STATISTICAL METHODS FOR THE EVALUATION OF A CANCER SCREENING PROGRAM

STATISTICAL METHODS FOR THE EVALUATION OF A CANCER SCREENING PROGRAM

BY

HUAN JIANG, M.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF CLINICAL EPIDEMIOLOGY & BIOSTATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

PHD OF HEALTH RESEARCH METHODOLOGY

PhD in Health Research Methodology (2014) McMaster University
(Clinical Epidemiology & Biostatistics) Hamilton, Ontario, Canada

TITLE:	STATISTICAL METHODS FOR THE EVALUATION OF A CANCER SCREENING PROGRAM
AUTHOR:	Huan Jiang M.Sc., (Health Research Methodology) McMaster University, Hamilton
SUPERVISOR:	Dr. Stephen Walter
NUMBER OF PAGES:	xiii, 99

Abstract

Estimation of the sensitivity and specificity of cancer screening tests using data from population-level databases is complicated by the lack of independent confirmation of test results using a "gold standard". The true sensitivity and specificity are unknown and errors in measurement can occur due to subjective clinical judgment, technical imperfections or interpretational differences. A further complication is clustered data (such as patients nested within examiners within screening centre), which are common in population-based screening. We propose a cancer screening model that accommodates the partially unobserved disease status, clustered data structures, general covariate effects, and the dependence between exams.

The model is applied to the estimation of the diagnostic accuracy of mammography and clinical breast examination using a cohort consisting of women 50 to 69 years of age screened at the OBSP between January 1, 2002 and December 31, 2003. When offered in addition to mammography, we found CBE may benefit women using hormone therapy but not likely benefit women with dense breast tissues.

The thesis also discusses two measures of interest, the length of the pre-clinical state and the false negative rate. Two estimation procedures are proposed to model the pre-clinical state duration, the false negative rate of screening exam, and the underlying incidence rate in the screened population. Both methods assume the sojourn time follows a negative exponential distribution, but we consider two different forms for the false negative rate: 1) constant with time and 2) an exponential function to compensate for the fact that lesions should become easier to detect the closer they are to become clinically evident. The proposed methods are illustrated with another cohort of women who were first screened through the OBSP between January 1, 2003, and December 31, 2004 and were followed up until December 31, 2009.

Acknowledgements

Completing this thesis has been thoroughly enjoyable. That enjoyment is largely a result of the interactions I had with my supervisors and colleagues.

I would like to thank Anna Chiarelli for the provision of the data. I also thank the Ontario Breast Screening Program, a program of Cancer Care Ontario, for use of its data for this study.

I feel privileged to have worked with my supervisors, Stephen Walter, Patrick Brown and Parminder Raina. To each of them I owe a great debt of gratitude for their patience, inspiration and friendship. They have taught me a great deal about the field of health research by sharing with me the joy of discovery.

The Unit of Research, Prevention and Cancer Control of Cancer Care Ontario has provided an excellent environment for my research. The project is supported by grants from the Cancer Care Ontario Population Study Network.

Thanks also to my family who have been extremely understanding and supportive of my studies. I feel very lucky to have a family that shares my enthusiasm for academic pursuits.

Declaration of academic achievement

Huan Jiang contributed to the development of the methods, analyzed and interpreted the data, wrote the source codes, and performed statistical analyses presented in Chapters 2, 3 and 4. Especially for the project presented in Chapter 4, she initiated the project, designed the study, and collected the data and obtained funding to support the study. Huan Jiang wrote all manuscripts chapters.

Dr. Patrick Brown contributed to the conception and design of the model presented in Chapter 2. He supported the development of the source codes, the finalization of the findings and critically reviewed the manuscripts presented in Chapter 2.

Dr. Stephen Walter contributed to the conception and design of the model presented in Chapter 4. He provided support for the analysis of the cohort data, and critically reviewed all manuscripts chapters.

Dr. Anna Chiarelli provided the data we used in Chapter 2 and 3. She reviewed all chapters in the thesis.

Notation and abbreviations

CBE: clinical breast exam

OBSP: Ontario Breast Screening Program

OCR: Ontario Cancer Registry

MCMC: Markov Chain Monte Carlo

CNBSS: Canadian National Breast Screening Studies

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Breast cancer is one of the most frequently diagnosed cancers and one of the leading causes of cancer death in Ontario women (Ontario Breast Screening Program, 2013). Regular screening for breast cancer is widely recommended to reduce mortality due to breast cancer. In light of this, the Ontario Breast Screening Program (OBSP) was initiated in 1990 and it has developed into a provincial program that provides high-quality, regular breast cancer screening and assessment services. From the time the program was launched in 1990 to 2013, the OBSP provided more than 5 million mammograms to over 1.4 million women. Between April 2007 and March 2008, approximately 43 million dollars were spent on this program. However, there is a growing recognition among medical professionals that while some individuals may benefit from early diagnosis and treatment as a result of screening, others may be inconvenienced from false positive results, be diagnosed and treated unnecessarily, or remain undiagnosed and lose the opportunity of being treated at an early stage (Schroen *et al.*, 1996; Brekelmans *et al.*, 1996).

A cancer screening test is not a diagnostic test, and it cannot be evaluated using conventional methods based on the assumption of a perfect reference test. In fact, the nature of the cancer process precludes exact observation of the time of onset of the preclinical disease state and the clinical disease state. A woman who has a negative screen may subsequently develop a so-called “interval” cancer, defined as a cancer diagnosed in the time intervals between routine exams. This may represent either a false-negative result in previous screening examinations or a case of cancer that has developed since the last examination. These two possibilities cannot be distinguished by direct observation. Moreover, the sensitivity of screening is not directly estimable except in the theoretical situation where all screening participants are subjected to a definitive diagnostic test, such as pathological classification of tissue biopsy; this would be highly unethical and therefore infeasible. In practice, the true sensitivity and specificity are unknown, and the uncertainties concerning the true cancer status of screened individuals might lead to biased estimation of the performance of an exam.

The potential benefit of a screening exam is the ability to detect the disease earlier than would otherwise occur without screening, enabling improved prognosis as a result of earlier intervention. Ideally, such benefit would be demonstrated by a comparison of the number of cancer-specific deaths in a randomized group of women being screened with a comparable unscreened population. The apparent effects of early diagnosis and intervention are always more favorable than the true, long-term effects: a comparison between screening-detected cases and non-screen-detected cases overestimates benefit because the former consists of cases that were diagnosed earlier, progress more slowly, and may never become clinically relevant. Some of those effects are due to lead-time bias, which reflects earlier diagnosis as a result of the screen even though the overall survival time of the patient may be unchanged. Others might relate to length bias, which refers to the fact that screening is more likely to detect slower-growing, less aggressive cancers. The extreme example of length bias is over-diagnosis, where a slow-growing cancer found by screening might never have caused harm or required treatment during a patient’s lifetime.

A public screening screening service is even harder to evaluate than screening in the context of a randomized trial, because of its observational nature. First, screening is often provided by different examiners in different centres. Screening outcome is influenced by the individual level risk factors such as age and family history. Those variations should be monitored and minimized to ensure the quality of a screening service. For example, in the OBSP, some screening centres offer both mammography and clinical breast exam (CBE) while others offer mammography only. However, the benefit of CBE has not been clearly established.

Second, maintaining high levels of regular attendance and timely participation of screenees are important but often difficult to achieve. For example, although biennial screening is recommended for women ages 50 to 69 years at average risk in Canada, only about 70% women return after a previous screening exam (Bobo *et al.*, 2004; Lees, 1996). The probability of attending routine rescreening might relate to the previous screening experience, knowledge about cancer and screening, cultural background and lifestyle. Irregular attendees could affect our understanding of the disease process and influence the estimate of overall screen accuracy.

Third, a complete evaluation of a screening exam may need 10 or more years of follow-up to observe the mortality rate. However, public health services need to respond in a timely fashion to public inquiries regarding the efficiency and effectiveness of their population-based screening programs. Besides, screening protocols require continual revision to allow for new research findings as well as technological developments. A program without any changes in its operation for more than 10 years is rarely found. Therefore an earlier and swift evaluation of the screening exams is often required to assess the appropriateness, efficiency and effectiveness of the program.

In recent years, there have been several debates about the effectiveness of screening for breast cancer as a public health policy: concerns have been raised about how often one should be screened for breast cancer, what should be the age to start regular screening, and whether one should be screened before age 50 years (Chiarelli *et al.* (2009); Oestreicher *et al.* (2005); Barton *et al.* (1999); US Preventive Services Task Force and others (2009); Miller *et al.* (2014)). This thesis explores these questions by modeling the cancer disease process and investigating the performance of screen exams in several ways:

- In Chapter two, a newly developed model is used to estimate the diagnostic error rates of cancer screening exams; to draw inferences about the accuracy of a screening test and to quantify the effect of explanatory variables, having accounted for heterogeneities and unobserved cancers which are inherent in population-level administrative data. The mixed model containing both fixed-effects and random-effects terms employed in this chapter accommodates the multilevel data structures, general covariate effects, partially unobserved disease status, and dependence between tests. A Markov chain Monte Carlo (MCMC) algorithm is described for estimating the posterior distributions of the sensitivity, specificity and prevalence when the reference standard is imperfect. The MCMC approach extends naturally to the case of unobserved cancers by including cancer status in the model as an unknown (or partially observed) latent variable.

- Chapter three is an extension of the second chapter. It presents an investigation of the effect of clinical breast examination in addition to mammography in the OBSP program. The joint mixed effects logistic regression model developed in Chapter 2 is used to estimate the effects of risk factors and overcome the difficulties in the analysis of clustered data with partly missing true cancer status. The author investigates and predicts the effects of age and other risk factors such as breast density, hormone use and family history on screening accuracy.

- Chapter four discusses a model to estimate the time interval between the onset of the detectable preclinical stage and the onset of the clinical stage (called “sojourn time”). We revisit and extend the Markov-type model developed by Day and Walter (1984) and apply the model to a cohort from the OBSP. We show how to estimate the sojourn time and the sensitivity of the screen using cohort data on the observed prevalence of breast cancer at successive screens and on the incidence of disease during intervals between screens. We further investigate the variation of screening sensitivity and the mean sojourn time for different age groups. Lastly, we apply the same method to a clinical trial (CNBSS-I) to compare the estimates from the OBSP study with those from randomized trials.

Chapter 2

Inference on Cancer Screening Exam Accuracy Using Population-Level Administrative Data

Abstract

This paper develops a model for cancer screening and cancer incidence data, accommodating the partially unobserved disease status, clustered data structures, general covariate effects, and dependence between exams. The true unobserved cancer and detection status of screening participants are treated as latent variables, and a Markov Chain Monte Carlo algorithm is used to estimate the Bayesian posterior distributions of the diagnostic error rates and disease prevalence. We show how the Bayesian approach can be used to draw inferences about screening exam properties and disease prevalence while allowing for the possibility of conditional dependence between two exams. The techniques are applied to the estimation of the diagnostic accuracy of mammography and clinical breast examination using data from the Ontario Breast Screening Program in Canada.

Keywords: Bayesian inference, test accuracy, cancer screening, clustered analysis, random

effect, latent class model

2.1 Introduction

Population-level administrative data from screening programmes for cancers and other diseases are a rich and extensive information source for evaluating screening effectiveness. The Ontario Breast Screening Program (OBSP) in Canada maintains one such database, containing information on hundreds of thousands of cancer screens, including individual-level characteristics such as age and family history (see Chiarelli *et al.*, 2003, 2009, 2010). In comparison with data from medical records or clinical trials, however, population-level data are more challenging to analyze due to two key difficulties not faced with clinical study data. The first is the "observational" nature of population-level data, with the test performance affected by heterogeneities in both the characteristics of the individuals being screened (McPherson *et al.*, 2000) and in the judgements and characteristics of the medical professionals carrying out the exams (Beam *et al.*, 1996). The second difficulty is the lack of a 'gold standard' reference test applied to all subjects, with the result that a number of cancers might remain unobserved or missed by the examiner and the health system during the followup periods.

The first of difficulties faced with population-level data, which in statistical terms is correlated data with varying probabilities of incidence and detection, can be accounted for through mixed effects models (see Diggle *et al.*, 2013, for example). The term mixed effects model refers to the use of both fixed and random effects in the same analysis. It allows a wide variety of correlation patterns to be modeled explicitly. Individual factors such as age and family history, examiner-level factors such as years of experience, and institution-level variables such as annual number of patients screened can be accounted for by their inclusion as fixed effects. Variations in detection rates and false positive rates amongst examiners and institutions can be captured by random effects, as in the two-level hierarchical model used by Woodard *et al.* (2007).

The solution to the second difficulty of some incident cases being unobserved is less straightforward, particularly when more than one screening test is administered and the sensitivity of this second test is variable. In the OBSP data considered here, each woman is examined twice with a radiologists performing a mammography and a nurse giving a Clinical Breast Exam (CBE). When mammography is performed first, a cancer missed by mammography might be either: (a) detected by CBE and diagnosed; (b) missed by CBE and later became symptomatic and diagnosed by clinical practitioners; or (c) missed by CBE and remained undetected. The consequence of misclassifying the cancers in category (c) as true negatives would be slight when considering a strata of the population for whom CBE is particularly effective and the number of such cancers is small. This misclassification would be consequential when comparing mammography sensitivity across different strata, i.e. age groups or health facilities, where the sensitivity of CBE is known to vary. Such an analysis should ideally acknowledge and account for the CBE test having created differences in misclassification rates among strata.

The goal of this paper is to make inferences on the performance of screening tests and to quantify the effect of specified explanatory variables using population-level administrative data. The particular objective considered is accommodating the heterogeneous nature of population-level data by explicitly modelling individual-level test result probabilities and creating and integrating out a “nusiance variable” for unobserved cancers. The random effects model employed in this paper accommodates multilevel data structures, general covariate effects, partially unobserved disease status, and dependence between screening tests. A Markov chain Monte Carlo (MCMC) algorithm is described for estimating the posterior distributions of sensitivity, specificity and prevalence when the reference standard is imperfect. The MCMC approach extends naturally to the case of unobserved cancers by including cancer status in the model as an unknown (or partially observed) latent variable. The method is illustrated through the analysis of data from the OBSP.

2.1.1 Data and motivating problem

The OBSP collects data on the women screened, the screening centres and practitioners carrying out the screen, and which women have been referred (or *recalled*) for medical treatment. Women may be referred for further assessment by either the radiologist or the nurse. For women diagnosed with a screen-detected or a cancer that is diagnosed in the interval after a normal screening examination and before the subsequent screen, pathological confirmation is obtained from regional staff through the recall process and through linkages with the Ontario Cancer Registry (OCR). A clinical diagnosis of breast cancer within 12 months of a screening exam is assumed to mean that the cancer existed at the time the exam was administered. This 12 months threshold is used by Chiarelli *et al.* (2009, 2010), references which can be consulted for further information on the OBSP program.

Case ascertainment in the OCR is estimated to be 98% complete for breast cancer (Holowaty *et al.*, 1995). Identification of vital status of participants is accomplished by record linkage with the Ontario Registrar General's mortality file (Marrett, 1990). All of the linkages were accomplished using a computerized probabilistic record linkage system known as Auto Match (Jaro, 1995). Further details of this data linkage is given in Appendix A.

The data considered here is a cohort of 234177 women screened, between January 1, 2002 and December 31, 2003 in the 73 OBSP screening centres which offer both CBE and mammography. Table 2.1 shows the number of positive and negative screening tests for each of CBE and mammography, subdivided according to whether cancers were subsequently diagnosed within 12 months of screening. There were 261 cancers diagnosed for women having negative preceeding test results for both mammography and CBE, or 14% of all cancer cases.Ignoring unobserved cancers, of the individuals having a breast cancer diagnosis, the sensitivity of mammography is much higher than CBE, being 100×1467/1822 = 80.5% versus 100×491/1822 = 26.9% for CBE. Of the individuals for whom no cancer was diagnosed, the specificity of mammography is 100×217299/232355 = 93.5% while the specificity of CBE is 100×225958/23255 = 97.2%. Table 2.2 summarizes the individual-level variables in OBSP dataset, with total number of women screened, number with positive screening tests, and number of cancers observed.

	CBE		
	Negative	Positive	Total
No cancer diagnosed			
Mammography negative	212190	5109	217299
Mammography positive	13768	1288	15056
Total	225958	6397	232355
Cancer diagnosed			
Mammography negative	261	94	355
Mammography positive	1070	397	1467
Total	1331	491	1822
Total			
Mammography negative	212451	5203	217654
Mammography positive	14838	1685	16523
Total	227289	6888	234177

Table 2.1: The number of women having negative screening tests and positive screening tests, and total number of screens administered for both mammography and Clinical Breast Exam (CBE), subdivided according to whether cancer was diagnosed within 12 months of screening.

