = (20^2+19^2+1^2)\sigma_e^2 = 762\sigma_e^2$, $V_2 = Var[a-b] = (1^2+1^2)\sigma_e^2 = 2\sigma_e^2$, and covariance $C_{1,2} = C[20c-19d-b, a-b] = \sigma_e^2$.

Further assuming $\mu_1 \approx 20 \times 100 = 2000$ points, $\mu_2 \approx 0.5$ points, and $\hat{n}(t_j) = 4000$, we have

$$Var[\hat{n}(t_j)] = \sigma_e^2 \times \frac{2000^2}{0.5^2}\Big[\frac{762}{2000^2} + \frac{2}{0.5^2} - 2\frac{1}{2000 \times 0.5}\Big] \approx \frac{2000^2}{0.5^2} \times \frac{2}{0.5^2} \times \sigma_e^2,$$

and coefficient of variation

$$CV[\hat{n}(t_j)] = 100\% \times \Big[\frac{2000}{0.5} \times \frac{2^{1/2}}{0.5} \times \sigma_e\Big]/4000 = 100\% \times 2.8\sigma_e.$$

Therefore, if the PostScript points are rounded to 3 decimal places, then 563.384 points probably lies somewhere (uniformly) between 563.3835 and 563.3845, so error range = 0.001 leads to $\sigma_e \approx 0/.001/\sqrt{12} = 0.0003$ and $CV[\hat{n}(t_j)] = 100\% \times 2.8 \times 0.0003 = 0.084\%$.

Similarly, if the PostScript points are rounded to 2 decimal places, then 563.38 points probably lies somewhere (uniformly) between 563.375 and 563.385, so error range = 0.01 leads to $\sigma_e \approx 0/.01/\sqrt{12} = 0.003$ and $CV[\hat{n}(t_j)] = 100\% \times 2.8 \times 0.003 = 0.84\%$.

CHAPTER 8
Summary and Discussion

This thesis presents theoretical and methodological developments for measuring the mortality reductions due to cancer screening. The aim is to give policy makers and funders more accurate evidence on how effective screening programs are and could be.

Our objective is to provide a way to project the time-specific reductions in mortality that would be produced by a sustained screening program, using data from randomized trials. Our contribution is having *formulated the estimand* (as the probability of being helped by screening given that the cancer would prove fatal otherwise), *stated the identifiability conditions* for this, and *proposed a corresponding estimation method*.

Unlike the prevailing approach that models the entire cancer progression, we focus on mortality only and thus avoid specifying parameters such as prevalence, incidence, test sensitivity and specificity, and sojourn time all together, as well as the modelling assumptions associated with each one of them. Also unlike the prevailing approach, we estimate the mortality impact using data from randomized screening trials, and thus our *probabilistic* projection is *evidence-based*.

We believe this is a major shift. Using only a few parameters, data from a single trial (or a set of trials containing variations in regimens) can be used to pursue a universal estimand that has meaning far beyond any specific regimen in any one

trial. It greatly extends the usefulness of the data generated by trials, and allows for extrapolation to other screening scenarios.

Before ending, I address some of the questions and concerns often raised by reviewers of our work.

> *1. Why does it not suffice to compare the cumulative incidences or cause-specific hazards between both arms in a trial, instead of focusing on the proposed parameter which demands more identification assumptions?*

Indeed estimating the mortality reductions in a trial does not require such a modelling effort, but our objective was more ambitious than this. A modelling approach decomposing the overall mortality reduction into the impacts of the individual rounds of screening is necessary for the purpose of projecting the impact of a screening program with more rounds of screening. The cumulative measure is not useful for our projection purpose, since it does not enable one to disentangle the impact of a single round of screening, which is central to our modelling approach. To quote again the first principle of Miettinen and Karp [63, p. 82]:

> The proportional reduction in mortality from the cancer is nothing like a constant over time from the beginning of the screening (for the generally short duration of it) to the end of the follow-up (for an arbitrary duration of it). It thus is logically inadmissible to quantify the reduction by pooling the experience across the entire duration of the follow-up. The proper concern in a trial like this [NLST] is to address the incidence density of death from the cancer as a function of time since the initiation of the

screening. And that function is, of course, different for different durations of the screening.

2. It would be interesting if you could compare with a nonparametric estimate, or an estimate obtained from a (possible piecewise) proportional hazards model.

As we note above, fitting non-parametric hazard ratio curves to the observed numbers of deaths was not of interest to us, since our modelling approach was based on extracting the mortality impact of a *single* round of screening. These in turn were used in the projection task by compounding the decomposed round-specific impacts according to the screening schedule of interest.

3. More discussion is needed on the plausibility of your assumptions and, ideally, also some investigation of the sensitivity of the results to violation of these assumptions.

We list four *identifying* assumptions, which are required to identify the estimand, the probability of being helped in terms of potential outcome variables. If violated, our estimand is simply not identifiable based on the observed data. Speaking of their plausibility, monotonicity and strongly ignorable treatment assignment are well justified, the former stating that screening does not shorten survival, and the latter being satisfied through the randomized assignment in a trial. Also, cancer screening is usually too specific to detect conditions other than the site-specific cancer. The curability assumption is the only potentially contentious one, but agrees with the second principle of Miettinen and Karp [63, p. 81–82]:

Reduction in mortality from a cancer subsequent to screening can occur only if the cancer's treatments under the screening - those early treatments - are more commonly curative than those in the absence of screening. In fact, attainment of enhanced curability - by earlier treatments - is the very idea in screening for a cancer. Thus the parameter of Nature that should be viewed as the proper object of any study on the intended consequence of screening for a cancer is one that meaningfully quantifies the *gain in the cancer's curability rate* when screening-associated early treatments replace the treatments on already symptomatic cases in the absence of screening. This is a proportion of the cases of the cancer that are fatal in the absence of screening, the proportion of these otherwise fatal cases that are curable by screening-associated early treatments.

4. You parametrize your model with the parameters γ, μ, and σ. Do you consider these parameters as meaningful in their own right? If so, it would be helpful with an interpretation of their estimates in the real data example.

These model parameters correspond to a *single* round of screening; although interpretable, they are not of primary interest, and thus we do not report the corresponding estimates. Instead, the end-product of the modelling effort, through compounding the round-specific impacts parametrized in terms of the above parameters, is a bathtub-shaped reduction curve, describing when the reductions begin, how large they are and how long they last. Our proposed model parametrizations should only be considered as suggestions or examples; since the actual inference is

likelihood-based, the model formulation is not restricted by what we propose in this thesis. In general, the appropriate complexity of the model parametrization depends on how much data are available.

> *5. If you do not consider the parameters the parameters γ, μ, and σ as meaningful in their own right, then why not use a more standard logistic parametrization of $Q_j(t)$, e.g. $Q_j(t) = \alpha + \beta(t - s_j)$?*

Although we do not interpret the corresponding estimates, our parametrizations were chosen to produce a biologically plausible reduction curve that would obey the 'curability-detectability tradeoff' of diagnosing cancers. We have explained why the reduction pattern produced by the two parametrizations suggested in Chapter 5 is reasonable in the screening context, in particular, why the impact of a given round of screening is first delayed, achieves a maximum reduction, and then fades away after screening is discontinued. Constant or linear effects over time, for example, are clearly inappropriate.