2.1.2 Literature Review

Uncertainty in true disease status is known to affect estimates of sensitivity and specificity, with misclassification of a few breast cancer cases being able to cause a dramatic change in estimates of parameters in a logistic regression model (Carroll, 2006). Several authors (Neuhaus, 1999; Rosychuk and Thompson, 2001; Roy *et al.*, 2005; Rosychuk and Islam, 2009) discuss the effects of misclassification on binary responses in different settings and claim that failure to account for measurement errors in covariates or responses causes biased and inconsistent parameter estimates. Approaches proposed for correcting the estimates of regression parameters include the SIMEX method (Kuchenhoff *et al.*, 2006) and a number of Bayesian methods (e.g. McGlothlin *et al.*, 2008; Rosychuk and Islam, 2009).

	Total		Positive		Cancer	
	screened		tests			
Breast cancer in family						
No	205639	(87.8%)	19008	(87.5%)	1517	(83.3%)
Yes	28538	(12.2%)	2718	(12.5%)	305	(16.7%)
Breast density						
low, 75%	220456	(94.1%)	19966	(91.9%)	1664	(91.3%)
high, $\geq$ 75%	13721	(5.9%)	1760	(8.1%)	158	(8.7%)
Age						
$50-59$	132000	(56.4%)	13316	(61.3%)	917	(50.3%)
$60-69$	102177	(43.6%)	8410	(38.7%)	905	(49.7%)

Table 2.2: Summary of individual-level variables in OBSP dataset, with total number of women screened, number with positive screening tests, and number of cancers observed.

To deal with imperfect reference tests, several authors have developed latent class models in evaluating accuracy of diagnostic tests (Rindskopf and Rindskopf, 1986; Hui and Walter, 1980; Walter and Irwig, 1988; Espeland and Handelman, 1989; Uebersax and Grove, 1990; Qu *et al.*, 1996). These methods generally require more than two independent tests in order to estimate parameters of interest. When there are fewer than two different tests, some parameters need to be constrained at fixed values based on model assumptions. An alternative is to use Bayesian inference in these situations, assigning a prior distribution to reflect the uncertainty in these parameter values. Assuming conditional independence between tests, Joseph *et al.* (1995) and Black and Craig (2002) use Bayesian inference to estimate disease prevalence as well as model parameters, specifying Beta distributions as the prior of the prevalence, sensitivity and specificity and obtaining the joint posterior distribution via Gibbs sampling. Dendukuri and Joseph (2001) extended their work and estimated the disease prevalence and test accuracy while adjusting for conditional dependence between two tests.

Heterogeneity in population-level data has been addressed with random effects models in a number of studies (Qu *et al.*, 1996; Rutter and Gatsonis, 2001; Macaskill, 2004; Arends *et al.*, 2008). Puggioni *et al.* (2008) introduced a joint model for the four cell probabilities that determined the screening test result when two tests were applied. Through the model, the stochastic dependence could be examined between the estimates of sensitivity and specificity. The methodology presented here is based on a similar random effects structure, and extends it to allow for multi-level correlation structures, covariates attached to observations as well as examiners, and partially unobserved disease status.

2.2 A latent variable model for cancer screening

2.2.1 Model description

Cancer screening outcomes are described with a random effects model, with observed data including test results and cancer diagnoses and unobserved latent variables, including the individual's true disease status and the true and false positive rates of the screening tests being administered. The model structure is illustrated graphically in Figure 2.1 and explained in detail in the following paragraphs.

The observed data used in the cancer screening model have the following elements.

- T_{ij} is the result of screening test j performed on individual $i = 1 \dots 234117$, with $j = 1$ indicating mammography and $j = 2$ denoting CBE. A positive or abnormal test is coded as $T_{ij} = 1$ with $T_{ij} = 0$ otherwise.
- Y_i is the observed cancer status of individual i, with $Y_i = 1$ if cancer was clinically diagnosed within 12 months of the screening exam and $Y_i = 0$ otherwise.
- e_{ij} is an identifier for the examiner administering test j on individual i.
- s_i identifies the health facility where the exams were performed.
- X_i, $R_{e_{ij}}$, and Q_{s_i} are vectors of covariates (or explanatory variables) associated with the individual, the examiners, and the health facility respectively.

There are no examiner-level variables $R_{e_{ij}}$ in the application of the model presented here, though this variable is retained in the model description for completeness.

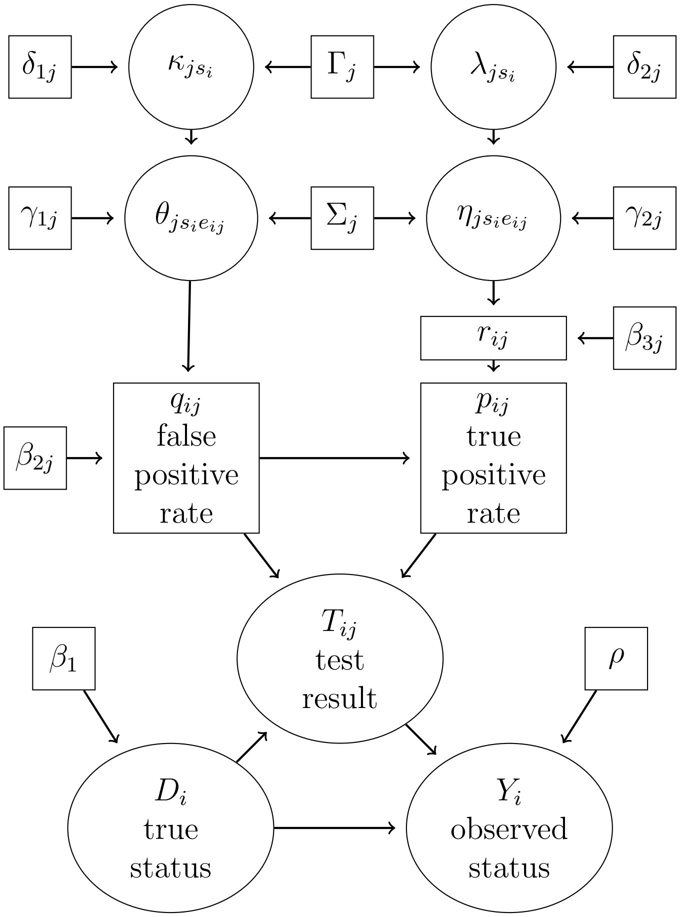


Figure 2.1: Graphical representation of the latent-variable screening model

The probability of observing any combination of test results T_{ij} and cancer incidence Y_i depends on the following, generally unobserved, quantities.

- D_i is the true disease status of subject i at the time of the screening tests. An observed cancer incidence with $Y_i = 1$ implies $D_i = 1$, though the converse is not always true with unobserved cancers $D_i = 1$ and $Y_{ij} = 0$ being possible.

- p_{ij} and q_{ij} are the true positive and false positive rates for exam j when administered to individual i.

- ρ is the probability that a cancer which remains undetected after screening tests will be subsequently diagnosed and recorded within 12 months.

The screening model is a four-level hierarchical model, with the first level of the model specifying a joint distribution for T_{ij} and Y_i conditional on D_i:

$$\text{Test results:} \qquad T_{ij}|D_i \begin{cases} \sim \text{Bernoulli}(q_{ij}) & \text{if } D_i = 0 (false positive or true negative) \\ \sim \text{Bernoulli}(p_{ij}) & \text{if } D_i = 1 (true positive or false negative) \end{cases}$$

$$\text{Observed cancers:} \quad Y_i|D_i, T_{ij} \begin{cases} = D_i & \text{if } T_{i1} = 1 \text{ or } T_{i2} = 1 \\ = 0 & \text{if } D_i = 0, T_{i1} = T_{i2} = 0 \\ \sim \text{Bernoulli}(\rho) & \text{if } D_i = 1, T_{i1} = T_{i2} = 0. \end{cases} \tag{2.1}$$

Each test j is assumed to be conditionally independent given a true disease status, and followup medical procedures resulting from a positive test ($T_{i1} = 1$ or $T_{i2} = 1$) are assumed to accurately identify disease status.

Although two tests T_{i1} and T_{i2} are assumed to be conditionally independent, with $pr(T_{i1} = 1|D_i, T_{i2} = 1) = pr(T_{i1} = 1|D_i, T_{i2} = 0)$, the assumption does not hold when conditioning on the observed cases Y_i instead of the latent true disease status D_i. As a consequence, the relationship between the "observed" true positive rate $pr(T_{i1} = 1|Y_i = 1)$ and the "actual" true positive rate $pr(T_{i1} = 1|D_i = 1)$ is non-trivial and related to the properties of the second test administered, regardless of the ordering of tests being applied. The odds ratio corresponding to these two probabilities is derived in Appendix B.1 and is equal to

$$\begin{aligned} \frac{pr(T_{i1} = 1|D_i = 1)}{pr(T_{i1} = 0|D_i = 1)} \Bigg/ \frac{pr(T_{i1} = 1|Y_i = 1)}{pr(T_{i1} = 0|Y_i = 1)} =& pr(Y_i = 1|T_{i1} = 0, D_i = 1) \\ =& \rho + p_{i2} - \rho p_{i2}. \end{aligned} \tag{2.2}$$

When ρ=1, meaning cancers missed by screening tests are always observed during follow-up, the odds ratio of “observed” false positive (a false positive given observed cancer) versus “true” false positive (a false positive given “true” cancer) is 1.

When $\rho < 1$, the odds ratio depends on both ρ and the true positive rate of the second test p_{i2}. When the second test has a constant true positive rate for all i with $p_{i2} = p_{02}$, this odds ratio is constant for all i. In this circumstance, a logistic regression with the observed cancer as the outcome would be able to measure the relationship between the disease status and its risk factors, and ρ would only affect the intercept parameter. Inferences made on odds ratios involving $pr(T_{i1} = 1|Y_i = 1)$ under either of these conditions also apply to the true positive rate $pr(T_{i1} = 1|D_i = 1)$, an intuitive result as the observed cases with $Y_i = 1$ would be a random thinning of all cases having $D_i = 1$.

When p_{i2} varies with individual i, making inference on true positive rates $pr(T_{i1} = 1|D_i = 1)$ based on observed cancer status Y_i is much less straightforward. The true positive rates of the second screening test p_{i2} and the probability of observing a cancer after a normal screening ρ will both affect the degree to which $pr(T_{i1} = 1|Y_i = 1)$ is an underestimate of $pr(T_{i1} = 1|D_i = 1)$.

In the second level of the model, the true and false positive probabilities p_{ij} and q_{ij} are assigned mixed-effects logistic models and a logistic regression model is specified for cancer incidence D_i (Equation (2.3)). The probabilities p_{ij}, q_{ij} and ψ_i, relating to true and false positives and cancer incidence respectively, depend on intercept parameters μ and regression coefficients β for individual-level covariates X_i. The test results T_{ij} additionally depend on random effects associated with examiner e_{ij} at health facility s_i, denoted $\eta_{js_ie_{ij}}$ and $\theta_{js_ie_{ij}}$ for the true positive and false positive probabilities for exam j, respectively. The resulting formulation is

$$\begin{array}{lrl}
\text{Cancer:} & D_i \sim & \text{Bernoulli}(\psi_i) \\
 & \text{logit}(\psi_i) = & \mu_1 + X_i\beta_1 \\
\text{False positives:} & T_{ij} = 1|D_i = 0 \sim & \text{Bernoulli}(q_{ij}) \\
 & \text{logit}(q_{ij}) = & \mu_{2j} + X_i\beta_{2j} + \theta_{js_ie_{ij}} \\
\text{True positives:} & T_{ij} = 1|D_i = 1 \sim & \text{Bernoulli}(p_{ij}) \\
 & p_{ij} = & 1 - (1 - r_{ij})(1 - q_{ij}) \\
 & \text{logit}(r_{ij}) = & \mu_{3j} + X_i\beta_{3j} + \eta_{js_ie_{ij}}.
\end{array} \tag{2.3}$$

Note that examiners do not work in set pairs, and individuals a and b who see the same radiologist for mammography ($e_{a1} = e_{b1}$) would typically see different nurses for CBE ($e_{a2} \neq e_{b2}$). This 'non-nestedness' is reflected in the use of the double-subscripts for the random effects η and θ above.

The relationship between p_{ij}, q_{ij} and r_{ij} can be rewritten as an odds ratio $1 - r_{ij} = (1 - p_{ij})/(1 - q_{ij})$, and r_{ij} can be interpreted as the information conveyed to a test from an individual's true cancer status. A value of $r_{ij} = 0$ implies the two probabilities p_{ij} and q_{ij} are equal and the true cancer status, D_i, has no effects on the test probabilities once X_i, $R_{e_{ij}}$, and Q_{s_i} are accounted for. An effective test should respond to cancer status by increasing the probability of a positive test when $D_i = 1$, and the r_{ij} quantify this increase in a way that ensures $0 \leq q_{ij} \leq p_{ij} \leq 1$. The extreme case where $r_{ij} = 1$ implies a perfectly sensitive test having a 100% true positive rate with $p_{ij} = 1$. Further explanation and interpretation of this parametrization is given in Appendix B.2.

The third level of the model specifies the distribution of the examiner-level random effects $\eta_{js_ie_{ij}}$ and $\theta_{js_ie_{ij}}$ from (2.3), with these distributions depending on site-level random effects κ_{js_i} and λ_{js_i}. It is assumed that these random effects are multivariate normal with following distributions:

$$\begin{pmatrix}\eta_{js_ie_{ij}}\\ \theta_{js_ie_{ij}}\end{pmatrix} \sim\text{N}\left(\begin{pmatrix}\kappa_{js_i}+R_{je_{ij}}\gamma_{j1}\\ \lambda_{js_i}+R_{je_{ij}}\gamma_{j2}\end{pmatrix},\Sigma_j\right) \tag{2.4}$$
$$\begin{pmatrix}\kappa_{js_i}\\ \lambda_{js_i}\end{pmatrix} \sim\text{N}\left(\begin{pmatrix}Q_{s_i}\delta_{1j}\\ Q_{s_i}\delta_{2j}\end{pmatrix},\Gamma_j\right).$$

As mentioned previously, the $R_{je_{ij}}$ and Q_{s_i} are vectors of covariates associated with examiners and health facilities (although no examiner-level covariates are used in this application). These covariates have regression coefficient parameters γ_{j1} and δ_{1j} affecting the true positive probabilities and γ_{j2} and δ_{2j} affecting the false positive probability of test j. There each test j has a variance matrix Σ_j for the examiner level random effects and Γ_j for the site-level random effects. The fourth and final level to the model specifies the prior distributions of model parameters, as detailed in Section 2.2.2.

The formulation above assumes conditional independence between the two tests, an assumption which can be relaxed through the inclusion of an additional random effect term. A random effect M_i can induce correlation between exams when added to the model for r_{ij} as

$$\begin{aligned}\text{logit}(r_{ij}) =&\mu_{3j}+X_i\beta_{3j}+\eta_{js_ie_{ij}}+M_i \\ M_i \sim&\text{N}(0,\sigma^2).\end{aligned} \tag{2.5}$$

2.2.2 Prior distributions

Informative priors are used for the intercept parameters μ_{31} and μ_{32} for the true detection probabilities for mammography and CBE respectively. American Cancer Society guidelines (Smith *et al.*, 2003) list sensitivities for mammography from various clinical studies with a range from 0.64 to 0.84. Smith *et al.* (2003) also cite the meta-analysis of Barton *et al.* (1999), who estimate the sensitivity CBE with a 95% confidence interval between 0.48 and 0.60. As the intercept parameter μ_{3j} for true positive probability p_{ij} in (2.3) relates to true positives in excess of the false positive rate, the priors for μ_{3j} target slightly lower probabilities than the values above.

- A prior of $\mu_{31} \sim \mathrm{N}(0.63, 0.24^2)$ is used for the intercept of mammography's true detection rate, giving a 95% prior interval for the baseline referral probability between 0.54 and 0.75 (Smith *et al.*, 2003).

- The prior $\mu_{32} \sim \mathrm{N}(0.08, 0.16^2)$ is used for the intercept of CBE, giving a 95% interval for its true positive probability between 0.44 and 0.60 (Barton *et al.*, 1999).

- The fixed effects parameters β, γ and δ refer to log-odds ratios for a 10 year change in age, one standard deviation change in the other continuous variables, or the presence/absence of a binary variable. For those parameters, prior distributions of $\mathrm{N}(0, 2^2)$ were chosen as a change of 4 on the logit scale roughly corresponds to a change in probabilities from 0.5 to 0.99.

- A prior of $\mathrm{N}(-4, 1^2)$ was used for the intercept of the cancer model μ_1 to give a 95% prior interval between 2.5 and 12 cases per 1000 women (Chiarelli *et al.*, 2009).

- The missed cancer probability ρ is assumed to have a uninformative uniform prior of $\mathrm{Beta}(1.2, 1.2)$ giving a 95% prior interval of 0.04 and 0.96.

Priors of cancer personal-level risk factors were motivated by results from Boyd *et al.* (2007) and Madigan *et al.* (1995). Family history of breast cancer has a prior of $\mathrm{N}(0.95, 0.21^2)$ giving 95% interval for the odds ratio of (1.7, 3.9). For breast density, an informative prior is used with a distribution of $\mathrm{N}(0.75, 0.17^2)$ corresponding to 95% interval for the odds ratio of (1.5, 3.0).

The inverse Wishart distribution was assumed for the variance matrices, with Γ^{-1} and Σ^{-1} distributed as $\mathrm{Wishart}(I/20, 6)$ giving 95% intervals for the standard deviations of

(0.06,0.25). While this is an informative prior concentrated at small values, the random effects operate on the log-odds scale and the upper limit allows for a substantial amount of between-examiner variation. A standard deviation of 0.25 results in examiners in the 2.5 percentile having an odds of producing a positive test being around 1/3 of the odds for an examiner in the 97.5 percentile, which is roughly calculated by $\exp(\mu - 1.96\sigma)/\exp(\mu + 1.96\sigma) = \exp(-1.96.25 - (1.96.25))$. The lower end of the prior, where the standard deviation is 0.06, yields very little variation between examiners with the corresponding odds ratio between the 2.5 percentile and 97.5 percentile close to 0.8, calculated by the same way.

The distributions used for the priors were chosen because of their conjugacy, meaning that the conditional distributions of parameters have closed form when possible. Wishart distributions are conjugate with the Gaussian random effects, the beta prior on ρ is conjugate with the binary Y_i, and Gaussian priors on the site-level covariates are conjugate with the Gaussian random effects. The choice of Gaussian distributions for the random effects was likewise a pragmatic decision.

2.2.3 MCMC algorithm

Inference on the model is performed with a Markov Chain Monte Carlo (MCMC) algorithm to sample from the posterior distributions of the parameters and the latent variables. Bayesian inference via MCMC is used partly because of the difficulty in computing marginal probabilities with random effects and latent variables, and partly in order to allow for the use of informative priors. A Gibbs sampling routine is described in detail in Appendix B.3, with each iteration consisting of the steps given below.

- First, any unknown cancer status variables $D_i^{(n)}$ are sampled from the full conditional distribution given Y_i and T_{ij}, and using parameter values $p_{ij}^{(n-1)}$, $q_{ij}^{(n-1)}$ and $\rho^{(n-1)}$ from the previous iteration.

- Second, the $\beta^{(n)}$, $\mu^{(n)}$ and $\psi^{(n)}$ parameters and examiner-level random effects $\theta^{(n)}$ and

$\eta^{(n)}$ receive Metropolis-Hastings updates (see Chib and Greenberg, 1995).

- Third, examiner-level covariances $\Sigma_j^{(n)}$ are sampled directly from Wishart distributions using their conjugate Wishart priors, examiner-level random effects $\theta^{(n)}$ and $\eta^{(n)}$, and the previous iteration's examiner-level random effects $\kappa^{(n-1)}$ and $\lambda^{(n-1)}$.

- Fourth, site-level random effects $\kappa^{(n)}$ and $\lambda^{(n)}$ and coefficients δ are sampled directly from Gaussian conditional distributions.

- Finally, site-level variances $\Gamma_j^{(n)}$ are drawn from conditional Wishart distributions.

The algorithm is coded in R (R Core Team, 2014), and the code and synthetic data are available `http://pbrown.ca/screening`. The package "glmmBUGS" (Brown and Zhou, 2010) was used for generating starting values, assuming no unobserved cancers and fitting four univariate models (true positive and true negative outcomes for each of the two tests). The starting values for ψ in the cancer model parameters are estimated using maximum likelihood, with the intercept term increased slightly to induce unobserved additional cancers: given the MLE of this intercept term as -4.8, we increased it and arbitrarily selected five starting values between -3 and -4 (in increments of 0.2) for five parallel MCMC chains. Each of the five chains was run for 3100 iterations, with the the first 100 iterations discarded as burn-in and the results shown use every 15th iteration.

2.2.4 Sensitivities, specificities, and joint probabilities

Cancer researchers and planners are often more concerned with sensitivities and specificities associated with a particular screening regimen than with any of the parameters or random effects comprising the model in Section 2.2.1. Sensitivity and specificity for a single test j performed on an individual i are defined as $pr(T_{ij} = 1|D_i = 1) = \mathrm{E}(p_{ij})$ and $pr(T_{ij} = 0|D_i = 0) = 1 - \mathrm{E}(q_{ij})$ respectively. A pair of tests performed on individual i have combined accuracy values of

$$\begin{aligned} \text{sensitivity:} \quad & pr(T_{i1} = 1 \text{ or } T_{i2} = 1 | D_i = 1) & = \mathrm{E}[1 - (1 - p_{i1})(1 - p_{i2})] \\ \text{specificity:} \quad & pr(T_{i1} = 0 \text{ and } T_{i2} = 0 | D_i = 0) & = \mathrm{E}[(1 - q_{i1})(1 - q_{i2})]. \end{aligned} \tag{2.6}$$

The expectations above are due to the p_{ij} and q_{ij} being random quantities, depending on site-level and examiner-level random effects $\eta_{js_ie_{ij}}$, $\theta_{js_ie_{ij}}$, κ_{js_i}, and λ_{js_i}. The expectations in (2.6) are non-linear combinations of the model parameters, and the ease with which posterior samples of these quantities can be assembled from MCMC output is an additional argument for the Bayesian-MCMC inference methodology adopted here.

How the random effects are treated depends on whether the question of interest concerns a specific pair of examiners e_1 and e_2 at a given health facility s_1 in the dataset, or rather the outcomes expected from consulting a random or hypothetical pair of examiners and averaging out the variations amongst different examiners and screening sites. This latter scenario concerns the evaluation of a screening program as a whole, whereas the first question would be useful for evaluating individual examiners and understanding the variation amongst these examiners.

When considering a specific examiner, the expectations in (2.6) are computed using joint posterior samples of all model parameters and the examiner's random effect variables. For example, specificity estimated for test j by examiner e at site s on an individual with covariates X_0 is the sample mean of a set of $p_{0j}^{(n)}$ computed from $n = 1 \ldots N$ posterior MCMC samples $\mu_{21}^{(n)}$, $\beta_{2j}^{(n)}$ and $\theta_{jse}^{(n)}$ as

$$\text{logit}\left(q_{0j}^{(n)}\right) = \mu_{21}^{(n)} + \theta_{jse}^{(n)} + X_0\beta_{2j}^{(n)}. \tag{2.7}$$

When an average or hypothetical examiner is desired, the expectation is with respect to the posterior distribution of the model parameters and the unconditional distribution of the random effects given these parameters. Random effects $\theta_{jse}^{(n)}$ used to obtain $q_{0j}^{(n)}$ in (2.7), for example, would be drawn from the distributions

$$\begin{pmatrix} \kappa_{js}^{(n)} \\ \lambda_{js}^{(n)} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} Q_0\delta_{1j}^{(n)} \\ Q_0\delta_{2j}^{(n)} \end{pmatrix}, \Gamma_j^{(n)} \right) \text{ and } \begin{pmatrix} \eta_{jse}^{(n)} \\ \theta_{jse}^{(n)} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \kappa_{js}^{(n)} + R_{j0}\gamma_{j1}^{(n)} \\ \lambda_{js}^{(n)} + R_{j0}\gamma_{j2}^{(n)} \end{pmatrix}, \Sigma_j^{(n)} \right).$$

Notice that the calculation requires specifying covariates Q_0 and R_{j0}, though a distribution for sampling random $Q_0^{(n)}$ could be specified.

An individual i faced with a decision on whether to obtain a breast cancer screen would necessarily have an unknown cancer status D_i, and their decision would be better informed by joint probabilities of cancer status and test results than by probabilities conditional on cancer status. Of particular interest would be the probabilities

$$\begin{aligned} pr\big[(T_{i1} = 1 \text{ or } T_{i2} = 1) \text{ and } D_i = 1\big] &= \mathrm{E}\big\{[1 - (1 - p_{i1})(1 - p_{i2})]\psi_i\big\} \\ pr\big[(T_{i1} = 1 \text{ or } T_{i2} = 1) \text{ and } D_i = 0\big] &= \mathrm{E}\big\{[1 - (1 - q_{i1})(1 - q_{i2})][1 - \psi_i]\big\}, \end{aligned}$$

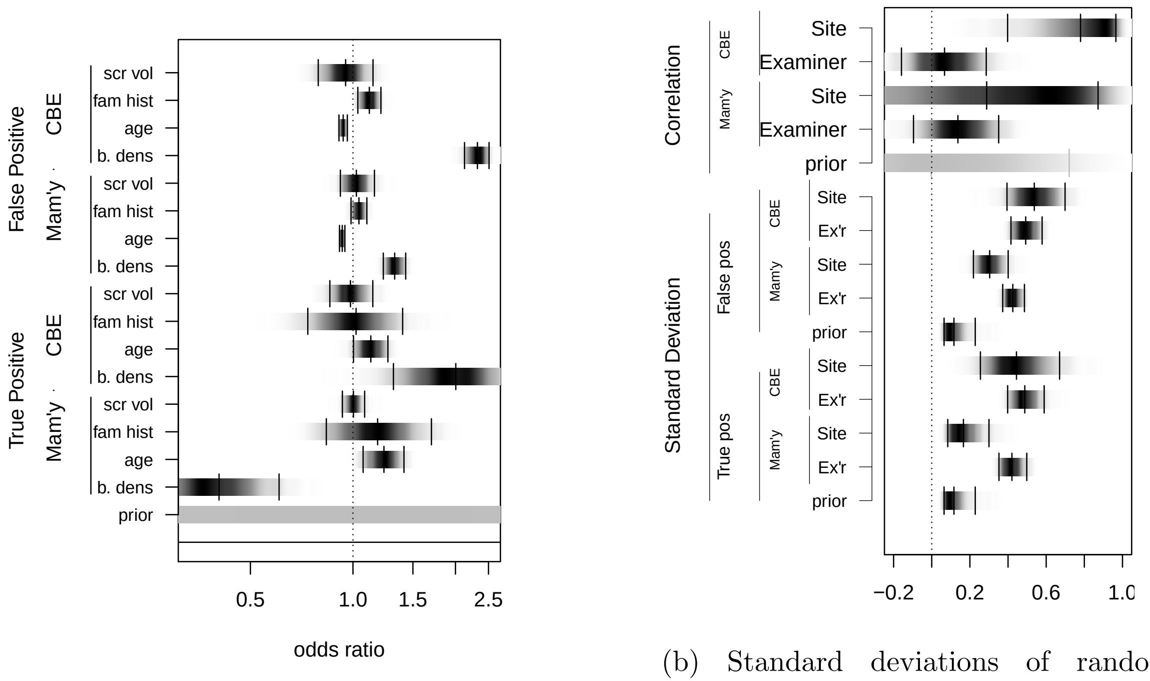
corresponding to: 1) having a cancer and having it detected by screening; and 2) not having cancer and obtaining a false positive test result. A high probability of outcome 1 is evidence in favour of the individual i obtaining screening tests, since this outcome would likely lead to early cancer detection and a consequent improvement in life expectancy. Outcome 2 can cause emotional stress, inconvenience, cost to the health care system, and in extreme cases lead to serious adverse outcomes from subsequent medical treatment. A high probability of this second outcome would argue against this individual being screened, though individuals will differ in their risk aversion for this second probability. Unlike sensitivities and specificities, these joint probabilities depend on the cancer prevalence rate for type of individual concerned and hence the model parameters governing the prevalence. A posterior sample of the cancer probability $\psi_i^{(n)}$ is calculated for each MCMC sample and used in conjunction with the $q_{0j}^{(n)}$ and $q_{0j}^{(n)}$ to create posterior samples of the joint probabilities.

2.3 Results

2.3.1 Parameters and random effects

Figure 2.2a contains posterior distributions for each regression coefficient β and δ. Four values are shown for each explanatory variable, with each type of test result (true positives $T_{ij} = 1|D_i = 1$ and false positives $T_{ij} = 1|D_i = 0$) having outcomes for each of the two screen tests (mammography for $j = 1$, and CBE when $j = 2$). The individual level covariates X_i are breast density (binary $\geq 75\%$ or $< 75\%$, with <75% as baseline), age (in decades), and family history as defined by a first degree relative having had breast cancer (and with no family history as baseline) (Chiarelli *et al.*, 2009). The site level covariate Q_s is the number of screens given annually at the health facility, log transformed and with the effect shown for the interquartile range. Note that all parameters have posterior distributions that are noticeably different from their non-informative prior distributions, indicating that the data has had a meaningful influence on determining the posteriors. Also, the posterior intervals are narrower for false positive outcomes, which is to be expected as there are many more false positives than true positives test results (cf. Table 2.1).

Breast density is the covariate with the strongest effect on test results, with high breast density having a pronounced effect on each of the four probabilities. High breast density increases the probability of a false positive for both screen tests, though the effect is much greater for CBE. The odds of a CBE false positive in the presence of high breast density are more than double the odds for a comparable woman with low breast density. The effect of high breast density on the true positive process is different for each screen test, with the odds of detecting a cancer with mammography on a woman with high breast density being roughly half of the comparable odds when breast density is low. Breast density increases the odds of cancer detection by CBE, and note that this true positive odds ratio reflects increased detection in excess of the corresponding increase in the false positive rate. Older women have lower odds of false positives and higher odds of a true positive than younger



(a) Regression coefficients for breast density (b. dens), 5.5 years or 1 SD change in age (age), Family history (fam hist), and IQR of annual screening volume for sites (scr vol).

(b) Standard deviations of random effects at the examiner-level sd($\eta_{js_ie_{ij}}$) and sd($\theta_{js_ie_{ij}}$) and site-level sd(κ_{js_i}) and sd(λ_{js_i}), along with correlations cor[$\eta_{js_ie_{ij}}, \theta_{js_ie_{ij}}$] and cor[$\kappa_{js_i}, \lambda_{js_i}$]

.

Figure 2.2: Prior and posterior distributions of model parameters for OBSP cancer screens by Mammography (Mam'y, $j = 1$) and Clinical Breast Exam (CBE, $j = 2$).

women, a result which applies to both screening modalities.

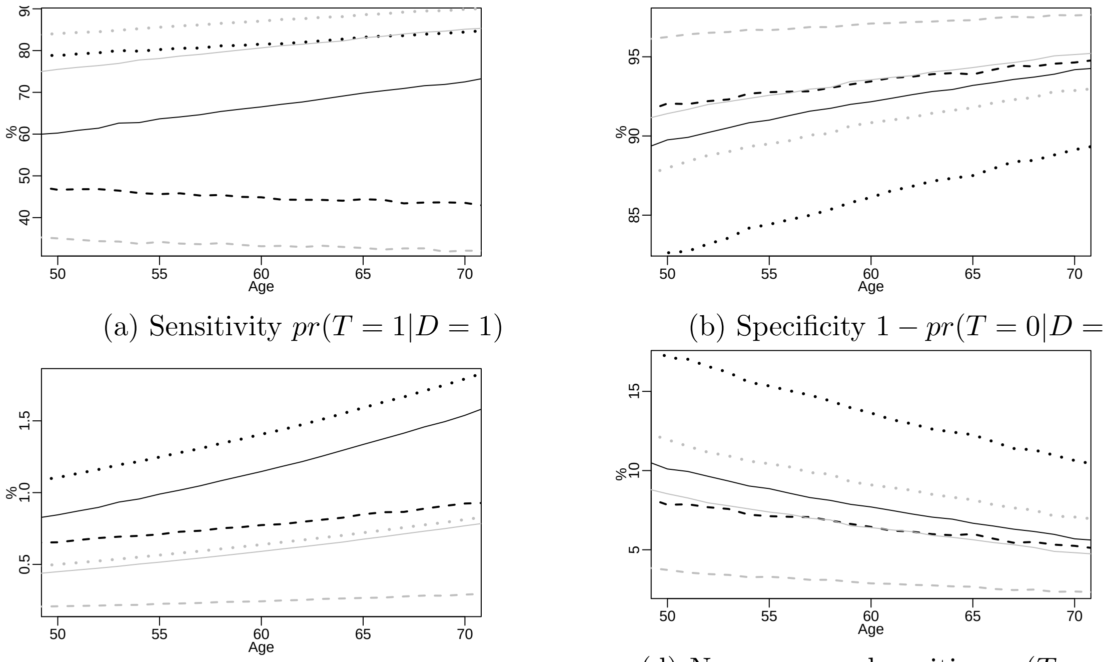
Figure 2.2b shows the prior and posterior distributions for the standard deviations and correlations of the random effects. Standard deviations for eight different random effects are shown: four at the examiner-level sd[$\eta_{js_ie_{ij}}$] and sd[$\theta_{js_ie_{ij}}$] for $j = 1, 2$; and four at the site level sd[κ_{js_i}] and sd[λ_{js_i}]. The amount of variation between sites and examiners is substantial, considering these random effects operate on the logit scale, with the posterior densities shifted to the right from their fairly informative prior distributions. Most of the posterior distributions are concentrated in excess of 0.5, a large value considering two standard deviations on the log-odds scale corresponds to a relative odds of $e^{2\cdot 0.5} = 2.7$ on the natural scale. The heterogenity of examiners and sites is generally greater for CBE than for mammography.

There are four correlation parameters shown in Figure 2.2b, showing the correlation between the odds of true positives and false positives cor$[\eta_{js_ie_{ij}}, \theta_{js_ie_{ij}}]$ (examiners) and cor$[\kappa_{js_i}, \lambda_{js_i}]$ (sites) for each screening modality ($j = 1, 2$). The positive site-level correlation indicates that a screening site s with a high false positive rate (and a positive λ_{js_i}) is likely to have a high true positive rate (with a positive κ_{js_i}). There is evidence for positive correlation at the site-level for CBE, suggesting screening sites obtain high true positive rates for CBE at the cost of higher accompanying false positive rates.