> *6. When there is noncompliance, please clarify what causal estimand is the interest and show the derivation of $cQ(t)$ in Section 5.3.4 as the estimator.*

When there is noncompliance, we take the potential outcome to correspond to being 'under screening' in the sense of being randomized into the screening arm of the trial, cf. the third principle of Miettinen and Karp [63, p. 82]:

> A quantitatively meaningful etiogenetic study on death from a cancer, with lack of screening for it the etiogenetic factor, can be based on a case and base series from the relevant segment of follow-up in a screening

trial with sufficiently long-term screening (and close adherence to the schedule). In the case series, the relevant history is about whether the person was under screening at the time of the cancer's detection (by virtue of being in the screening arm of the trial, irrespective of whether the diagnosis was derived on the prompting of a positive result of the initial test at issue or due to symptoms emerging between the scheduled tests).

We do not intend to relate the potential outcome to actually being screened; unless otherwise specified, our causal estimand should be understood as an intention-to-treat type of effect. However, we present the $cQ(t)$ formulation for the purpose of the projections, where it makes sense to upscale or downscale the mortality impact of the screening program by the expected participation rate. In addition, a relevant quantity for individual level decision making is the mortality reduction conditional on participation. In particular, we note that multiplying by the constant c only accounts for non-compliance that is completely at random, but we do allow for different non-compliance rates between the different screening rounds.

A full treatment of potentially non-random non-compliance is an important problem, and can be addressed in the proposed framework when data on individual-level screening histories are available, but this would be a topic for further work. Other directions for future research include applying the outlined methods to observational data routinely collected under existing screening programs. The main difference compared to applications using trial data would be the need to control for confounding, since without randomization, those who decide to undergo screening

may be a highly selected group. This in turn would proceed by modelling the probabilities of individual-level screening histories conditional on relevant covariates, and using these as propensity score [79] analogues in removing the confounding due to the observed individual-level characteristics.

References

[1] Aalen, O., Borgan, Ø., Gjessing, H., and Gjessing, S. (2008). *Survival and Event History Analysis: A Process Point of View*. Statistics for biology and health. Springer.

[2] ACR (2014). BMJ Article on Breast Cancer Screening Effectiveness Incredibly Flawed and Misleading. *ACR News*. `www.acr.org/News-Publications/News/News-Articles/2014/Quality-Care/BMJ-Article-on-Breast-Cancer-Screening-Effectiveness-Incredibly-Flawed-and-Misleading`, accessed on Feb. 25, 2014.

[3] Albert, A., Gertman, P. M., Louis, T. A., and Liu, S. (1978). Screening for the early detection of cancer – ii. the impact of screening on the natural history of the disease. *Mathematical Biosciences*, 40:61–109.

[4] Alexander, F. E., Anderson, T. J., Brown, H. K., Forrest, A. P., Hepburn, W., Kirkpatrick, A. E., Muir, B. B., Prescott, R. J., and Smith, A. (1999). 14 years of follow-up from the Edinburgh randomised trial of breast-cancer screening. *Lancet*, 353(9168):1903–1908.

[5] Andersson, I., Aspegren, K., Janzon, L., Landberg, T., Lindholm, K., Linell, F., Ljungberg, O., Ranstam, J., and Sigfusson, B. (1988). Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. *BMJ*, 297(6654):943–948.

[6] Andriole, G. L., Crawford, E. D., Grubb, R. L., Buys, S. S., Chia, D., Church, T. R., and *et al.* (2012). Prostate cancer screening in the randomized prostate, lung, colorectal, and ovarian cancer screening trial: mortality results after 13 years of follow-up. *Journal of the National Cancer Institute*, 104:125–132.

[7] Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):pp. 444–455.

[8] Atkin, W. S., Edwards, R., Kralj-Hans, I., Wooldrage, K., Hart, A. R., Northover, J. M., and et al (2010). Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *Lancet*, 375(9726):1624–1633.

[9] Baines, C. J. (2014). Re: Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ*. `http://www.bmj.com/content/348/bmj.g366/rr/687252`, accessed on Feb. 25, 2014.

[10] Baker, S. G., Kramer, B. S., and Prorok, P. C. (2008). Early reporting for cancer screening trials. *Journal of Medical Screening*, 15(3):122–129.

[11] Berry, D. A., Cronin, K. A., Plevritis, S. K., Fryback, D. G., Clarke, L., Zelen, M., Mandelblatt, J. S., Yakovlev, A. Y., Habbema, J. D., and Feuer, E. J. (2005). Effect of screening and adjuvant therapy on mortality from breast cancer. *New England Journal of Medicine*, 353(17):1784–1792.

[12] Bjurstam, N., Bjorneld, L., Warwick, J., Sala, E., Duffy, S. W., Nystrom, L., Walker, N., Cahlin, E., Eriksson, O., Hafstrom, L. O., Lingaas, H., Mattsson, J.,

Persson, S., Rudenstam, C. M., Salander, H., Save-Soderbergh, J., and Wahlin, T. (2003). The Gothenburg Breast Screening Trial. *Cancer*, 97(10):2387–2396.

[13] Branswell, H. (2014). Contentious Canadian study says mammography doesnt cut deaths from breast cancer. *National Post*. `http://news.nationalpost.com/2014/02/12/contentious-canadian-study-says-mammography-doesnt-cut-deaths-from-breast-cancer/`, accessed on Feb. 25, 2014.

[14] Caro, J. J. (1990). Screening for breast cancer in quebec: estimates of health effects and of costs. *Montreal: CETS*.

[15] Charalampoudis, P. (2014). Re: Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ*. `http://www.bmj.com/content/348/bmj.g366/rr/686969`, accessed on Feb. 25, 2014.

[16] CISNET Breast Cancer Collaborators (2006). The Impact of Mammography and Adjuvant Therapy on U.S. Breast Cancer Mortality (1975–2000): Collective Results from the Cancer Intervention and Surveillance Modeling Network. *Journal of the National Cancer Institute Monographs*, 2006(36):1–126.

[17] Cole, S. R. and Frangakis, C. E. (2009). The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20:3–5.

[18] Corrado, D., Basso, C., Pavei, A., Michieli, P., Schiavon, M., and Thiene, G. (2006). Trends in sudden cardiovascular death in young competitive athletes after implementation of a preparticipation screening program. *JAMA*, 296(13):1593–1601.

[19] CTFPHC (2011). The Canadian Task Force on Preventive Health Care: Recommendations on screening for breast cancer in average-risk women aged 40-74 years. *CMAJ*, 183(17):1991–2001.

[20] Dammin, T. C. (2014). Re: Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ*. `http://www.bmj.com/content/348/bmj.g366/rr/687263`, accessed on Feb. 25, 2014.

[21] Day, N. and Warren, R. (2000). Mammographic screening and mammographic patterns. *Breast Cancer Res.*, 2(4):247–251.

[22] Day, N. E. and Walter, S. D. (1984). Simplified models of screening for chronic disease: Estimation procedures from mass screening programmes. *Biometrics*, 40:1–13.

[23] Duchateau, L., Collette, L., Sylvester, R., and Pignon, J. (2000). Estimating number of events from the kaplan-meier curve for incorporation in a literature-based meta-analysis: What you don't see you can't get! *Biometrics*, 56(3):886–892.