2.3.2 Sensitivities and specificities

Screening outcomes are affected by individual characteristics such as breast density and family history, and results are shown for women in a "high risk" group (with high breast density and having a first-degree relative with breast cancer) and a "low risk" group (having neither). Figures 2.3a and 2.3b show predicted sensitivity and specificity of screening tests involving CBE and/or mammography as a function of an individual's age for the two groups. Mammography (solid lines) has consistently higher sensitivity than CBE (dashed lines), though specificity is somewhat lower for the former. Unsurprisingly, undergoing both exams (dotted lines) increases sensitivity and decreases specificity. Tests on women in the high risk group (black lines) have lower specificity than tests on women in the low risk group (grey lines) for both screening modalities, while sensitivity in the former group is lower for mammography and higher for CBE.

Figures 2.3c and 2.3d show joint probabilities of an individual having cancer and obtaining a positive screening test (Figure 2.3c) and an individual not having cancer incurring a false positive test (Figure 2.3d). The joint cancer-true-positive probability in Figure 2.3c increases and the joint non-cancer-false-positive probability in Figure 2.3d decreases for both risk profiles. A pertinent practical question these Figures can inform is whether an individual should opt for obtaining both CBE and Mammography or forgo CBE in favour of Mammography alone. Women in the high risk group have a substantially higher joint cancer-true positive probability with two exams than when undergoing Mammography alone, as evidenced by the large gap between the dotted and solid black lines in Figure 2.3c. The increase in this joint probability for low risk women, shown by the dotted and solid grey lines, is slight. The rationale for undergoing both exams is therefore stronger for the high-risk women, though the combination of exams also increases the undesirable joint false-positive probability in 2.3d.



(a) Sensitivity $pr(T = 1|D = 1)$

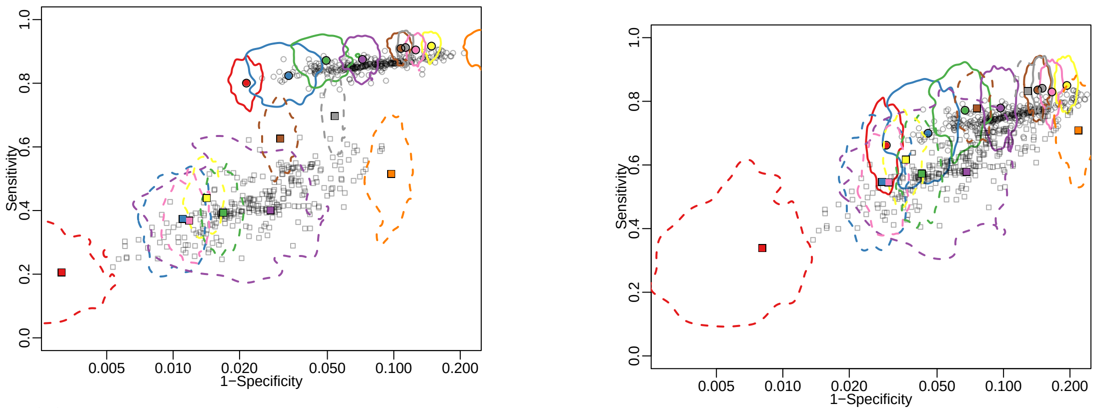
(b) Specificity $1 - pr(T = 0|D = 0)$

(c) Cancer and positive $pr(T = 1, D = 1)$

(d) No cancer and positive $pr(T = 1, D = 0)$

Figure 2.3: Sensitivity and specificity for mammography (—), CBE (- - -) and both (. . .) for high risk women with high breast density and a family history of cancer (black lines) and low risk women with low density without family history (grey lines).

The nature of variability amongst examiners can be inferred from the posteriors of covariance parameters in Figure 2.2b, though a more intuitive illustration using sensitivities and specificities is presented in Figure 2.4. Each plotting symbol in Figure 2.4 shows the posterior mean of the sensitivity and specificity for an individual examiner seeing a 65 year old individual, with Figure 2.4a relating to low risk (low breast density, no family history) individuals with probabilities for high risk individuals (high density, with a family history) shown in Figure 2.4b. For an arbitrary selection of examiners, 75% posterior credible regions



(a) Low breast density, without family history

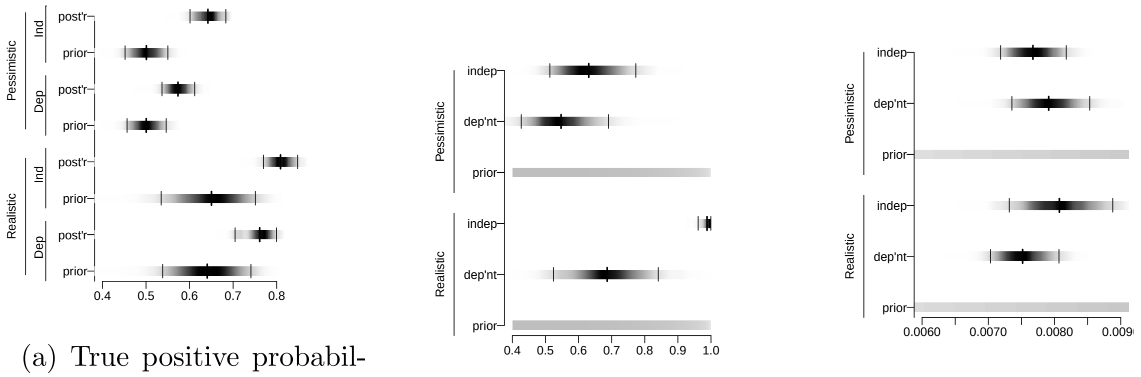
(b) High breast density, with family history

Figure 2.4: Posterior means for sensitivity and specificity of individual examiners screening a 65 year-old patient with CBE (□) and Mammography (○). Coloured elements show 75% posterior credible regions and their corresponding posterior means for CBE (dashed lines) and Mammography (solid lines). Note that the x-axis is on the log scale.

are shown as contour lines. Notice the clear separation between Mammography and CBE, with Mammography examiners having uniformly higher sensitivity than CBE and specificity which is on average lower. Also note that sensitivity is fairly stable amongst examiners (although more variable for CBE), whereas specificity is extremely heterogeneous. The differences in accuracy amongst examiners is therefore most evident in their false positive rates, with the cancer-detecting abilities of the most and least effective examiners being reasonably similar. The positive correlation parameters in Figure 2.2b are reflected in upward slopes in Figure 2.2b.

2.3.3 Prior sensitivity

To assess the influence and importance of prior distributions and the assumption of independence between screen types, posterior distributions from four variations on the screening model were computed. A “Pessimistic” model specifies a prior distribution for the μ_1 parameter which results in a true positive probability for mammography on a baseline individual tightly distributed around 50%, as opposed to the 54% to 73% range under the “Realistic prior” assumptions in Section 2.2.2. The independence assumption was relaxed by fitting the “Dependence model” from (2.5), with an individual-level random effect allowing for the possibility that a tumor which escapes detection from one screening modality is also more likely to escape detection by a second. This model and the standard “Independence model” were fit using both the “Realistic” and “Pessimistic” set of prior distributions. Figure 2.5a shows the distribution of the true positive probability $pr(T_{ij} = 1|D_i = 1)$ for mammography, on an individual with baseline covariates, from both prior and posterior samples. The priors for the dependence model differ slightly from the independence model as for the former the individual-level random effect is integrated out. Figure 2.5b shows the posterior probability for the proportion ρ of cancers missed by screening which are subsequently detected by other means. The baseline cancer rates on the natural scale are contained in Figure 2.5c.



(a) True positive probability $pr(T_{ij} = 1|D_i = 1)$ for mammography on an individual in the baseline group.

(b) Clinical detection rate for missed cancers (ρ)

(c) Baseline cancer rate $\exp(\mu_1)/(1 + \exp(\mu_1))$, see Equation (2.3)

Figure 2.5: Prior and posterior distributions for models with realistic and pessimistic prior distributions, assuming independence between radiographers and nurses and allowing for dependence via an individual-level random effect.

In each of the four models mammography sensitivity is predicted to be at the upper end of the prior distribution. The Dependence model leads to estimates that are lower and closer to the priors than the independence model, a somewhat intuitive result as the additional σ parameter allows for greater flexibility. Also unsurprising is a lower posterior sensitivity in a model is accompanied by a higher number of unobserved cancers (a lower ρ) and a higher cancer rate. As the number of missed cancers is small, the baseline cancer rate need only shift by a small amount in order to accommodate lower sensitivity. For example, the 95% credible interval moves from (0.0070, 0.0081) to (0.0074, 0.0085) for the dependence model.

One conclusion to draw from Figure 2.5 is that the choice of prior does affect the inferences made regarding sensitivity of a screening test when the number of unobserved cancers is unknown. The posterior distribution for ρ suggests an analysis of the OBSP data using a simpler model with an implicit assumption of all cancers being observed is not unreasonable if independence and the "Realistic" priors are also assumed. However, one would not be able to assess the appropriateness of this simpler model without first having carried out an analysis where the number of unobserved cancers is allowed to vary.

With respect to the independence assumption, Figure 2.5a suggests that this assumption does not artificially inflate the estimate of sensitivity if one accepts the findings of clinical research on mammography sensitivity (the "Realistic" priors). Were the "Pessimistic" prior assumptions to be believed, the data would be inconsistent with the independence assumption and an individual-level random effect in the Dependence model would be required to induce a number of missing and unobserved cancers.

2.4 Discussion

With electronic medical records and linked health databases becoming increasingly common, large population-based observational datasets similar to the OBSP records will become increasingly available. The individuals in such databases are heterogeneous and the outcome variables can be incomplete, though this paper has demonstrated with the OBSP data that explanatory variables, random effects, and latent variables for disease status can accommodate these complexities. Understanding the sources of variation in large screening databases can yield useful information for managing and improving screening programs, one example of which is provided by the estimates of examiner-level variance parameters. Examiner-level variations in false positive rates are substantially greater than the corresponding variations in true positive rates, implying that training and resources would be most effective when targeted at those examiners for whom false positive rates are the highest. Quantifying the influence of individual-level characteristics on screening outcomes can aid the development of individualized cancer screening policies and recommendations. Current guidelines from the American Cancer Society and the Canadian Cancer Society make breast cancer screening recommendations based on a woman's risk factor profile. The impact a woman's personal characteristics have on test sensitivity and specificity could also be reflected in screening recommendations with, for example, low-risk individuals encouraged to obtaining screening only if their predicted false positive probability is small.

There is a fundamental non-identifiability involved in estimating both the cancer detection rate and the prevalence of unobserve cancers from a single dataset, as the number of cancers observed may increase due to an increase in the detection rate or a decrease in the number of unobserved cancers presumed. This issue is addressed here through the use of reasonably strong prior information on test accuracy obtained from clinical studies. Some of this non-identifiability would be reduced when more than one screening test is administered and an assumption of independence between screens is made, though Dendukuri and Joseph (2001) discuss in some detail how substantial prior sensitivity remains when the number of screening tests is less than three. Informative prior distributions are therefore necessary for some of the model parameters. Inferences using this model are sensitive to these prior assumptions, with the 95% posterior credible interval (1%, 25%) for the proportion of screen-undetected cancers which are unobserved during follow-up changing to (25%, 50%) when a pessimistic prior positing a low true positive rate is substituted. Drawing conclusions from an analysis which relies on subjective prior information does provoke skepticism, though conclusions from an identifiable model with an implicit assumption that all screen-undetected cancers are observed should also be treated skeptically.

There can be subtleties involved with incorporating prior information into complex random effects models, with Menten *et al.* (2008) demonstrating how different dependence models may result in similar fits to the data while resulting in different inferences. They concluded that the selection of appropriate latent class models should be based on substantive subject matter knowledge. Our modelling framework can assist in this regard by explicitly separating true and observed disease status (D_i and Y_i), and specifying a model for test results conditional on true cancer status. To some degree, sensitivity to model specification is explored with the inclusion of a random effect term allowing for dependence between screening tests performed on the same individual. Although this dependence term results in a substantial increase in the number of unobserved cancers predicted (from 1% - 20% to 20% - 40%), the increase is not sufficiently large to cause more than a modest effect on the posterior for test sensitivity. There is certainly scope for a more detailed exploration of the possible dependence structures between screening tests with population-level data, such as the negative correlation between mammography and CBE noted by Walter *et al.* (2012), though the fundamental non-identifiability expressed by Dendukuri and Joseph (2001) is a substantial limitation for research in this regard.

A number of model extensions would be possible to more faithfully reflect the underlying biological processes of cancer incidence and screening outcomes. Variation in cancer risk throughout the population could be accommodated by including additional covariates or random effects terms in the model for ψ_i (Equation 2.3) and testing for bimodality or skewness in examiner-level random effects would result from using more flexible distributions or mixtures in place of the Gaussian. Such extensions are conceptually straightforward, and the size of the dataset would likely support the identification of these additional parameters. The computational challenges would be greatly increased, however, and a more sophisticated MCMC updating scheme would be required in place of the Metropolis-within-Gibbs algorithm used here.

Finally, no discussion of cancer screening methodology can avoid addressing the controversy regarding the degree to which cancer detection via screening ultimately provides a health benefit (see Bleyer and Welch, 2012, for example). Our model provides an evaluation of a screening program or test based on the simplest and most immediately obtainable success criterion, which is detection of lesions leading to a clinical diagnosis of cancer. However, the accommodation of heterogeneities amongst subjects and examiners in this model could form the basis of extended and expanded models for a more complex set of hypotheses and health outcomes. For example, the degree to which screen detection of cancers is overdiagnosis could be assessed by exploring the relationship between an examiner's test sensitivity and the long-term health status of the individuals they screen. Although the variation in examiners' sensitivities is shown here to be modest (in the range of 10% to 20%), the large sample sizes in population-level datasets with random assignment of individuals to examiners within centres suggest that such an analysis would be insightful.

Acknowledgements:

We would like to thank Anna Chiarelli for the provision of the data. The research project was supported by grants from the Canadian Breast Cancer Foundation Ontario Region and the Natural Sciences and Engineering Research Council. The authors also thank the Ontario Breast Screening Program, a program of Cancer Care Ontario, for use of its data for this study. The Associate Editor and two anonymous reviewers deserve particular recognition for their attention and suggestions for improvements to this paper.

Chapter 3

Estimation of the Benefit of Including Clinical Breast Examination in an Organized Breast Screening Program

Abstract

Background: There is controversy about the value of clinical breast examination (CBE) in addition to mammography. Estimation of the accuracy of breast screening using population-level data is complicated by heterogeneity of the case mix and by the lack of independent confirmation of exams. **Methods:** To compare the accuracy of referral with and without CBE, a cohort was identified from information collected by the Ontario Breast Screening Program (OBSP). The cohort consists of women 50 to 69 years of age screened at the OBSP between January 1, 2002 and December 31, 2003. The associations between patients, radiologists and screening centres were investigated using a joint logistic regression model that accommodates the partially unobserved disease status, clustered data structures, individual risk factors, and the dependence between true and false detection. Bayesian inference was applied to approximate the posterior distributions of the diagnostic error rates and disease prevalence. We assumed there were four groups of women varying by age and risk factors including hormone therapy, breast density, family history, examined by an average radiologist from an average centre. Measures of test accuracy were predicted using the estimated coefficients and random effects obtained from the model. **Results:** Among the four assumed groups, in the group of women at high risk, the rate of cancer detection can be increased from an average of 1.46 to 25.3 per 1,000 compared with the group at low risk. We found that women who are currently using hormone therapy gain most sensitivity among the four groups. Women with dense breast tissue benefit the least from increased sensitivity when CBE is added, and experience the second most harm due to false-positive results among four groups of women. **Conclusions:** When CBE is offered in addition to mammography, it may benefit older high-risk women and women using hormone therapy but probably does not benefit women with dense breast tissue.

KEYWORDS: mammography, breast clinical examination, clustered analysis, detection rate, sensitivity, specificity

3.1 Introduction

Although there are studies which have found that CBE can detect cancers that are missed by mammography (Miller *et al.*, 1991; Green and Taplin, 2003; Barton *et al.*, 1999), the value of CBE in addition to mammography has not been thoroughly established. An evaluation conducted by the International Agency for Research on Cancer (IARC) in 2002 showed there is a lack of evidence that screening with CBE, either alone or in addition to mammography, can reduce mortality from breast cancer. The IARC noted that CBE may be important in countries where there are insufficient resources for mammography, or where disease is usually at an advanced stage at the time of diagnosis. Two studies examining the evidence from a number of breast screening trials of mammography (Humphrey *et al.*, 2002; Kerlikowske *et al.*, 1995) found that the decrease in breast cancer mortality in the four trials that included

CBE in addition to mammography were similar to those in trials that included mammography only. The contribution of CBE has varied significantly by study: the contribution of CBE alone to breast cancer detection has ranged from 3% to 45% in randomized clinical trials (RCTs) (Humphrey *et al.*, 2002).