[24] Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium On Mathematical Statistics and Probability*, pages 831–853.

[25] Ericson, K. (2014). Re: Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ*. `http://www.bmj.com/content/348/bmj.g366/rr/686656`, accessed on Feb. 25, 2014.

[26] Etzioni, R. (2014). Re: Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ*. http://www.bmj.com/content/348/bmj.g366/rr/686666, accessed on Feb. 25, 2014.

[27] Ferguson, S. A. (2014). Re: Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ*. http://www.bmj.com/content/348/bmj.g366/rr/686676, accessed on Feb. 25, 2014.

[28] Fey, M. F. and Tobler, A. (2011). Marriage risk of cancer research fellows. *Lancet*, 378(9809):2070.

[29] Frisell, J., Lidbrink, E., Hellstrom, L., and Rutqvist, L. E. (1997). Followup after 11 years–update of mortality results in the Stockholm mammographic screening trial. *Breast Cancer Res. Treat.*, 45(3):263–270.

[30] Galton, F. (1901). Biometry. *Biometrika*, 1:7–10.

[31] Goss, P., Ingle, J., Ales-Martinez, J., and *et al.* (2011). Exemestane for breast-cancer prevention in postmenopausal women. *The New England Journal of Medicine*, 364:2381–2391.

[32] Gøtzsche, P. C. (2011). Time to stop mammography screening? *CMAJ*, 183(17):1957–1958.

[33] Gøtzsche, P. C. (2012). *Mammography screening: truth, lies, and controversy.* Radcliffe, London, UK.

[34] Guyot, P., Ades, A. E., Ouwens, M. J., and Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*, 12:9.

[35] Hanley, J. A. (2005). Analysis of mortality data from cancer screening studies: looking in the right window. *Epidemiology*, 16(6):786–790.

[36] Hanley, J. A. (2010). Mortality reductions produced by sustained prostate cancer screening have been underestimated. *Journal of Medical Screening*, 17(3):147–151.

[37] Hanley, J. A. (2011). Measuring mortality reductions in cancer screening trials. *Epidemiologic Reviews*, 33(1):36–45.

[38] Hanley, J. A., McGregor, M., Liu, Z., Strumpf, E. C., and Dendukuri, N. (2013). Measuring the mortality impact of breast cancer screening. *Canadian Journal of Public Health*, 104(7):e437–442.

[39] Heijnsdijk, E. A., Wever, E. M., Auvinen, A., Hugosson, J., Ciatto, S., Nelen, V., Kwiatkowski, M., Villers, A., Paez, A., Moss, S. M., Zappa, M., Tammela, T. L., Makinen, T., Carlsson, S., Korfage, I. J., Essink-Bot, M. L., Otto, S. J., Draisma, G., Bangma, C. H., Roobol, M. J., Schroder, F. H., and de Koning, H. J. (2012). Quality-of-life effects of prostate-specific antigen screening. *The New England Journal of Medicine*, 367(7):595–605.

[40] Hernán, M. A., Cole, S. R., Margolick, J., Cohen, M., and Robins, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety*, 14:477–491.

[41] Hu, P. and Zelen, M. (1997). Planning clinical trials to evaluate early detection programmes. *Biometrika*, 84:817–830.

[42] Humphrey, L. L., Helfand, M., Chan, B. K., and Woolf, S. H. (2002). Breast cancer screening: a summary of the evidence for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, 137(5 Part 1):347–360.

[43] Jansen, J. (2011). Network meta-analysis of survival data with fractional polynomials. *BMC Medical Research Methodology*, 11(61).

[44] Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, Second Edition*. Wiley, New Jersey.

[45] Kofteridis, D. P., Alexopoulou, C., Valachis, A., Maraki, S., Dimopoulou, D., Georgopoulos, D., and Samonis, G. (2010). Aerosolized plus intravenous colistin versus intravenous colistin alone for the treatment of ventilator-associated pneumonia: a matched case-control study. *Clinical Infectious Diseases*, 51(11):1238–1244.

[46] Kolata, G. (2014). Vast Study Casts Doubts on Value of Mammograms. *New York Times*. `http://www.nytimes.com/2014/02/12/health/study-adds-new-doubts-about-value-of-mammograms.html`, accessed on Feb. 25, 2014.

[47] Kopans, D. B. (2014). Re: Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ*. `http://www.bmj.com/content/348/bmj.g366/rr/686292`, accessed on Feb. 25, 2014.

[48] Kopans, D. B. and Feig, S. A. (1993). The Canadian National Breast Screening Study: a critical review. *American Journal of Roentgenology*, 161(4):755–760.

[49] Kösters, J. P. and Gøtzsche, P. C. (2003). Regular self-examination or clinical examination for early detection of breast cancer (Review). *Cochrane Database of Systematic Reviews*, (CD003373).

[50] Law, M. (2009). What now on screening for prostate cancer? *Journal of Medical Screening*, 16(3):109–111.

[51] Lee, S. and Zelen, M. (2006). A stochastic model for predicting the mortality of breast cancer. *Journal of the National Cancer Institute Monographs*, 2006(36):79–86.

[52] Levman, J. (2014). Re: Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ*. `http://www.bmj.com/content/348/bmj.g366/rr/686749`, accessed on Feb. 25, 2014.

[53] Liu, Z., Hanley, J. A., and Strumpf, E. C. (2013a). Projecting the yearly mortality reductions due to a cancer screening program. *Journal of Medical Screening*, 20(3):156–164.

[54] Liu, Z., Rich, B., and Hanley, J. A. (2013b). Recovering the raw data behind a non-parametric survival curve. *Under Review*, 0(0):0.

[55] Mandel, J. S., Church, T. R., Bond, J. H., Ederer, F., Geisser, M. S., Mongin, S. J., Snover, D. C., and Schuman, L. M. (2000). The effect of fecal occult-blood screening on the incidence of colorectal cancer. *New England Journal of Medicine*, 343(22):1603–1607.

[56] Mandelblatt, J. S., Cronin, K. A., Bailey, S., Berry, D. A., de Koning, H. J., Draisma, G., Huang, H., Lee, S. J., Munsell, M., Plevritis, S. K., Ravdin, P.,

Schechter, C. B., Sigal, B., Stoto, M. A., Stout, N. K., van Ravesteyn, N. T., Venier, J., Zelen, M., and Feuer, E. J. (2009). Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. *Annals of Internal Medicine*, 151(10):738–747.

[57] Mandelblatt, J. S., Cronin, K. A., Berry, D. A., Chang, Y., de Koning, H. J., Lee, S. J., Plevritis, S. K., Schechter, C. B., Stout, N. K., van Ravesteyn, N. T., Zelen, M., and Feuer, E. J. (2011). Modeling the impact of population screening on breast cancer mortality in the United States. *Breast*, 20 Suppl 3:75–81.

[58] Markus, L. (2012). *ReadImages: Image Reading Module for R.* R package version 0.1.3.2.

[59] Miettinen, O. S. (2000). Screening for lung cancer: Can it be cost-effective? *Canadian Medical Association Journal*, 162:1431–1436.