Even though a complete assessment of screening has to be in terms of mortality, measures of screening accuracy such as sensitivity and specificity can be used as interim indicators of effectiveness so that screening methods can be evaluated and compared sooner. “The cancer detection rate and screen sensitivity determine whether screening is effective in detecting women with breast cancer”. When the sensitivity is low, the screening modality produces a high rate of false-negative results or interval cancers that contribute significantly to mortality in the screened population (Schroen *et al.*, 1996; Brekelmans *et al.*, 1996). Screen specificity is related to the efficiency, or the ability to quickly detect women with breast cancer. When the specificity is low, the screening modality produces a high rate of false-positive results. In such a scenario, more women may be inappropriately referred for further diagnostic tests that are potentially invasive. (Chiarelli *et al.*, 2006, 2009)

Three studies in community settings have looked at the contribution of CBE in addition to mammography for women 50 to 69 years of age. A US study (Bobo *et al.*, 2000) found an additional 2.6 cancers per 10,000 mammography screenings. A Canadian study (Bancej *et al.*, 2003) found an additional 3 cancers per 10,000 screenings. A recent Ontario study, using highly trained nurses, found an additional 4 cancers per 10,000 screened. However, achieving these results comes at a cost, as there would be an estimated additional 219 false positives cases being found (Chiarelli *et al.*, 2009). It has been found that the performance of CBE and mammography is influenced by patient age and breast density and changes with risk factors related to examiners, such as training background, working experience, and length of the examination (Oestreicher *et al.*, 2005; Barton *et al.*, 1999). However, there was no strong evidence that greater practice volume or experience at interpreting mammograms is associated with better performance (Barlow *et al.*, 2004). It is often difficult to estimate the contribution of CBE in an organized screening program since the screening techniques are performed by different examiners in different locations. The accuracy of mammography or CBE varies according to factors related to the patients, examiners and centres. Statistical methodology needs to be selected carefully in order to deal with the clustered data structure.

The other difficulty in assessing the sensitivity of an exam with population-level screening data is that the true cancer status of screened participants is often unknown. Ideally, a screening exam would be compared to an accepted “gold standard”, for instance pathological classification of tissue biopsy for women having undergone a breast cancer screen. However, it is infeasible to apply the gold standard to all individuals due to limited resources, ethics or practicality issues, and therefore the true disease status could be detected by screening exams, missed by screening exams but shown clinically as interval cancers, or missed without clinical diagnosis. In this circumstance, the true sensitivity and specificity are unknown, and the partly missing data on true cancer status might lead to biased estimation of the performance measures of a screening exam.

This chapter describes an epidemiological study that builds on the methods developed in Chapter 2. The objective is to investigate and estimate the effects of various risk factors such as age, breast density, hormone use and family history on the value of CBE in addition to mammography in the OBSP program. A joint mixed-effects logistic regression model developed in Chapter 2 is used to estimate the effects and overcome the difficulties in the analysis of clustered data with partly missing true cancer status.

3.2 Subjects and methods

3.2.1 Study population

The OBSP has operated since 1990 to deliver a population-based breast screening program. The OBSP offers service to a population with geographic and social diversity, which leads to a unique base for building a cohort to investigate the effect of CBE and mammography in a community setting. A complete description of the operation of the OBSP and the cohort has been published elsewhere (Chiarelli *et al.*, 2006, 2009), and these are summarized and defined below.

The explanation of this data linkage can be found in Chapter 2 and Appendix A. The two types of screening protocols included in various OBSP centres were defined as mammogram only and CBE in addition to mammogram. At centres without a nurse, a woman could only be referred by a radiologist. At centres with a nurse, a CBE is performed "in the upright, supine and lateral oblique positions and includes palpation for nodes in both the axillae and supra/infra clavicular areas". The nurse makes an independent decision based on the CBE findings as to whether the woman requires further assessment. A woman may be referred by either the radiologist or the nurse. Referral information is collected in a standardized method on the screening report and recorded as "normal" or "recommended clinical assessment by physician" by the nurse and as "normal/benign" or "needs additional evaluation by imaging and/or surgical consultation" by the radiologist" (Chiarelli *et al.*, 2006, 2009).

Women 50 to 69 years of age screened at the OBSP between January 1, 2002 and December 31, 2003 was selected and followed for up to 12 months after their last screening examination. "Women who participate in OBSP must be residents of Ontario, have no history of breast cancer or augmentation mammoplasty and be free of acute breast symptoms. The OBSP offers all eligible women biennial screening consisting of two-view mammography. Women considered at high risk of breast cancer are recalled in one year" (Chiarelli *et al.*, 2006, 2009). Of the 102 OBSP screening centres in operation during the study period, 73 (71.6%) centres offers CBE in addition to mammography.

Between 2002 and 2003, the OBSP provided 343711 screens to 301362 women aged 50 to 69 who had complete follow-up. The cohort excludes women who took screening exams not consistent with the current protocols of the centres they attended. Of these women, 4604 were excluded as they had a mammogram and CBE from an OBSP centre that no longer offered CBE; 6520 women were excluded as they only had a mammogram in centres that offered both mammography and CBE, 8 women were excluded as they attended an affiliate that screened less than 10 eligible women, and 4 women were excluded as they only had a CBE in screening centres. The final sample size for analyses was 290226 women.

3.2.2 Risk factors

We adopted the definitions of risk factors and screening visit characteristics described by Chiarelli *et al.* (2009), except for the definitions of certain performance measures, as described later: “Information on risk factors for breast cancer was based on self-reported data collected at the last screening appointment through a personal interview with the nurse or technologist. Women with at least one first degree relative with breast cancer were classified as having a positive family history. Women were defined as current users of hormone therapy if they were taking it at the time of their last screen. Mammographic density was recorded by the radiologist as $\leq$75% or $>$ 75%, based on findings from the mammogram”. For women with a breast cancer (screen-detected=2196; interval=197), the screen prior to diagnosis was included and for women with more than one screen during the study period the last screen (N=290006) was included and recorded as “rescreen=1”. For women with one screen during the study period, that single screen was included and recorded as “rescreen=0”.

Age at last screen was defined as the age of the woman at her last OBSP screening examination from January 2002 to December 2003. The average annual volume of screens was calculated as the number of screens delivered by the centre between 2002 and 2003 divided by the number of years in operation over the time period. A logarithmic transformation has been applied to produce an approximately normal distribution for the transformed data. Years of operation refers to the exact number of years the centre operated as part of the OBSP from 1990 to 2003. Age, annual volume of screens, and years of operation were standardized by subtracting their means then dividing by their standard deviations.

Table 3.1: Definition of performance measures

Characteristic	Mammogram only	CBE in addition to Mammogram
True positive	Cancer after abnormal mammogram	Cancer after either abnormal
False negative	Cancer after negative mammogram	Cancer after both negative
True negative	No cancer after negative mammogram	No cancer after both negative
False positive	No cancer after abnormal mammogram	No cancer after either abnormal

3.2.3 Performance measures

“Cancer detection rate” was defined as the number of women detected with invasive or ductal carcinoma in situ cancer per 1,000 women screened (True positive/Total Participants). For both types of screening protocols, sensitivity was defined as the proportion of women with a breast cancer who had a positive screening exam (True positive/ (True positive + False negative)). The true positive results or screen-detected cancers included breast cancers diagnosed within 12 months after a recall by a radiologist or a nurse. The false negative results included breast cancers diagnosed after a negative recall by a radiologist when only mammogram was applied; or a negative mammogram and a negative CBE when both exams were applied. Specificity was defined as the proportion of women without a breast cancer who had a negative screening exam (True Negative/(True Negative + False Positive)). The true negative results included women without a breast cancer diagnosis within 12 months after a normal mammogram and a normal CBE, if applied. The false positive results included women without a breast cancer diagnosis after a positive mammogram or a positive CBE result. Those definitions are summarized in Table 3.1.

3.3 Statistical model

The model from Chapter 2 considers three partially unobserved processes: the cancer incidence D_i for individual i; recall due to correct identification of a cancerous nodule as a result of mammography or CBE, denoted $pr(T_{i1} = 1|D_i = 1)$ and $pr(T_{i2} = 1|D_i = 1)$ respectively; and recall without having identified a nodule by mammography or CBE, or

$pr(T_{i1} = 1|D_i = 0)$ and $pr(T_{i2} = 1|D_i = 0)$. The probability $pr(T_i = 1|D_i = 1)$ is termed a *true positive rate*, and it applies to individuals with cancer. The probability $pr(T_i = 1|D_i = 0)$ is termed a *false positive rate*, and it most commonly results from an examiner recalling a patient without cancer, having misinterpreted the screening results.

Instead of modeling sensitivity and specificity directly, we modeled the latent variables true positive, false positive and real cancer status jointly and derived the performance measures such as sensitivity and specificity. Understanding the behavior of the true positive, false positive and real cancer status enables enhanced insight into the joint behaviors of the performance measures such as sensitivity and specificity. For example, high breast density can influence screening outcomes in three possible ways: breast density affecting cancer risk through the cancer incidence; high density making lesions more difficult to detect on screening images through the true positive process; and high density causing spurious features on screening images causing recalls through the false positive process. Using three logistic models (Chapter 2 Formula 2.3) for these processes yields three coefficients for the breast density variable, disentangling these three effects at the same time. In addition, by modeling true positives and false positives jointly, the model is able to assess the nature of the dependence between these two probabilities at both the examiner and screening centre levels (see Chapter 2 for further statistical details).

The same models were applied separately to the two populations attending screen centres offering both mammography and CBE or mammography only. Inference on the mixed-effects logistic regression models was done using a Bayesian MCMC algorithm in the software package R (www.r-project.org R Core Team (2014)). Computational details had been presented in Chapter 2. The mixed-effects logistic regression models were adjusted for factors related to the woman and to the screening centre she attended. To control for clustering of women and providers within screening centres all models included random effects for each radiologist and centre. The results from these models are presented in Section 3.4.

To assess the performance of radiologists, one can compute posterior distributions of sensitivities, specificities and predictive values through the true positive and false positive probabilities (see Chapter 2 for computational details). Since the recall accuracies of radiologists depends on their own pools of screenees, a better way of assessing performance of radiologists is to investigate case ascertainment when considering equivalent screenees. In light of this, we predicted the performance measures for populations of women varying by age and risk factors including hormone therapy, breast density, family history. Based on the estimated risk factor effects, four groups of women age 50 to 69 are considered: first women at high risk (defined as women with breast density ≥75%, having family cancer history and current using hormone therapy); second women at low risk (defined as women without high breast density, family history or current hormone therapy); third women with high breast density but without family history or hormone therapy; fourth women currently using hormone therapy but without family history or high breast density. Screening performance for all groups of women was assessed by adopting the estimate of an average radiologist (or nurse) from an average centre, and using the estimated coefficients and random effects obtained from the joint logistic regression model.

3.4 Results

Among the 290226 women selected, 56049 women had a mammogram only among which 3936 (7.02%) women were referred to further exams; 234177 women had both mammogram and CBE among which 16523 (7.06%) women were recalled. Of 102 centres included for analysis, 73 centres offered both mammogram and CBE (Table 3.2). On average, those centres offering CBE have almost twice the annual screening volumes compared to centres providing mammogram only. They are more evenly geographically distributed than centres without CBE that mainly locate at the central east of the province. For example, all 18 centres in the south west provided both exams. A total of 218 radiologists work in the centres providing both modalities, while 106 radiologists work in the centres providing mammography only.

Table 3.2: Characteristics of radiologists and OBSP screening centres

Characteristic	Centres without CBE	Centres with CBE
	N=56049	N=234177
Number of participants being recalled	3936	16523
Number of centres	29	73
Average annual screening volume		
median(range)	962.47(69.91-5402.59)	1296.66(129.26-13717.17)
Administrative Region, N(%)		
Central East	24(82.75)	10 (13.70)
Central West	2(6.89)	11 (15.07)
Eastern	1(3.44)	7(9.59)
North East	2(6.89)	10(13.70)
North West	0(0.00)	4(5.48)
South	1(3.44)	4(5.48)
South East	1(6.89)	9(12.33)
South West	0(0.00)	18(24.66)
Number of radiologists	106	218
Radiologist years of experience		
in screening program, mean(SD)	3.27(2.98)	4.77(3.98)
Number of nurses		167
Nurse years of experience		
in screening program, mean(SD)		3.94(3.26)

The average working experience in screening program is 4.77 years in the former, which is slightly longer than the average of 3.27 years in the latter.

Table 3.3 summarizes the screening referral patterns in subgroups of participants with different risk factors. The recall rates for women with two exams are close to the rates with one exam for most of the subgroups. For example, the recall rate is 6.82% for women with low breast density when screened in centres without CBE and 6.97% when screened in centres with CBE. However, when screened for the first time, the recall rate was 8.01% for women with one exam and 10.12% with two exams, which is to be expected because of a larger proportion of prevalent cases.

The result of the mixed-effects model was presented in Table 3.4. Similar to the regular logistic model, the odds ratio represents the odds that a true positive or a false positive

Table 3.3: Summary of cancer outcomes and characteristics of patient

	Centres without CBE n=56049				Centres with CBE n=234177			

	N(%)	Recalled
		Rate(per 100)
			Cancer	Interval
				Cancer
					N(%)	Recalled(%)
						Rate(per 100)
							Cancer	Interval
								Cancer

First degree relative with breast cancer								
No	50501(90.1)	7.04	316	45	205639(87.8)	0.74	1517	116
Yes	5548(9.9)	6.79	58	7	28538(12.2)	1.07	305	29
Breast density								
$<$75%	49657(88.6)	6.82	331	11	220456(94.1)	0.75	1664	118
$\geq$ 75%	6392(11.4)	8.57	43	41	13721(5.9)	1.15	158	27
Rescreen								
First	30532(54.5)	8.01	229	28	65936(28.2)	0.89	586	38
Subsequent	25517(45.5)	5.84	145	24	168241(71.8)	0.73	1236	107
Current use of hormone therapy								
No	42424 (75.7)	6.93	270	32	163650(69.9)	0.71	1170	82
Yes	12458 (22.2)	7.33	95	19	69370(29.6)	0.92	640	63
Unknown	1167 (2.1)	6.94	9	1	1157(0.5)	1.04	12	0

occurs given the presence of a particular exposure, compared to the odds occurring in the absence of that exposure. After adjusting for personal and facility characteristics, women using hormone therapy who attended centres offering mammogram only had a significantly lower odds of a true positive (OR=0.47; 95% credible interval(CI), 0.23-0.95) than women not using hormone therapy, and had a significantly higher risk of being recalled without cancer being detected (OR=1.15; 95% CI, 1.06-1.25) (Table 3.4). Similarly, women using hormone therapy who attended centres offering both exams also had a significantly lower risk of being recalled with detection (OR = 0.67; 95%CI, 0.47-0.95) and a higher risk of being falsely recalled (OR = 1.26; 95%CI, 1.22-1.30). Women with high breast density had a significantly higher risk of being recalled without cancer being detected (OR=1.42; 95% CI, 1.27-1.59 for mammogram only; OR=1.59; 95% CI, 1.50-1.69 for both exams).

There was more variation among the examiners than among the centres. At the radiologist's level, the probabilities of being recalled with or without cancer being detected were negatively correlated, which means the more true detections a radiologist made, the fewer false detections he/she made. However, at the level of screening centres, recall with or without cancer being detected were not highly correlated, and the variation among centres was fairly small.