[60] Miettinen, O. S. (2008). Screening for a cancer: a sad chapter in today's epidemiology. *European Journal of Epidemiology*, 23(10):647–653.

[61] Miettinen, O. S. (2014). Screening for Breast Cancer: What Truly Is the Benefit? *Canadian Journal of Public Health*, 104(7):e435–e436.

[62] Miettinen, O. S., Henschke, C. I., and Pasmantier, M. W. (2002). Mammographic screening: no reliable supporting evidence? *Lancet*, 359:404–405.

[63] Miettinen, O. S. and Karp, I. (2012). *Epidemiological Research: An Introduction.* Springer, Dordrecht.

[64] Miller, A. B., Baines, C. J., To, T., and Wall, C. (1992). Canadian National Breast Screening Study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. *CMAJ*, 147(10):1477–1488.

[65] Miller, A. B., To, T., Baines, C. J., and Wall, C. (2000). Canadian National Breast Screening Study-2: 13-year results of a randomized trial in women aged 50-59 years. *J. Natl. Cancer Inst.*, 92(18):1490–1499.

[66] Miller, A. B., Wall, C., Baines, C. J., Sun, P., To, T., and Narod, S. A. (2014). Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ*, 348:g366.

[67] Morrison, A. S. (1985). *Screening in Chronic Disease*. New York: Oxford University Press, first edition.

[68] Morrison, A. S. (1992). *Screening in Chronic Disease*. New York: Oxford University Press, second edition.

[69] Moss, S. M., Cuckle, H., Evans, A., Johns, L., Waller, M., and Bobrow, L. (2006). Effect of mammographic screening from age 40 years on breast cancer mortality at 10 years' follow-up: a randomised controlled trial. *Lancet*, 368(9552):2053–2060.

[70] Mukherjee, S. (2010). *The emperor of all maladies: a biography of cancer*. Scribner, New York.

[71] Murrell, P. (2009). Importing vector graphics: The grimport package for r. *Journal of Statistical Software*, 30:1–37.

[72] National Lung Screening Trial Research Team (2011a). Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England Journal of Medicine*, 365(5):395–409.

[73] National Lung Screening Trial Research Team (2011b). The National Lung Screening Trial: overview and study design. *Radiology*, 258(1):243–253.

[74] Nelson, H. D., Tyne, K., Naik, A., Bougatsos, C., Chan, B. K., and Humphrey, L. (2009). Screening for breast cancer: an update for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, 151(10):727–737.

[75] Ouwens, M., Philipsa, Z., and Jansen, J. (2010). Network meta-analysis of parametric survival curves. *Research Synthesis Methods*, 1:258–271.

[76] Parmar, M., Torri, V., and Stewart, L. (1998). Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine*, 17:2815–2834.

[77] Raffle, A. E. and Gray, J. A. M. (2007). *Screening - Evidence and practice*. Oxford University Press.

[78] Ridker, P. M., MacFadyen, J. G., Fonseca, F. A., Genest, J., Gotto, A. M., Kastelein, J. J., Koenig, W., Libby, P., Lorenzatti, A. J., Nordestgaard, B. G., Shepherd, J., Willerson, J. T., and Glynn, R. J. (2009). Number needed to treat with rosuvastatin to prevent first cardiovascular events and death among men and women with low low-density lipoprotein cholesterol and elevated high-sensitivity C-reactive protein: justification for the use of statins in prevention: an intervention trial evaluating rosuvastatin (JUPITER). *Circ Cardiovasc Qual Outcomes*, 2(6):616–623.

[79] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 6:41–55.

[80] Schröder, F. H., Hugosson, J., Roobol, M. J., Tammela, T. L. J., Ciatto, S., Nelen, V., Kwiatkowski, M., Lujan, M., Lilja, H., Zappa, M., and others (2009). Screening and prostate-cancer mortality in a randomized european study. *New*

England Journal of Medicine, 360:1320–1328.

[81] Self, S. G. (1991). An adaptive weighted log-rank test with application to cancer prevention and screening trials. *Biometrics*, 47(3):975–986.

[82] Self, S. G. and Etzioni, R. (1995). A likelihood ratio test for cancer screening trials. *Biometrics*, 51(1):44–50.

[83] Shapiro, S. (1977). Evidence on screening for breast cancer from a randomized trial. *Cancer*, 39(6 Suppl):2772–2782.

[84] Shaukat, A., Mongin, S. J., Geisser, M. S., Lederle, F. A., Bond, J. H., Mandel, J. S., and Church, T. R. (2013). Long-term mortality after screening for colorectal cancer. *New England Journal of Medicine*, 369(12):1106–1114.

[85] Shen, Y. and Zelen, M. (1999). Parametric estimation procedures for screening programmes: Stable and nonstable disease models for multimodality case finding. *Biometrika*, 86(3):503–515.

[86] Shen, Y. and Zelen, M. (2001). Screening sensitivity and sojourn time from breast cancer early detection clinical trials: mammograms and physical examinations. *Journal of Clinical Oncology*, 19:3490–3499.

[87] Shen, Y. and Zelen, M. (2005). Robust modeling in screening studies: Estimation of sensitivity and pre-clinical sojourn time distribution. *Biostatistics*, 6(3):604–614.

[88] Tabár, L., Fagerberg, C. J., Gad, A., Baldetorp, L., Holmberg, L. H., Grontoft, O., Ljungquist, U., Lundstrom, B., Manson, J. C., and Eklund, G. (1985). Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish

National Board of Health and Welfare. *Lancet*, 1(8433):829–832.

[89] Tabar, L., Vitak, B., Chen, T. H., Yen, A. M., Cohen, A., Tot, T., Chiu, S. Y., Chen, S. L., Fann, J. C., Rosell, J., Fohlin, H., Smith, R. A., and Duffy, S. W. (2011). Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiology*, 260(3):658–663.

[90] Tarone, R. (1995). The excess of patients with advanced breast cancers in young women screened with mammography in the canadian national breast screening study. *Cancer*, 75:997–1003.

[91] Tartter, P. I. (2014). Re: Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ*. `http://www.bmj.com/content/348/bmj.g366/rr/686641`, accessed on Feb. 25, 2014.

[92] Therneau, T. and Lumley, T. (2011). survival: Survival analysis, including penalised likelihood. *R package version 2.36-10*.

[93] Tierney, J., Stewart, L., Ghersi, D., Burdett, S., and MR, S. (2007). Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials*, 8(16).

[94] Tudur, C., Williamson, P., Khan, S., and Best, L. (2001). The value of the aggregate data approach in meta-analysis with time-to-event outcomes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164:357–370.

[95] USPSTF (2014). Recommendations for adults. *US Preventive Services Task Force*. `www.uspreventiveservicestaskforce.org/adultrec.htm#cancer`.

[96] Walter, S. D. and Day, N. E. (1983). Estimation of the duration of a pre-clinical disease state using screening data. *American Journal of Epidemiology*, 118:865–886.

[97] Welch, H. G., Schwartzl, L., and Woloshin, S. (2011). *Overdiagnosed*. Beacon Press, Boston.

[98] Williamson, P., Smith, C., Hutton, J., and Marson, A. (2002). Aggregate data meta-analysis with time-to-event outcomes. *Statistics in Medicine*, 21:3337–3351.

[99] Yankelevitz, D. F. and Smith, J. P. (2013). Understanding the core result of the National Lung Screening Trial. *The New England Journal of Medicine*, 368(15):1460–1461.

[100] Zauber, A. G., Lansdorp-Vogelaar, I., Knudsen, A. B., Wilschut, J., van Ballegooijen, M., and Kuntz, K. M. (2008). Evaluating test strategies for colorectal cancer screening: a decision analysis for the u.s. preventive services task force. *Annals of Internal Medicine*, 149(9):659–669.

[101] Zelen, M. (1993). Optimal scheduling of examinations for the early detection of disease. *Biometrika*, 80:279–293.

[102] Zelen, M. and Feinleib, M. (1969). On the theory of screening for chronic diseases. *Biometrika*, 56:601–614.

[103] Zucker, D. M. and Lakatos, E. (1990). Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika*, 77:853–864.

Appendix A: R code for the Hu-Zelen model (Chapter 3)

```
# Updated 2012-08-30

# LIST OF INPUT VALUES:
# number invited to each arm N=25,000;
# incidence rate w=0.006;
# prevalence at randomization P_0(t_0)=0.003;
# sensitivity beta=0.7
# median survivals for those not in S_p at t_0 are 10 and 17 years
# in the control and screening arms respectively;
# median survivals for those in S_p at t_0 are 11 and 20 years
# in the control and screening arms respectively;
# screening exams are in year 0, 1, 2, 3;

N=25000; w=6/1000; p=3/1000; beta=0.7;
lambdac0=log(2)/10; lambdac=log(2)/11;
lambdas0=log(2)/17; lambdas=log(2)/20;
t0=0; ts=c(0.001,1,2,3,100); K=length(ts)-1;
T=15;

###############control arm###############
Dc1=function(x){
    p*w*lambdac0*exp(-lambdac0*x)/(w-lambdac0*p)*
    (1-exp(lambdac0*x-w*x/p))
}

Dc2=function(x){
    w*exp(-lambdac*x)*(exp(lambdac*x)-1-(lambdac*p/(w-lambdac*p))*
    (1-exp(lambdac*x-w*x/p)))
}

theta9=rep(0,T)
for (i in 1:T){
    theta9[i]=integrate(Dc1, lower=0,upper=i)$value+
    integrate(Dc2, lower=0,upper=i)$value
}
theta9[T]
```

```
#cumulative no. deaths
Dc=theta9*N;Dc
#yearly no. deaths:
d.c=c(Dc[1],Dc[2:T]-Dc[1:T-1]);d.c

###############screening arm (screen-detected)###############
#formula 3.4 part1 in Hu and Zelen (1997)
Dr1=function(x,r){
    ((1-beta)^(r-1))*beta*w*lambdas0*
    (p/(lambdas0*p-w))*exp(-lambdas0*x)*
    (exp(lambdas0*x-w*x/p)-exp(lambdas0*ts[r]-w*ts[r]/p))
}

theta7=rep(0,T)
for (i in 1:T){
    for (r in 1:K){
        theta7[i]=theta7[i]+ifelse(ts[r]>i,0,
        integrate(Dr1,lower=ts[r],upper=i,r=r)$value)
    }
}

#formula 3.4 part2
Dr2=function(x,r){
    sum=0
    for (j in 1:r) {
        sum <- sum + ((1-beta)^(r-j))*beta*lambdas*w*
        (p/(lambdas*p-w))*exp(-lambdas*x)*
        (exp(w*ts[j]/p)-exp(w*(ifelse((j-1)==0,0,ts[j-1]))/p))*
        (exp(lambdas*x-w*x/p)-exp(lambdas*ts[r]-w*ts[r]/p))
    }
    return(sum)
}

theta8=rep(0,T)
for (r in 1:K){
    for (i in 1:T){
        theta8[i]= theta8[i]+ifelse(ts[r]>i,0,
```

```
        integrate(Dr2, lower=ts[r],upper=i,r=r)$value)
    }
}

#formula 3.4: cumulative number of screen-detected deaths
Dr=(theta7+theta8)*N;Dr
#yearly no. screen-detected deaths:
d.r=c(Dr[1],Dr[2:T]-Dr[1:T-1]);d.r

###############screening arm (interval cancers)###############
#using our proposed shortcut

Dr1_star=function(x,r){
    ((1-beta)^(r-1))*beta*w*lambdac0*
    (p/(lambdac0*p-w))*exp(-lambdac0*x)*
    (exp(lambdac0*x-w*x/p)-exp(lambdac0*ts[r]-w*ts[r]/p))
}

theta7_star=rep(0, T)
for (i in 1: T){
    for (r in 1:K){
        theta7_star[i]=theta7_star[i]+ifelse(ts[r]>i,0,
        integrate(Dr1_star,lower=ts[r],upper=i,r=r)$value)
    }
}

Dr2_star=function(x,r){
    sum=0
    for (j in 1:r) {
        sum <- sum + ((1-beta)^(r-j))*beta*lambdac*w*
        (p/(lambdac*p-w))*exp(-lambdac*x)*
        (exp(w*ts[j]/p)-exp(w*(ifelse((j-1)==0,0,ts[j-1]))/p))*
        (exp(lambdac*x-w*x/p)-exp(lambdac*ts[r]-w*ts[r]/p))
    }
    return(sum)
}

theta8_star=rep(0, T)
```

```
for (r in 1:K){
    for (i in 1: T){
        theta8_star[i]= theta8_star[i]+ifelse(ts[r]>i,0,
        integrate(Dr2_star, lower=ts[r],upper=i,r=r)$value)
    }
}

#cumulative no. deaths from the screening arm
theta1=theta7+theta8+(theta9-theta7_star-theta8_star)
Ds=(theta1)*N;Ds
#yearly no. deaths from screening arm:
d.s=c(Ds[1],Ds[2:T]-Ds[1:T-1]); d.s

#Output (cumulative no. of deaths):
round(cbind(Dc,Ds))
```

Appendix B: Validating the R code (Chapter 3)

To validate my R code, I contacted the first author Ping Hu and checked my programming in R against hers in FORTRAN by comparing the output using the same input values. Below I present the 3 sets of inputs we used. Our results are almost identical.

```
--------------------------INPUT 1-----------------------------
- number invited to each arm N=25,000;
- incidence rate w=0.006;
- prevalence at randomization P_0(t_0)=0.003;
- sensitivity beta=0.7;
- median survivals for those not in S_p at t_0 are 10 and 17 years
  in the control and screening arms, respectively;
- median survivals for those in S_p at t_0 are 11 and 20 years
  in the control and screening arms, respectively;
- screening exams are in year 0, 1, 2, 3;

--------------------------My R Output 1-----------------------------
Year  Dc  Ds
  1   5   4
  2  19  16
  3  41  35
  4  71  61
```

```
  5 108  94
  6 152 134
  7 203 182
  8 260 235
  9 322 295
  10 390 360
  11 462 431
  12 540 506
  13 621 587
  14 707 671
  15 797 760

--------------------Ping Hu's FORTRAN Output 1--------------------
  mu    med_c0  med_c   med_s0  med_s    alpha(one-side)
  0.50   10.00   11.00   17.00   20.00     0.050

     w      p_0     beta_S    t0      t1      t2      t3
  0.0060   0.0030    0.70    0.00    1.00    2.00    3.00

  # of exam =  4     sample size =  25000   25000

     T      death_c   death_s     Mort Red(%)  POWER(%)
  3.00         41.       35.          14.        16.
  4.00         71.       61.          14.        22.
  5.00        108.       94.          13.        25.
  6.00        152.      135.          12.        28.
  7.00        203.      182.          10.        29.
  8.00        260.      235.           9.        29.
  9.00        322.      295.           8.        29.
 10.00        390.      360.           8.        28.
 11.00        462.      431.           7.        28.
 12.00        540.      506.           6.        27.
 13.00        621.      587.           6.        26.
 14.00        707.      671.           5.        25.
 15.00        797.      760.           5.        24.


-------------------------INPUT 2-------------------------------
```