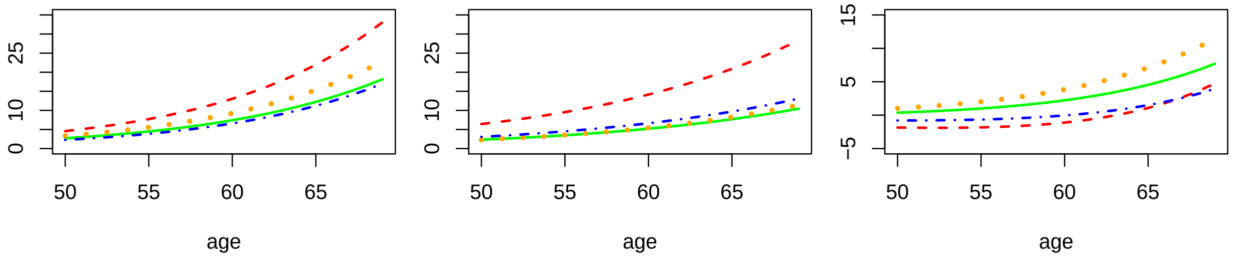
Types of Exams	Mammogram n = 54882	Mammogram and CBE n = 233020
True positive		
breast density(y/n)	0.80(0.24, 2.38)	0.74(0.42, 1.37)
age	1.41(1.00, 2.01)	1.50(1.26, 1.79)
family history(y/n)	0.71(0.28, 1.81)	1.38(0.87, 2.12)
hormone therapy(y/n)	0.47(0.23, 0.95)	0.67(0.47, 0.95)
rescreen(y/n)	0.68(0.30, 1.50)	0.72(0.51, 1.02)
radiologist experience(in yrs)	1.02(0.55, 1.92)	0.98(0.73, 1.31)
screen volume(log-transformed)	1.00(0.72, 1.40)	1.01(0.86, 1.19)
False positive		
breast density(y/n)	1.42(1.27, 1.59)	1.59(1.50, 1.69)
age	0.97(0.94, 1.01)	0.93(0.92, 0.95)
family history(y/n)	1.02(0.91, 1.17)	1.06(1.01, 1.11)
hormone therapy(y/n)	1.15(1.06, 1.25)	1.26(1.22, 1.30)
rescreen(y/n)	0.55(0.51, 0.60)	0.59(0.57, 0.60)
radiologist experience(in yrs)	1.02(0.58, 1.82)	1.01(0.74, 1.38)
screen volume(log-transformed)	1.01(0.79, 1.31)	1.00(0.90, 1.12)
Variance at radiologist level		
true positive	3.11(2.52, 3.79)	2.15(1.85, 2.49)
false positive	3.28(2.79, 3.81)	2.31(2.06, 2.60)
correlation	-0.35(-0.63, -0.06)	-0.38(-0.61,-0.14)
Variance at centre level		
true positive	0.89(0.37, 1.64)	0.57(0.30, 0.95)
false positive	0.64(0.31, 1.13)	0.38(0.23, 0.59)
correlation	-0.10(-0.83, 0.75)	-0.03(-0.68,0.61)

Table 3.4: Odds ratios with 95 percent credible intervals and estimated variance for risk of being recalled with cancer being detected and without cancer being detected, by modalities of referral

3.5 Estimation of performance measures

3.5.1 Cancer detection rate

The benefit of offering CBE in addition to mammography varies by age. For women ages 50 to 69 and at high risk, the rate of cancer detection increases from an average of 5.0 to 32.8 per 1000 depending on age at the first screen with both exams, compared to an average of 6.5 to 27.5 with mammography only. However, when examining the gain of the detection rate (Figure 3.1), women who currently having hormone therapy benefit most among all four groups. Their detection rates increase from a range of 1.0 to 11.4 per 1000 depending on age at the first screen.



(a) Mammography and CBE

(b) Mammography only

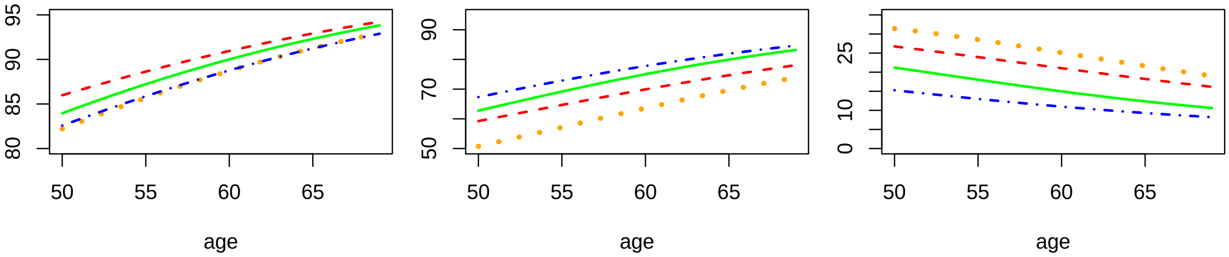
(c) The difference between two protocols

Figure 3.1: The detection rates for first screens: low risk women with low breast density, without family history and without hormone therapy (—), high risk women with high breast density, having family history of cancer and currently using hormone therapy (- - -) , women currently using hormone therapy, with low breast density and without family history (. . .), women with high density breast, without family history and without hormone therapy (- · - · -).

3.5.2 Sensitivities

For all women aged 50 to 69, the sensitivity increases when CBE is provided in addition to mammography and the gain in sensitivity decreases with age (Figure 3.2). It appears that CBE generally adds incrementally more to sensitivity among women with high risk. For the group of women aged 50 to 69 at low risk being screened for the first time, the sensitivity of mammography with CBE ranges from 83.9% to 93.8% while that of mammography alone ranges from 62.7% to 83.2%(Figure 3.2(a)(b)). For women at high risk, it ranges from 85.9% to 94.2% for both exams and from 59.2% to 78.1% for mammography alone. For a woman aged 60 at low risk, an average of 14.0% can be gained by adding CBE while a much higher increase of 20.0% is estimated for a woman of the same age at high risk.

Women with dense breast tissue benefit least among the four groups, with an average increment of 0.8% to 15%. There is also a notable levelling-off effect among groups: differences among groups become smaller as participants grow older (Figure 3.2(c)). For example, the difference in sensitivities between women currently using hormone therapy and women with dense breast tissue is 16.2% at age 50 compared with 10.8% at age 65.



(a) Mammography and CBE

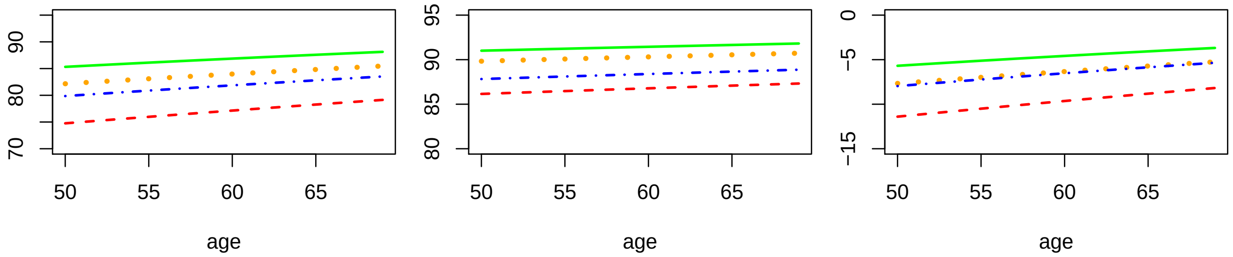
(b) Mammography only

(c) The difference between two protocols

Figure 3.2: The sensitivities with and without CBE for first screens: low risk women with low breast density, without family history and without hormone therapy (—), high risk women with high breast density, having family history of cancer and currently using hormone therapy (- - -) , women currently using hormone therapy, with low breast density and without family history (. . .), women with high density breast, without family history and without hormone therapy (- · - · -).

3.5.3 Specificity

Specificity declines when CBE is used in conjunction with mammography, and this decrement is more pronounced in women with high risk (Figure 3.3). For the high risk group, the average loss of specificity is 8.1 - 11.4% and it is 3.7 - 5.6% for the low risk group for women ages 50 to 69. Similar to sensitivity, the gap between specificities of the high risk and low risk women is slightly narrower at higher ages when CBE is provided in addition to mammography: for example, the gap decreases from an average of 5.8% in women aged 50 to an average of 4.5% in women aged 69 for the first screen.



(a) Mammography and CBE

(b) Mammography only

(c) The difference between two protocols

Figure 3.3: The specificities with and without CBE for first screens: low risk women with low breast density, without family history and without hormone therapy (—), high risk women with high breast density, having family history of cancer and currently using hormone therapy (- - -) , women currently using hormone therapy, with low breast density and without family history (. . .), women with high density breast, without family history and without hormone therapy (- · - · -).

3.6 Discussion

This chapter uses the methodology developed in Chapter 2 to conduct a thorough investigation of the outcome of breast screening examinations in the OBSP cohort, with a specific emphasis on the incremental benefit of CBE. Our study found a number of individual-level risk factors affecting screening results for both protocols. Having high breast density, family history and using hormone therapy generally increase a woman's probability of being recalled correctly, though neither breast density nor previous screens appear to significantly affect detection probabilities. Our findings agree with several studies that have shown that current use of hormone therapy substantially lowered the true detection rate of cancers while it increased the probability of false detection compared to non-use for both screening protocols (Séradour *et al.*, 1999; Litherland *et al.*, 1999; Kavanagh *et al.*, 2000; Banks, 2001). Although high radiographic breast density was associated with decreased probability of being recalled with cancer being detected and increased probability of being falsely recalled, our findings showed that the effects were not significant after adjusting for other risk factors.

After adjusting for various provider and facility characteristics, this study found that the addition of CBE to mammography improved the sensitivity of women receiving both exams, but these women were also subject to increased risk of a false positive test. Among the four groups of women we examined, the results suggest that CBE generally improves sensitivity and reduces specificity most among women at high risk. They also suggest that women with dense breast tissue gain least sensitivity among the four groups of women when CBE is applied in addition to mammography. One previous study found that women who benefit most from including CBE is women with dense breasts (Oestreicher *et al.*, 2005). However, our model shows that the group of women ages 50 to 69 with dense breast tissue is the group that benefits the least from increased tumor ascertainment when CBE is added to a screening program; further, the group experiences the second most harm from CBE due to false-positive results among four groups of women.

In the context of an organized screening program, the potential added contribution of CBE to tumor detection increases with age, with breast cancer becoming more prevalent and mammography becoming more accurate at detecting breast cancers as a woman becomes older. However, our results suggest that the incremental benefit of CBE in sensitivity does not vary much by age, especially for women at low risk or with dense breast tissue, even though sensitivities themselves do increase with age. The detrimental effect in specificity when CBE is included decreases slightly with age.

Recently the Canadian Task Force on Preventive Health Care updated their breast screening guidelines, recommending routine screening for breast cancer with mammography every 2 to 3 years for women ages 50 to 69. Routinely performing a CBE alone or in conjunction with mammography was not recommended, even though this advice was based on low quality evidence. With our results, participants and their physicians may carefully weigh the change of detection rates and the trade-offs between sensitivity and specificity based on the risk status. CBE can be targeted to the groups in which it has the highest net benefit (e.g., women 50 to 69 currently having hormone therapy).

There is considerable variation among examiners' recall probabilities after adjusting for various factors, and this variation is not accounted for by years of experience or variation between screening sites. However, the small number of cancers in the dataset and the consequent lack of information on cancer detection rates for women not being recalled might reduce the precision in true positive rates. False positive rates for examiners are, by contrast, well identified by extensive data on recalls of cancer-free participants. As a result, examiners can be more clearly differentiated by their specificities and positive predictive values than by their sensitivities and negative predictive values. The variation in estimated specificity is substantial, with a number of examiners recalling 20% of the healthy individuals who they see while others recalling only 2% of the healthy individuals. For a population-level screening program examining tens of thousands of healthy individuals annually, these variations are not inconsequential. After adjusting for various factors, we also find that the true positive and false positive rates for mammography referrals are not significantly different among centres with CBE compared to those without CBE.

The study had several strengths. First, it accounts for the effects of provider experience and facility characteristics on the accuracy of an exam. Secondly, by applying advanced statistical models, possible biases created by missing cancers during follow-up and the cancers detected by chance have been taken into account. An earlier study of the same population found that women referred by CBE alone were significantly more likely to have incomplete follow-up information in the OBSP (Chiarelli *et al.*, 2009). This would lead to underestimated accuracy measures if the missing cancers were not taken into account. Lastly, since sensitivity of mammography may be overestimated as CBE results are available to radiologists, the study estimates the effect of CBE on diagnostic accuracy of mammography by directly comparing two screening protocols, with and without CBE.

The study has some limitations. Due to lack of data access, years of experience for the provider or years of operation for the affiliates outside of the OBSP were unavailable for the analysis. Even though our analysis shows that years of experience inside the OBSP has no significant impact on either true positive or false positive rates, more evidence would be needed to draw a conclusion that radiologists' overall working experience is unrelated to diagnostic accuracy. As some of the characteristics of the women, such as family history and hormone therapy use, were based on self report, misclassification might have occurred. Although a previous study (Halapy *et al.*, 2005) reported that the accuracy in first degree relatives or hormone therapy use was relatively high, any misclassification may have been non-differential and would reduce the reliability of our estimates. Effect sizes of risk factors related to false positive probabilities are much better estimated than those related to cancer detection, likely due to the much greater number of women without cancer than those with cancer in the dataset. The lack of significance for the risk factors related to true positive probabilities may be due to low power in our study.

Overall, our study found that CBE may benefit older, currently using hormone therapy and high-risk women when offered in addition to mammography. However, the increase in sensitivity from the addition of CBE to mammography should be compared to the decrease in specificity resulting from false positive screening exams and costs of further follow-up and greater number of diagnostic exams. These risks and benefits may differ by characteristics of the women such as age, current hormone therapy use, breast density and family history. In addition, the accuracy of mammography and CBE significantly varies by examiners and this variation is not accounted for by years of experience of examiners inside OBSP or variation among screening sites. Further investigation is required to determine the factors associated with differences in screening accuracy among examiners.

Chapter 4

Estimation of Screening Sensitivity and Sojourn Time from an Organized Screening Program

Abstract

Regular screening for breast cancer with mammography is widely recommended to reduce mortality due to breast cancer. However, whether breast cancer screening does more harm than good has been debated continuously (Chiarelli *et al.*, 2009; Oestreicher *et al.*, 2005; Barton *et al.*, 1999; US Preventive Services Task Force and others, 2009; Miller *et al.*, 2014). Since a full evaluation of the effect on mortality will take a follow-up of about 10 to 15 years to provide a reliable estimate of the benefits and harms, it is unrealistic to expect each new modification of a screening technique to be evaluated in this way. Therefore, one needs to rapidly estimate the measures of effect. In this chapter, two measures of interest, the length of the pre-clinical state and the false negative rate, are discussed. Two estimation procedures are proposed to model the pre-clinical state duration, the false negative rate of screening exam, and the underlying incidence rate in the screened population. Both methods

assume the sojourn time, time spent in the preclinical detectable phase, follows a negative exponential distribution while we consider two different forms for the false negative rate: 1) being constant with time and 2) an exponential distribution to compensate for the fact that lesions would become easier to detect the closer they are to the clinical stage. We show how to jointly estimate these measures by using data on the observed prevalence of disease at a screen and on the incidence of disease during intervals between screens. We illustrate the proposed methods with breast cancer screening data from the Ontario Breast Screening Program in Canada.

4.1 Introduction

Breast cancer was the most frequently diagnosed cancer in Ontario women in 2012, and ranks second among all cancers only to lung cancer among causes of cancer deaths in women (Ontario Breast Screening Program, 2013). Regular screening for breast cancer with mammography and clinical breast examinations are widely recommended for reducing mortality due to breast cancer. However, whether breast cancer screening does more harm than good has been debated continuously. For population screening programs, the debate has focused on the reduction in mortality attributable to screening, the numbers of women overdiagnosed, and the accuracy of the screening exams (Chiarelli *et al.*, 2009; Oestreicher *et al.*, 2005; Barton *et al.*, 1999; US Preventive Services Task Force and others, 2009; Miller *et al.*, 2014). The arguments have become polarized between those who believe that the benefits of screening outweigh the harms and those who believe the opposite. In 2009, the US Preventive Services Task Force updated their recommendations on breast cancer screening in the general population. They recommended that women younger than 50 years do not need to be screened routinely and women between the ages of 50 and 74 years should have biennial screening mammography (US Preventive Services Task Force and others, 2009). The Canadian Taskforce on Preventative Health Care updated their guidelines in 2011 and found that the reduction in mortality associated with screening mammography is small for women at average risk of breast cancer (Canadian Task Force, 2011). Some reviews on screening for breast cancer with mammography concluded that screening was likely to reduce breast cancer mortality at the expense of 30% overdiagnosis and overtreatment (Gøtzsche and Olsen, 2000; Olsen and Gøtzsche, 2001). The Independent UK Panel on Breast Cancer Screening suggested that breast screening extended lives and concluded that the UK breast screening program, where women aged 50 to 70 years are invited for screening every 3 years, conferred significant benefit and should continue (Independent UK Panel, 2012). One recent Canadian paper (Miller *et al.* (2014)) argued that annual mammography in addition to physical examination in women aged 40 to 59 does not reduce mortality from breast cancer.

The best demonstration of the benefit would be a reduction in cancer-specific mortality in a screened group. It would ideally come from a comparison of the number of cancer-specific deaths in a randomized group of women being screened with that in an unscreened comparable population, which is similar in terms of cancer risk factors and quality of treatment, and followed up until death. However, such randomized trials exist only rarely. In fact, Olsen and Gøtzsche (2001) have assessed seven randomized trials of screening mammography, and have concluded that no trial data were of high quality. In addition to this, most of the randomized controlled trials started at least twenty years ago (Peer *et al.*, 1996; Kerlikowske *et al.*, 1996a; Shapiro, 1977; Shapiro *et al.*, 1988; Uk Trial Of Early Detection and Group, 1988; Tabar *et al.*, 1995; Tabár *et al.*, 2000; Miller *et al.*, 1992). Observational studies might provide more contemporary estimates. Since a full evaluation of the effect on mortality will take a follow-up about 10 to 15 years to provide reliable estimates of the benefits and harms, it is often unrealistic to expect each new modification of a screening technique to be evaluated in this way, especially for population based screening services, where the implementation of the service changes by time and screening sites. Therefore, an early evaluation of the effect is important.