- number invited to each arm N=25,000;
- incidence rate w=0.006;
- prevalence at randomization P_0(t_0)=0.003;
- sensitivity beta=0.9
- median survivals for those not in S_p at t_0 are 10 and 17 years
  in the control and screening arms, respectively;
- median survivals for those in S_p at t_0 are 11 and 20 years
  in the control and screening arms, respectively;
- screening exams are in year 0, 1, 2, 3;

----------My R OUTPUT2----------

Year	Dc	Ds
1	5	4
2	19	15
3	41	34
4	71	58
5	108	90
6	152	130
7	203	176
8	260	229
9	322	288
10	390	353
11	462	423
12	540	498
13	621	578
14	707	662
15	797	750

----------Ping Hu's FORTRAN Output 2----------

mu	med_c0	med_c	med_s0	med_s	alpha(one-side)
0.50	10.00	11.00	17.00	20.00	0.050

w	p_0	beta_S	t0	t1	t2	t3
0.0060	0.0030	0.90	0.00	1.00	2.00	3.00

# of exam = 4 sample size = 25000 25000

T	death_c	death_s	Mort Red(%)	POWER(%)

3.00	41.	34.	18.	22.
4.00	71.	58.	18.	29.
5.00	108.	91.	16.	35.
6.00	152.	130.	15.	38.
7.00	203.	176.	13.	39.
8.00	260.	229.	12.	40.
9.00	322.	288.	11.	39.
10.00	390.	353.	9.	39.
11.00	462.	423.	9.	38.
12.00	540.	498.	8.	36.
13.00	621.	578.	7.	35.
14.00	707.	662.	6.	34.
15.00	797.	750.	6.	33.

----------INPUT 3----------

- number invited to each arm N=50,000;
- incidence rate w=0.006;
- prevalence at randomization P_0(t_0)=0.003;
- sensitivity beta=0.9
- median survivals for those not in S_p at t_0 are 10 and 17 years in the control and screening arms, respectively;
- median survivals for those in S_p at t_0 are 11 and 20 years in the control and screening arms, respectively;
- screening exams are in year 0, 1, 2, 3;

----------My R OUTPUT 3----------

Year	Dc	Ds
1	10	8
2	38	30
3	82	67
4	142	117
5	216	181
6	305	260
7	406	352
8	519	458
9	644	576
10	779	705
11	925	845

```
 12 1079  995
 13 1243 1155
 14 1415 1324
 15 1594 1500

-------------------Ping Hu's FORTRAN Output 3-------------------
     mu    med_c0  med_c   med_s0  med_s    alpha(one-side)
  0.50   10.00   11.00   17.00   20.00     0.050

     w      p_0     beta_S    t0      t1      t2      t3
  0.0060   0.0030    0.90    0.00    1.00    2.00    3.00

  # of exam =  4     sample size =  50000   50000

     T      death_c   death_s     Mort Red(%)  POWER(%)
  3.00         82.       67.          18.        33.
  4.00        142.      117.          18.        46.
  5.00        216.      181.          16.        55.
  6.00        305.      260.          15.        60.
  7.00        406.      352.          13.        62.
  8.00        519.      458.          12.        62.
  9.00        644.      576.          11.        62.
 10.00        779.      705.           9.        61.
 11.00        925.      845.           9.        59.
 12.00       1079.      996.           8.        58.
 13.00       1243.     1155.           7.        56.
 14.00       1415.     1324.           6.        54.
 15.00       1594.     1501.           6.        52.
```

Appendix C: NLST reanalysis R code (Chapter 6)

```
#Some R code
#updated 2014-01-27

PATIENT=read.table("patient.csv", sep=",", head=T)
pt=PATIENT[,c("pid", "rndgroup", "age", "death_days",
"deathcutoff", "finaldeathLC")]
dim(pt); length(pt$pid)

participation.rate=c(98.5+94+92.9+98+92.6+91.2)/600
participation.rate

ds=subset(pt,pt$finaldeathLC==1&deathcutoff %in% c(1,2))
dim(ds)

ds$deathtime=as.numeric(as.character(ds$death_days))
summary(ds$deathtime/365)

ds=subset(ds,ds$deathtime<366*7)
dim(ds)

#save
ds0=ds

##########functions##########

logL.chisq <- function (x, n.year.follow.up, screen.time,
participation.rate, randomization.ratio)
{
    max.reduction = exp(x[1])/(1 + exp(x[1]))
    degree = exp(x[2]) + 2
    y = 1:n.year.follow.up - 0.5
    one.reduction = participation.rate *
     max.reduction * dchisq(x = y, df = degree)/
     dchisq(x = degree - 2, df = degree)
    mat = matrix(0, ncol = n.year.follow.up,
    nrow = length(screen.time))
    a = n.year.follow.up:1
```

```
    for (i in 1:length(screen.time)) {
        mat[i, 1:length(screen.time[i]:length(a))] <-
         a[screen.time[i]:length(a)]
    }
    cell.index = t(mat[, ncol(mat):1])
    prob.not.helped = NULL
    if (length(screen.time) > 1) {
        for (i in 1:nrow(cell.index)) {
            prob.not.helped[i] <-
            cumsum(prod(1 - one.reduction[cell.index[i, ]]))
        }
        R <- 1 - prob.not.helped
    }
    else R <- one.reduction
    p = randomization.ratio * (1 - R)/
     (randomization.ratio * (1 - R) + 1)
    like = dat[, 2] * log(p) + dat[, 1] * log(1 - p)
    sum(like)
}

logL.half.year.chisq <- function (x, n.interval, screen.time,
participation.rate, randomization.ratio)
{
    max.reduction = exp(x[1])/(1 + exp(x[1]))
    degree = exp(x[2]) + 2
    y = (1:n.interval - 0.5)/2
    one.reduction = participation.rate *
     max.reduction * dchisq(x = y, df = degree)/
     dchisq(x = degree - 2, df = degree)
    mat = matrix(0, ncol = n.interval,
    nrow = length(screen.time))
    a = n.interval:1
    for (i in 1:length(screen.time)) {
        mat[i, 1:length(screen.time[i]:length(a))] =
         a[screen.time[i]:length(a)]
    }
    cell.index = t(mat[, ncol(mat):1])
    prob.not.helped = NULL
```

```
    if (length(screen.time) > 1) {
        for (i in 1:nrow(cell.index)) {
            prob.not.helped[i] = cumsum(prod(1 -
             one.reduction[cell.index[i, ]]))
        }
        R <- 1 - prob.not.helped
    }
    else R <- one.reduction
    p = randomization.ratio * (1 - R)/
     (randomization.ratio * (1 - R) + 1)
    like = dat[, 2] * log(p) + dat[, 1] * log(1 - p)
    sum(like)
}

logL.individual.chisq <- function (x, screen.time,
participation.rate, randomization.ratio)
{
    max.reduction = exp(x[1])/(1 + exp(x[1]))
    degree = exp(x[2]) + 2
    shift = screen.time[2:length(screen.time)] -
     screen.time[1:(length(screen.time) - 1)]
    Q = function(y) {
        q1 = participation.rate * max.reduction * dchisq(x = y,
            df = degree)/dchisq(x = degree - 2, df = degree)
        y = y - shift[1]
        q2 = participation.rate * max.reduction * dchisq(x = y,
            df = degree)/dchisq(x = degree - 2, df = degree)
        y = y - shift[2]
        q3 = participation.rate * max.reduction * dchisq(x = y,
            df = degree)/dchisq(x = degree - 2, df = degree)
        y = y + shift[1] + shift[2]
        return(ifelse(y <= screen.time[2], q1,
         ifelse(y > screen.time[2] & y <= screen.time[3],
         q2 * (1 - q1) + q1, (1 - q1) * (1 - q2) * q3 +
         q2 * (1 - q1) + q1)))
    }
    H = Q(ds$deathtime/365.25)
    p = randomization.ratio * (1 - H)/
```

```
     (randomization.ratio * (1 - H) + 1)
    like = sum(log(p[which(ds$rndgroup == "1")])) +
     sum(log(1 - p[which(ds$rndgroup == "2")]))
    sum(like)
}

reduction.chisq <- function (x, follow.up.time, screen.time,
participation.rate)
{
    max.reduction = exp(x[1])/(1 + exp(x[1]))
    degree = exp(x[2]) + 2
    one.reduction = participation.rate *
     max.reduction * dchisq(x = follow.up.time, df = degree)/
     dchisq(x = degree - 2, df = degree)
    mat = matrix(0, ncol = length(follow.up.time),
     nrow = length(screen.time))
    a = length(follow.up.time):1
    for (i in 1:length(screen.time)) {
        mat[i, 1:length(screen.time[i]:length(a))] <-
         a[screen.time[i]:length(a)]
    }
    cell.index = t(mat[, ncol(mat):1])
    prob.not.helped = NULL
    if (length(screen.time) > 1) {
        for (i in 1:nrow(cell.index)) {
            prob.not.helped[i] =
             cumsum(prod(1 -one.reduction[cell.index[i, ]]))
        }
        R <- 1 - prob.not.helped
    }
    else R <- one.reduction
    return(R)
}

##########data##########

xray=subset(ds0,ds0$rndgroup==2);dim(xray)
ct=subset(ds0,ds0$rndgroup==1);dim(ct)
```

```
postscript("~/Dropbox/1-work/1-PhD/thesis/Figures/
NLST_cumulative.eps", width=8.5, height=5,
paper="special", horizontal=FALSE)
plot(sort(xray$deathtime)/365,cumsum(rep(1,nrow(xray))),col=2)
points(sort(ct$deathtime)/365,cumsum(rep(1,nrow(ct))),col=1)
dev.off()

##########using yearly data##########

d0=as.numeric(table(floor(sort(xray$deathtime/365))))
d1=as.numeric(table(floor(sort(ct$deathtime/365))))
dat=cbind(d0,d1);dat
colSums(dat)

fit=optim(par=c(-2.5,1),fn=logL.chisq,n.year.follow.up=7,
screen.time=c(1,2,3),participation.rate=participation.rate,
randomization.ratio=1,method="BFGS",hessian=T,
control=list(fnscale=-1))
fit
(est=fit$par)
(covmat=solve(-fit$hessian))
cov2cor(covmat)
se=sqrt(diag(covmat));
CI.lo=est-qnorm(.975)*se; CI.hi=est+qnorm(.975)*se;

(param=c(exp(est[1])/(1+exp(est[1])),2+exp(est[2])))
(SE=c(se[1]*exp(est[1])/((1+exp(est[1]))^2),exp(est[2])*se[2]))
c(exp(CI.lo[1])/(1+exp(CI.lo[1])),2+exp(CI.lo[2]));
c(exp(CI.hi[1])/(1+exp(CI.hi[1])),2+exp(CI.hi[2]));

postscript("~/Dropbox/1-work/1-PhD/thesis/Figures/
NLST_fit_chisq.eps",width=8.5,height=5,paper="special",
horizontal=FALSE)
n.FU=9
R=reduction.chisq(est, follow.up.time=seq(0, n.FU, by=.1),
screen.time=c(1,11,21), participation.rate=participation.rate)
plot(seq(0, n.FU, by=.1), 1-R, ylim = c(0, 1.6),
```

```
ylab="Reduction", xlab="Follow-up year", type = "l",
col=2, lwd=2, yaxt="n", xaxt="n");

axis(2, at=1-seq(0,1,.2), labels=paste(100*seq(0,1,.2),"%"),
las=2,cex=.8)