To fully understand the harms and benefits of a screening program, one needs to study the natural history of the disease in question. As represented in Figure 4.1, it may be supposed that breast cancer is initiated by a change in a single cell and then may reach a size that is potentially detectable by screening (T_0). If a woman is not screened, her disease may progress to the phase where it becomes symptomatic and clinically detectable (T_1). If a woman is screened between T_0 and T_1, the disease will possibly be detected in the prevalent phase, but there may be a false negative result. The interval between T_0 and T_1 is called sojourn time and it constitutes the detectable pre-clinical phase of the disease. If a screen happens at time T_2 and the disease is detected, the interval T_1 - T_2 represents the lead time; the interval T_2 - T_0 is known as the delay time.

The purpose of a screening program is to advance the time of diagnosis to the "pre-clinical phase" so that prognosis can be improved by earlier intervention. To evaluate the efficacy of a screening program, there are two important parameters: the false negative rate and the sojourn time. Sojourn time refers to the time interval between the onset of the detectable preclinical phase and the onset of the clinical phase. The sojourn time measures how much earlier the disease might be detected by the screening procedure. Lead time, as a part of sojourn time, is the interval by which diagnosis is actually brought forward; the longer the sojourn time is, the greater the potential for detecting disease in an early phase. Another parameter of a screening programs is the sensitivity or one minus the false negative rate, which is the probability that a screening examination detects disease in the pre-clinical phase. A screen-detected case might represent a newly developed case since the last examination or a false-negative on a previous screening examination. Knowledge of these two parameters facilitates the development of optimal breast cancer screening strategies. However, neither sojourn time nor the false negative rate is directly observable. As a consequence, various investigations have been carried out to model the screening process and estimate those quantities; see e.g Zelen and Feinleib (1969), Day and Walter (1984), Lee and Zelen (1998) , Shen and Zelen (2005), Auvinen *et al.* (2002), Straatman *et al.* (1997), and Shen and Zelen (1999).

Evidence from many breast cancer screening studies have shown that the sensitivity of

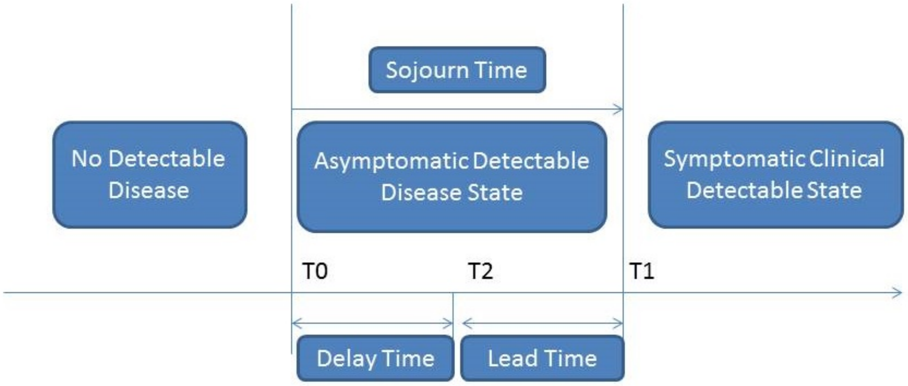

Sojourn Time
No Detectable Disease
Asymptomatic Detectable Disease State
Symptomatic Clinical Detectable State
T0
T2
T1
Delay Time
Lead Time

Figure 4.1: Disease progression with a screening exam (Day and Walter, 1984)

screening exams is higher among women aged 50 years and older compared to those under age 50 (Peer *et al.*, 1996; Kerlikowske *et al.*, 1996a; Miller *et al.*, 1992, 2000). This might relate to the fact that breast tissue of older women is less dense than that of younger women (Kerlikowske *et al.*, 1996a,b). The age at detection has also been found to be related to the length of the pre-clinical phase. Several studies have found that younger women tend to have shorter sojourn times due to rapid tumor growth (Tabar *et al.*, 1995; Tabár *et al.*, 2000; Shen and Zelen, 1999). For example, Shen and Zelen (2001) estimated the mean sojourn time as 1.9 years for the 40-49 age group versus 3.1 years for the 50-59 age group using data from the Canadian National Breast Screening Studies (CNBSS). They concluded that the interval between screenings for the younger age group should be shorter compared to that for the older age group. The differences in screening sensitivity and sojourn time among different age groups raise a challenge for the design of a population screening program. If a screening interval is too long, some of detectable cancers might have advanced into the clinical phase and be missed by screening. However, too short an interval might result in an unnecessary burden on the health care system. Therefore, it is of great interest in epidemiology to estimate the sojourn time and screening sensitivity by age. However, little has been done to model such relationships based on organized service screening data.

In this paper, we revisited and extended the Markov-type model developed by Day and

Walter (1984) and applied the model to a cohort from the Ontario Breast Screening Program. We show how to estimate the sojourn time and the sensitivity of the screen using cohort data on the observed prevalence of breast cancer at screens and on the incidence of disease during intervals between screens. We further investigated the variation of screening sensitivity and the mean sojourn time for different age groups. Lastly, we applied the same method to a clinical trial (CNBSS-I) to test if our methods generate the same estimates as previous studies.

4.2 Methods

4.2.1 Expressions of incidence and prevalence

Day and Walter (1984) proposed a method to estimate test sensitivity and sojourn-time distribution simultaneously, allowing for different distributions of sojourn time. The results indicated that the sojourn time distribution for breast cancer was well approximated by the exponential distribution after examining three distributions: exponential, step function and log-normal. When considering clinical trial data, Zelen and Feinleib (1969) proved that the necessary and sufficient condition for the sojourn time to have an exponential distribution was that the standardized mean age of incidence in the first examination at time t in a control group is the same as the standardized mean age at diagnosis for individuals detected in the pre-clinical phase at time t in the study group. With data from a randomized clinical trial for breast cancer conducted by the Health Insurance Plan of Greater New York, they showed that the estimated standardized mean age at diagnosis in the pre-clinical phase was very close to the mean age of incidence, though this condition cannot be easily verified given an observational study without a “control group”. Therefore, we made a pragmatic decision and assumed that the sojourn time follows the exponential distribution for the purpose of convenience.

Following Day and Walter (1984), we write $J(t)$ as the underlying incidence of the preclinical phase at age t and $f(y)$ as the density function of the length of sojourn time y. The "false negative rate", i.e. the probability that a test returns a negative result when the individual is in the preclinical phase, is written as β. We assume that β changes with screening times, denoted by $\beta_1, \beta_2, \cdots, \beta_I$ at times $t_1, t_2, \cdots, t_n$. The intensities of observed incidence $I(t)$ and of pre-clincial incidence $J(t)$ are related through $f(y)$ by the equation

$$``I(t) = \int_0^t J(s) f(t-s) ds''$$

Supposing that every individual in the population is screened at t_1, the cancer incidence observed afterward consists two parts: individuals entering the preclinical phase before t_1 but having a false negative result at t_1, and individuals entering the preclinical phase after t_1. Thus the incidence $I_1(t)$ after t_1 is written as

$$I_1(t) = \int_0^{t_1} J(s)\beta_1 f(t-s) ds + \int_{t_1}^{t} J(s) f(t-s) ds$$

When screens occur at times $t_1, t_2, \cdots, t_n$, the incidence $I_n(t)$ after the nth screen can be written as

$$\begin{aligned} I_n(t) = & \int_0^{t_1} J(s)\beta_1\beta_2\cdots\beta_n f(t-s) ds + \int_{t_1}^{t_2} J(s)\beta_2\cdots\beta_n f(t-s) ds + \cdots \\ & + \int_{t_{n-1}}^{t_n} J(s)\beta_n f(t-s) ds + \int_{t_n}^{t} J(s) f(t-s) ds \end{aligned}$$

or in shorter format

$$I_n(t) = \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} J(s) \prod_{j=i+1}^{n} \beta_j f(t-s) ds + \int_0^{t-t_n} J(s) f(t-s) ds$$

where $t_0 = 0$.

Assuming that $J(t)$ is uniform for an individual across the whole study period, the above expression can now be written as

$$I_n(t) = J\left(\sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}}\prod_{j=i+1}^{n}\beta_j f(t-s)ds + \int_{t_n}^{t} f(t-s)ds\right) \tag{4.1}$$

Expressions for the prevalence of the preclinical condition can be derived similarly. The prevalence, P_1, observed at the first screen at time t_1, is given by

$$P_1 = J\left(\int_0^{t_1}\int_{t_1-s}^{\infty}(1-\beta_1)f(y)dyds\right) \tag{4.2}$$

If there is one previous negative screen at time t_1, its contribution to the prevalence observed at the second screen at time t_2 made by individuals entering the preclinical phase in the interval $(0, t_1)$ is given by

$$P_2^{(0-t_1)} = J\left(\int_0^{t_1}\int_{t_1-s}^{\infty}\beta_1(1-\beta_2)f(y)dyds\right)$$

The contribution made by individuals entering the preclinical phase in the interval (t_1, t_2) is given by

$$P_2^{(t_1-t_2)} = J\left(\int_{t_1}^{t_2}\int_{t_2-s}^{\infty}(1-\beta_2)f(y)dyds\right)$$

If there are $n-1$ previous negative screens, at times $t_1, \cdots, t_{n-1}$, then the contribution to the prevalence observed at the nth screen at time t_n made by individuals entering the preclinical phase in the interval (t_{i-1}, t_i) is given by

$$J \int_{t_n-t_i}^{t_n-t_{i-1}} \int_{t_n-t_{i-1}-s}^{\infty} \beta_i \cdots \beta_{n-1}(1-\beta_n) f(y) dy ds$$

Summation over the n intervals $(0, t_1), (t_1, t_2), \cdots, (t_{n-1}, t_n)$ gives the following expression for the total prevalence P_n at the nth screen at time t_n:

$$P_n = \sum_{i=1}^{n} J \left(\int_{t_n-t_i}^{t_n-t_{i-1}} \int_{t_n-t_{i-1}-s}^{\infty} \beta_i \cdots \beta_{n-1}(1-\beta_n) ds f(y) dy \right) \tag{4.3}$$

4.2.2 Two models with different forms for the false negative rate

Estimates of the sojourn time $f(y)$ and of the false-negative rate $\beta(t)$ can be obtained by maximum likelihood. We have considered two different forms for $\beta(t)$: a constant for a certain age group, and an exponential distribution to reflect the fact that lesions may become easier to detect the closer in time that they are to being detected clinically.

Parameter Estimation with $\beta(s)$ as constant: Model I

We assume that underlying incidence rate J is uniform for an individual in a certain age group. Given a relatively short follow-up period, this assumption holds approximately for cancer. Similarly we assume β to be independent of sojourn time and uniform within the age group. Since the age of the participants upon entering the study varies between 50 and 69, the following analysis is stratified by three age groups: age 50-55, age 56-64 and age 65+ years.

With these two assumptions, estimation of $f(y)$ and β is relatively straightforward. For a group of women in a given age group, at the ith screen $(i = 1, \ldots, n)$, one observes s_i screen-detected cases and N_i the number screened; between screen i and $i + 1$, a total of c_i cases are diagnosed clinically (not by a screening exam) from a total of y_i person-years at risk. Then the incidence $I_n(t)$ after the nth screen is given by

$$\begin{aligned} I_n(t) &= \sum_{i=0}^{n} J\beta^{n-i}\lambda \int_{t-t_{i+1}}^{t-t_i} e^{-\lambda y} dy \\ &= J\sum_{i=0}^{n} \beta^{n-i}(e^{-\lambda(t-t_{i+1})} - e^{-\lambda(t-t_i)}) \end{aligned}$$

The probability Q_i of a case developing between screen i and screen $i+1$ and being diagnosed clinically is given by:

$$Q_i = 1 - e^{-\int_{t_i}^{t_{i+1}} I_i(t)dt}$$

Since the incidence between two screens $\int_{t_i}^{t_{i+1}} I_i(t)dt$ is much smaller than 1, Q_i can be well approximated by

$$Q_i \approx \int_{t_i}^{t_{i+1}} I_i(t)dt$$

The probability p_n of a case being found by the nth screen is given by

$$P_n = (1-\beta)J\lambda^{-1} \sum_{i=0}^{n} \beta^{j-i}(e^{-\lambda(t-t_{i+1})} - e^{-\lambda(t-t_i)})$$

Parameter estimation with $\beta(s)$ taken as exponential: Model II

We have assumed that the false-negative rate is independent of both the lead time and the sojourn time in model I. However, a lesion might become easier to be detected when it is closer to being detected clinically, that is, the false negative rate is low at the end of the pre-clinical phase and high at the beginning. Therefore, we write the false-negative rate as follows:

$$\beta(s) = e^{-\theta(t-s)}$$

Where θ is the parameter of the exponential function and reflects the rate of change of $\beta(s)$ when $t-s$ changes. At the beginning of the pre-clinical phase (where the value of $t-s$ is small), the false negative rate is relatively high and even close to 1; at the end of the pre-clinical phase, the false negative rate would be very low. The exponential function also suggests that a change in the sojourn time gives the same rate of change in the false negative rate, in other words, the false negative rate decreases faster as sojourn time increases.

After one screen, the incidence $I_1(t)$ is given by

$$\begin{aligned}
I_1(t) &= J\left(\int_0^{t_1} e^{-\theta(t_1-s)} f(t-s)ds + \int_{t_1}^{t} f(t-s)ds\right)\\
&= J\left(\int_0^{t_1} e^{-\theta(t_1-s)} \lambda e^{-\theta(t-s)}ds + \int_{t_1}^{t} \lambda e^{-\lambda(t-s)}ds\right)\\
&= J\left(\frac{\lambda}{\theta+\lambda}[e^{-\lambda(t-t_1)} - e^{-\theta t_1-\lambda t}] + [1-e^{-\lambda(t-t_1)}]\right).
\end{aligned}$$

The incidence (4.1) after the nth screen can be written as

$$\begin{aligned}
I_n(t) = J\Bigg\{&\int_0^{t_1} e^{-\theta(t_1+t_2+\cdots+t_n-ns)} f(t-s)ds + \int_{t_1}^{t_2} e^{-\theta(t_2+t_3+\cdots-(n-1)s)} f(t-s)ds\\
&+\cdots+\int_{t_{n-1}}^{t_n} e^{-\theta(t_n-s)} f(t-s)ds + \int_{t_n}^{t} f(t-s)ds\Bigg\}\\
&= J\Bigg\{\frac{\lambda}{n\theta+\lambda}\lambda e^{-\theta(t_1+\cdots+t_n)-\lambda t}(e^{(n\theta+\lambda)t_1} - e^0)\\
&+\frac{\lambda}{(n-1)\theta+\lambda}\lambda e^{-\theta(t_2+\cdots+t_n)-\lambda t}\left(e^{[(n-1)\theta+\lambda]t_2} - e^{[(n-1)\theta+\lambda]t_1}\right)\\
&+\cdots+\frac{\lambda}{\theta+\lambda}\lambda e^{-\theta t_n-\lambda t}\left(e^{(\theta+\lambda)t_n} - e^{(\theta+\lambda)t_{n-1}}\right) + 1 - e^{-\lambda(t-t_n)}\Bigg\}.
\end{aligned}$$

The prevalence P_1 observed at the first screen at times t_1 is given by

$$\begin{aligned} P_1 &= J\int_0^{t_1}\int_{t_1-s}^{\infty}\left(1-e^{-\lambda(t_1-s)}\right)f(y)dyds \\ &= J\left\{\int_0^{t_1}\left(1-e^{-\theta(t_1-s)}\right)\lambda e^{-\lambda y}dyds\right\} \\ &= \frac{J}{\lambda}(e^{-\lambda t_1}-1)-\frac{J}{\theta+\lambda}(e^{-(\theta+\lambda)t_1}-1). \end{aligned}$$

If there are $n-1$ previous screens at times $t_1,\cdots,t_{n-1}$, the prevalence observed at the nth screen at time t_n made by individuals entering the preclinical phase in the interval (t_{i-1},t_i) is given by

$$\begin{aligned} P_n = J\sum_{i=1}^{n}\Bigg\{ & \frac{1}{(n-1)\theta+\lambda}\left[e^{-\theta(t_1+\cdots+t_{i-1})-\lambda t_n+((n-1)\theta+\lambda)t_i}-e^{-\theta(t_1+\cdots+t_{i-1})-\lambda t_n+((n-1)\theta+\lambda)t_{i-1}}\right] \\ & -\frac{1}{n\theta+\lambda}\left[e^{-\theta(t_1+\cdots+t_{i-1})-(\lambda+\theta)t_n+(n\theta+\lambda)t_i}-e^{-\theta(t_1+\cdots+t_{i-1})-(\lambda+\theta)t_n+(n\theta+\lambda)t_{i-1}}\right]\Bigg\} \end{aligned}$$

4.2.3 Likelihood function and estimation of the preclinical incidence

Recall that for all women in a given group at the ith screen $(i=1,\ldots,n)$, one observes s_i screen-detected cases among N_i the number screened; between screen i and $i+1$, a total of c_i cases are diagnosed clinically. Estimation of the distribution $f(y)$ and of the false-negative rate β can be found by the likelihood function L:

$$L(\beta,\lambda;y,s,c)\propto\prod_{i=1}^{n}P_i^{s_i}Q_i^{c_i}\left(1-P_i^{N_i-s_i}\right)\left(1-Q_i^{y_i-c_i}\right)$$

Since the value of $\left(1-P_i^{N_i-s_i}\right)\left(1-Q_i^{y_i-c_i}\right)$ is very close to 1, the likelihood function can be approximated by $\prod_{i=1}^{n}P_i^{s_i}Q_i^{c_i}$. Then the approximate log-likelihood function can be written as:

$$logL(\beta, \lambda; y, s, c) \propto \sum_{i=1}^{n} \left[s_i log(P_i) + c_i log(Q_i)\right]. \tag{4.4}$$

It is then straightforward to compute the expected number of screen-detected cases E_{is} and interval cases E_{ic} by calculating expected incidence and prevalence. Note that we have set up the range of β as (0, 1) and limit λ to be positive. To calculate the expected numbers of cases E_{is} and E_{ic} given the observed cases s_i and c_i, one needs an numeric value for the preclinical incidence J. However, J is a constant and is unrelated to the optimization process. Instead we obtained the optimal value of J by trying different values of J to minimize the fit statistic. Since the number of cancer cases can reasonably be assumed to follow a Poisson distribution, $(s_i - E_{is})/\sqrt{E_{is}}$ will be approximately normally distributed. From this we see that $(s_i - E_{is})^2/E_{is}$ approximately follows a chi-square distribution with 1 degree of freedom. Noting the independence between and among s_i and c_i, the test statistic 4.5 has a chi-square distribution with $2n$ degrees of freedom.

$$\chi^2_{2n} = \sum_{i=1}^{n} \left[\frac{(s_i - E_{is})^2}{E_{is}} + \frac{(c_i - E_{ic})^2}{E_{ic}}\right]. \tag{4.5}$$

The optimization was performed using the software package R(www.r-project.org/ R Core Team (2014)), and the "optim" function was adopted for the maximization of the likelihood function. The "optim" function is based on a limited-memory quasi-Newton method for bound-constrained optimization (Byrd *et al.*, 1995), and is used to optimize the likelihood and obtain the estimates $\hat{\lambda}$ and $\hat{\beta}$.