axis(1, at=seq(0,n.FU,1),labels=paste(0: n.FU), las=1, cex=.8)
points(seq(0.5,6.5, by=1), dat[,2]/dat[,1],
cex=.3/sqrt(1/dat[,2]+1/dat[,1]), pch=19, col=2)
abline(h=1,lty=1)
abline(h=seq(.2,1,by=.2),lty='dotted')

##########using half-yearly data##########

d0=as.numeric(table(floor(sort(xray$deathtime/182.5))))
d1=as.numeric(table(floor(sort(ct$deathtime/182.5))))
round(100*(1-d1/d0),di=2)
dat=cbind(d0,d1);dat
colSums(dat)

fit=optim(par=c(-2.5,1), fn=logL.half.year.chisq, n.interval=14,
screen.time=c(1,3,5), participation.rate=participation.rate,
randomization.ratio=1,method="BFGS",hessian=T,
control=list(fnscale=-1))
(est=fit$par)
(covmat=solve(-fit$hessian))
cov2cor(covmat)
se=sqrt(diag(covmat));
CI.lo=est-qnorm(.975)*se; CI.hi=est+qnorm(.975)*se;

(param=c(exp(est[1])/(1+exp(est[1])),2+exp(est[2])))
(SE=c(se[1]*exp(est[1])/((1+exp(est[1]))^2),exp(est[2])*se[2]))
c(exp(CI.lo[1])/(1+exp(CI.lo[1])),2+exp(CI.lo[2]));
c(exp(CI.hi[1])/(1+exp(CI.hi[1])),2+exp(CI.hi[2]));

R=reduction.chisq(est, follow.up.time=seq(0,n.FU,by=.1),
screen.time=c(1,11,21), participation.rate=participation.rate)
lines(seq(0, n.FU,by=.1), 1-R, lwd=2, col=4)
points(seq(.5,13.5,by=1)/2, dat[,2]/dat[,1],
```

```
cex=.3/sqrt(1/dat[,2]+1/dat[,1]), pch=19, col=4)

##########using individual data##########

rm(fit)
fit=optim(par=c(-2.5,1), fn=logL.individual.chisq,
screen.time=c(0,1,2), participation.rate=participation.rate,
randomization.ratio=1, method="BFGS", hessian=T,
control=list(fnscale=-1))
fit
(est=fit$par)
(covmat=solve(-fit$hessian))
cov2cor(covmat)
se=sqrt(diag(covmat));
CI.lo=est-qnorm(.975)*se; CI.hi=est+qnorm(.975)*se;

(param=c(exp(est[1])/(1+exp(est[1])),2+exp(est[2])))
(SE=c(se[1]*exp(est[1])/((1+exp(est[1]))^2),exp(est[2])*se[2]))
c(exp(CI.lo[1])/(1+exp(CI.lo[1])),2+exp(CI.lo[2]));
c(exp(CI.hi[1])/(1+exp(CI.hi[1])),2+exp(CI.hi[2]));

R=reduction.chisq(est, follow.up.time=seq(0,n.FU,by=.1),
screen.time=c(1,11,21), participation.rate=participation.rate)
lines(seq(0,n.FU,by=0.1), 1-R, lwd=2, lty=2)

legend("bottomright",col=c(2,4,1),c("Using yearly data",
"Using half-yearly data","Using individual-level data"),
lty=c(1,1,2),lwd=2)
text(0,.05,"S1",cex=1); text(1,.05,"S2",cex=1);
text(2,.05,"S3",cex=1);
dev.off()

##########projection##########

fit=optim(par=c(-2.5,1), fn=logL.individual.chisq,
screen.time=c(0,1,2), participation.rate=participation.rate,
randomization.ratio=1,method="BFGS",hessian=T,
control=list(fnscale=-1))
```

```
fit
(est=fit$par)
(covmat=solve(-fit$hessian))
cov2cor(covmat)

se=sqrt(diag(covmat));
CI.lo=est-qnorm(.975)*se; CI.hi=est+qnorm(.975)*se;
(param=c(exp(est[1])/(1+exp(est[1])),2+exp(est[2])))
c(exp(CI.lo[1])/(1+exp(CI.lo[1])),2+exp(CI.lo[2]));
c(exp(CI.hi[1])/(1+exp(CI.hi[1])),2+exp(CI.hi[2]));

postscript("~/Dropbox/1-work/1-PhD/thesis/Figures/
NLST_projection.eps",width=8.5,height=5,paper="special",
horizontal=FALSE)

n.FU=15;
R=reduction.chisq(est, follow.up.time=seq(0,n.FU,by=.1),
screen.time=seq(1,91,by=10), participation.rate=participation.rate)
plot(seq(0, n.FU, by=.1), 1-R, ylim = c(0, 1), xlim=c(0,15),
ylab="Reduction", xlab="Follow-up year", type = "l",
col=2, lwd=2, yaxt="n", xaxt="n");

R=reduction.chisq(est, follow.up.time=seq(0,n.FU,by=.1),
screen.time=seq(1,21,by=10),
participation.rate=participation.rate)
lines(seq(0, n.FU, by=.1), 1-R, lty=2, lwd=2);

axis(2,at=1-seq(0,1,.2), labels=paste(100*seq(0,1,.2),"%"),
las=2, cex=.8)
axis(1,at=seq(0,n.FU,1), labels=paste(0: n.FU), las=1, cex=.8)

text(0:2,rep(.1,3),"S",col=1)
text(0:9,rep(.015,10),"S",col=2)
abline(h=seq(.2,1,by=.2),lty='dotted')
legend("bottomright",legend=c("Fitted (3 rounds)",
"Projection (10 rounds)"), lty=c(2,1), col=c(1,2), lwd=c(2,2))
dev.off()
```

```
##########projection (using subgroup data)##########

xray=subset(ds0,ds0$rndgroup==2&ds0$age<65);
ct=subset(ds0,ds0$rndgroup==1&ds0$age<65);
ds=subset(ds0,ds0$age<65)

logL.individual.chisq <- function (x, screen.time,
participation.rate, randomization.ratio)
{
    max.reduction = exp(x[1])/(1 + exp(x[1]))
    degree = exp(x[2]) + 2
    shift = screen.time[2:length(screen.time)] -
     screen.time[1:(length(screen.time) - 1)]
    Q = function(y) {
        q1 = participation.rate * max.reduction * dchisq(x = y,
            df = degree)/dchisq(x = degree - 2, df = degree)
        y = y - shift[1]
        q2 = participation.rate * max.reduction * dchisq(x = y,
            df = degree)/dchisq(x = degree - 2, df = degree)
        y = y - shift[2]
        q3 = participation.rate * max.reduction * dchisq(x = y,
            df = degree)/dchisq(x = degree - 2, df = degree)
        y = y + shift[1] + shift[2]
        return(ifelse(y <= screen.time[2], q1,
         ifelse(y > screen.time[2] & y <= screen.time[3],
         q2 * (1 - q1) + q1, (1 - q1) *
            (1 - q2) * q3 + q2 * (1 - q1) + q1)))
    }
    H = Q(ds$deathtime/365.25)
    p = randomization.ratio * (1 - H)/
     (randomization.ratio * (1 - H) + 1)
    like = sum(log(p[which(ds$rndgroup == "1")])) +
     sum(log(1 - p[which(ds$rndgroup == "2")]))
    sum(like)
}

fit=optim(par=c(-2.5,1),fn=logL.individual.chisq,
screen.time =c(0,1,2), participation.rate=participation.rate,
```

```
randomization.ratio=1, method="BFGS", hessian=T,
control=list(fnscale=-1))
fit
(est=fit$par)
(covmat=solve(-fit$hessian))
cov2cor(covmat)
(param=c(exp(est[1])/(1+exp(est[1])),2+exp(est[2])))

postscript("~/Dropbox/1-work/1-PhD/thesis/Figures/
NLST_projection_below65.eps", width=8.5, height=5,
paper="special", horizontal=FALSE)

n.FU=15
R=reduction.chisq(est, follow.up.time=seq(0,n.FU,by=.1),
screen.time=seq(1,91,by=10),
participation.rate=participation.rate)
plot(seq(0, n.FU, by=.1), 1-R, ylim = c(0, 1),
ylab="Reduction", xlab="Follow-up year", type = "l",
col=2, lwd=2, yaxt="n", xaxt="n");

R=reduction.chisq(est, follow.up.time=seq(0,n.FU,by=.1),
screen.time=seq(1,21,by=10),
participation.rate=participation.rate)
lines(seq(0, n.FU, by=.1), 1-R, lty=2,lwd=2);

axis(2,at=1-seq(0,1,.2), labels=paste(100*seq(0,1,.2),"%"),
las=2, cex=.8)
axis(1,at=seq(0,n.FU,1), labels=paste(0: n.FU), las=1, cex=.8)
text(0:2,rep(.1,3),"S",col=1)
text(0:9,rep(.015,10),"S",col=2)
abline(h=seq(.2,1,by=.2),lty='dotted')
legend("bottomright",legend=c("Fitted (3 rounds)",
"Projection (10 rounds)"),lty=c(2,1),col=c(1,2),lwd=c(2,2))
dev.off()
```