4.3 Application

4.3.1 Study population

A cohort of women aged 50-69 years who were first screened through the OBSP between Jan 1, 2003 and Dec 31, 2004, was identified from information routinely collected by an integrated client management system on all women screened in the OBSP, who had been followed up until Dec 31, 2009. Women who participate in the OBSP must be residents of Ontario, have no history of breast cancer or augmentation mammoplasty, and have no acute breast symptoms. The OBSP participation rate is around 29% for women aged 50 to 74 in 2003-2004. From Jan 1, 2003 to Dec 31, 2009, the OBSP provided 402674 screens to 120357 women who were aged 50-69 years, who had their first screen between Jan 1, 2003, and Dec 31, 2004. Among those women, 59472 women attended the program regularly. We have omitted the irregular attenders in the following analysis.

For women diagnosed with a screen-detected or interval breast cancer, pathological confirmation was obtained from regional staff through the recall process and through linkages with the Ontario Cancer Registry (OCR). Case ascertainment in the OCR was estimated to be 98% complete for breast cancer (Holowaty *et al.*, 1995). All of the linkages were accomplished using a computerized probabilistic record linkage system known as Auto Match (Jaro, 1995). Further explanation of this data linkage is given in Appendix A.

We used the data from the first six years of follow-up after the start of screening. We consider the idealized situation where the population is screened at regular intervals. When the screening histories of the individuals in the population follow a less regular pattern, each individual could be treated as a separate sample with size one. However, that creates a heavy computational burden, especially for an observational study with relatively large sample size.

(a) Prevalence

Years	Negative	Women screened	Observed	Prevalence	Expected	Expected
since start	screens		cases	per 1000	model I	model II
0	1	30608.00	214.00	6.99	215.23	255.63
1	2	1055.00	0.00	0.00	2.49	0.63
2	2	22193.00	77.00	3.47	78.26	39.24
3	2	732.00	0.00	0.00	3.24	2.20
3	3	260.00	0.00	0.00	0.68	0.30
4	3	20562.00	62.00	3.02	66.92	36.36
5	3	591.00	0.00	0.00	2.24	1.42
6	4	8845.00	35.00	3.96	37.81	26.63

(b) Incidence

Years since	Previous	Women-months	Observed	Annual	Expected	Expected
start	negative	of followup	Cases	Incidence	mode I	model II
	screens			per 1000		
0-1	1	322881.60	9.00	0.33	12.52	18.14
1-2	1	317304.00	40.00	1.51	22.03	40.10
2-3	1	195811.20	30.00	1.84	18,29	31.25
3-4	1	102884.40	16.00	1.87	11.54	18.03
4-5	1	83268.00	5.00	0.72	10.56	15.22
5-6	1	47190.00	13.00	3.31	6.53	8.79
1-2	2	3276.00	0.00	0.00	0.09	0.18
2-3	2	138610.80	1.00	0.09	6.33	13.55
3-4	2	239802.00	11.00	0.55	14.92	30.30
4-5	2	163260.00	5.00	0.37	14.32	26.05
5-6	2	48362.40	2.00	0.50	4.79	8.17
2-3	3	36.00	0.00	0.00	0.00	0.00
3-4	3	840.00	0.00	0.00	0.02	0.05
4-5	3	93262.80	0.00	0.00	2.69	5.23
5-6	3	151825.20	4.00	0.32	9.36	19.18

Table 4.1: Prevalence and incidence rates (per 1000 person-years) of breast cancer for women aged 50-55 years

4.3.2 Parameter estimates: Model I

We first applied the model to data for women aged 50 to 55 from the OBSP study. The data are summarized in Table 4.1. In this age group, over thirty thousand women had been screened. Figure 4.2(a) shows the joint confidence regions for $1/\lambda$ and β. There is a strong negative correlation between $\hat{\lambda}$ and $\hat{\beta}$. Calculated from the profile likelihood, we estimated $\hat{\lambda}$ and $\hat{\beta}$ to be 0.25 and 0.11, respectively. One can derive the estimated screening sensitivity (1 - β) to be 0.88 (95% CI: (0.69,1)) and estimated mean sojourn time ($1/\lambda$) 3.94 (95% CI:(2.78, 5.5)) years. The numbers of cases expected to be detected at screening are shown in the last two columns of Table 4.1. The overall goodness of fit is $\chi^2_{21} = 65.75$ (P-value < 0.01) with the incidence of the preclinical phase estimated as 2.16 cases per 1000 person-years.

(a) Prevalence

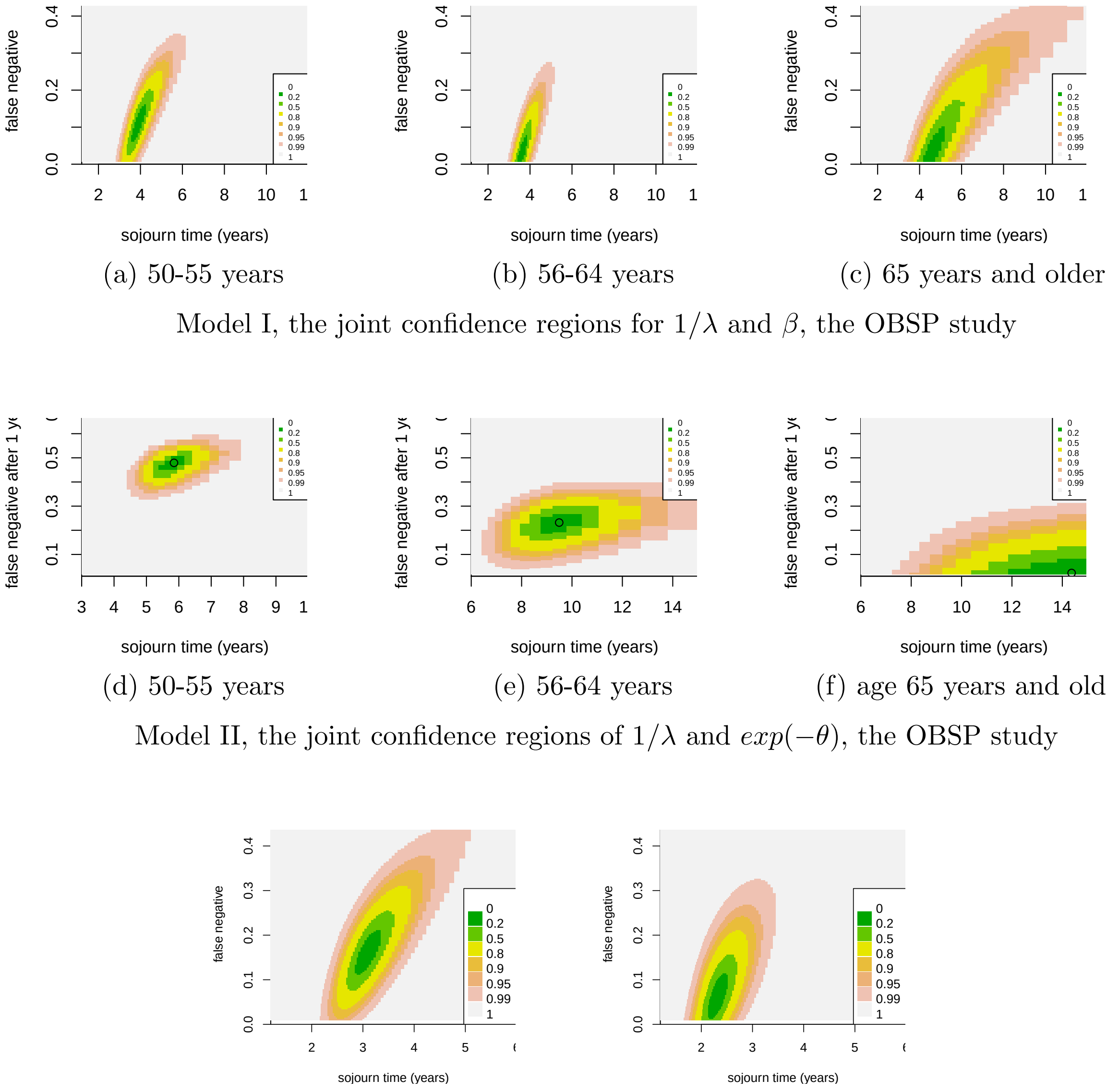
Years since start	Negative screens	Women screened	Observed cases	Prevalence per 1000	Expected modelI 1	Expected model II
0	1	20949.00	227.00	10.84	243.41	268.24
1	2	677.00	0.00	0.00	2.77	0.70
2	2	14705.00	72.00	4.90	83.18	40.66
3	2	424.00	0.00	0.00	2.96	1.90
3	3	163.00	0.00	0.00	0.63	0.31
4	3	13419.00	62.00	4.62	64.67	37.10
5	3	361.00	0.00	0.00	2.04	1.32
6	4	5853.00	26.00	4.44	37.62	26.30

(b) Incidence

Years since start	Previous negative screens	Women-months of followup	Observed cases	Annual Incidence per 1000	Expected model I	Expected model II
0-1	1	225900.00	6.00	0.32	12.24	20.72
1-2	1	219946.80	31.00	1.69	18.21	37.29
2-3	1	131824.80	30.00	2.73	14.12	24.72
3-4	1	66537.60	17.00	3.07	8.50	12.75
4-5	1	55885.20	12.00	2.58	8.12	10.77
5-6	1	33577.20	11.00	3.93	5.38	6.48
1-2	2	2112.00	0.00	0.00	0.07	0.19
2-3	2	96650.40	2.00	0.25	4.86	13.95
3-4	2	164365.20	9.00	0.66	11.09	27.87
4-5	2	108684.00	7.00	0.77	10.22	20.38
5-6	2	29157.60	4.00	1.65	3.13	5.55
3-4	3	492.00	0.00	0.00	0.02	0.45
4-5	3	63548.40	0.00	0.00	2.13	5.83
5-6	3	103969.20	9.00	1.04	6.79	17.63

Table 4.2: Prevalence and incidence rates of breast cancer for women aged 56-64 years

There are over 20000 women in the cohort who had their first screen between the age of 56 to 654 (Table 4.2). Figure 4.2(b) shows the joint confidence regions of $1/\lambda$ and β. The corresponding estimated screening sensitivity is 0.97 (95% CI:(0.78, 1)), and mean sojourn time is 3.70 (95% CI:(3.03, 4.76)) years. The overall goodness of fit is χ^2_{20}=43.32 (P-value $<$ 0.01) with an estimated incidence rate of the population being 3.18 cases per 1000 person-years.



(a) 50-55 years

(b) 56-64 years

(c) 65 years and older

Model I, the joint confidence regions for $1/\lambda$ and β, the OBSP study

(d) 50-55 years

(e) 56-64 years

(f) age 65 years and older

Model II, the joint confidence regions of $1/\lambda$ and $exp(-\theta)$, the OBSP study

(g) CNBSS

(h) HIP

Model I, the joint confidence regions for $1/\lambda$ and β, clinical studies

Figure 4.2: Maximum likelihood estimates and associated confidence regions

(a) Prevalence

Years	Negative	Women screened	Observed	Prevalence	Expected I	Expected II
since start	screens		cases	per1000	model 1	model 2
0	1	7915.00	107.00	13.52	118.17	128.72
1	2	187.00	0.00	0.00	0.65	0.31
2	2	5351.00	36.00	6.73	32.95	20.28
3	2	127.00	0.00	0.00	1.05	0.73
4	3	4576.00	29.00	6.34	27.92	17.34
5	3	109.00	0.00	0.00	0.79	0.52
6	4	1765.00	9.00	5.10	10.77	6.69

(b) Incidence

Years since	Previous	Women-months	Observed	Annual	Expected	Expected
start	negative	of followup	cases	Incidence	modelI	modelII
	screens			per 1000		
0-1	1	86143.20	1.00	0.14	3.09	10.84
1-2	1	82278.00	10.00	1.46	7.62	17.02
2-3	1	48118.80	12.00	2.99	6.64	10.46
3-4	1	25916.40	9.00	4.51	4.17	5.67
4-5	1	22546.80	6.00	3.19	4.58	4.93
5-6	1	14115.60	2.00	1.70	3.19	3.09
1-2	2	612.00	0.00	0.00	0.02	0.08
2-3	2	36175.20	1.00	0.33	2.33	6.71
3-4	2	58930.80	3.00	0.61	5.39	12.19
4-5	2	37321.20	3.00	0.96	5.11	8.11
5-6	2	11080.80	0.00	0.00	1.73	2.42
2-3	3	12.00	0.00	0.00	0.00	0.00
3-4	3	132.00	0.00	0.00	0.00	0.02
4-5	3	23512.80	1.00	0.51	0.81	2.96
5-6	3	35113.20	1.00	0.34	3.21	7.26

Table 4.3: Prevalence and incidence rates of breast cancer for women aged 65+ years

For the group of women starting their first OBSP screen at the age of 65 or older, their screening histories are summarized in Table 4.3. Figure 4.2(c) shows the joint confidence regions of $1/\lambda$ and β with $\hat{\lambda}$ to be 0.22 and $\hat{\beta}$ to be 0.01 respectively. Similar to the other two groups, there is a strong negative correlation between $\hat{\lambda}$ and $\hat{\beta}$. The estimated screening sensitivity is 0.99 (95% CI: (0.64,1)) and mean sojourn time is 4.4 (95% CI,(3.44, 9.09)) years respectively. The wide confidence interval for sensitivity is the result of it being the smallest group among the three groups considered. The overall goodness of fit statistic is χ^2_{20}=48.87 (P-value $<$ 0.01) with the population incidence estimates as 3.36 cases per 1000 persons.

4.3.3 Parameter estimates: Model II

In model II, we assume that the false negative rate follows an exponential decline over time, specifically $\beta = exp(-\theta(t-s))$; that is, the false negative rate is high at the beginning of the pre-clinical phase and decreases when it is close to the clinical phase. The larger the absolute value of θ, the faster the false negative rate drops from its original value. We applied model II to the three age groups. Figure 4.2(d) shows the joint confidence regions of $1/\lambda$ and $exp(-\theta)$ for the group of women aged 55 and younger. The estimated sojourn time $1/\hat{\lambda}$ is 5.88 (98% CI:(4.54, 7.68)) years and the estimated false negative rate after one year in the pre-clinical phase, $exp(-\theta)$, is 0.48 (98% CI:(0.35, 0.57). The overall goodness of fit statistic is χ^2_{21}=148.96 (P-value<0.01) with the preclinical population incidence estimated as 2.27 cases per 1000 person-years. In the last column of Table 4.1, we give the estimated number of incident and prevalent cases calculated from model II.

The estimated number of incident and prevalent cases calculated from model II are shown in the last column of Table 4.2 for women screened for the first time between 56 and 64 years old. Figure 4.2(e) shows the joint confidence regions of $1/\lambda$ and $exp(-\theta)$. The estimated sojourn time $1/\hat{\lambda}$ is 9.98 (95% CI: (6.66, 14.29)) years and the false negative rate after one year in the pre-clinical phase is $exp(-\theta)$ is 0.23 (95% CI: (0.07, 0.37)). The overall goodness of fit statistic is much poorer (χ^2_{20}=148.95 with P-value<0.01). The estimated population incidence rate is 2.32 cases per 1000 persons, which is higher than in the younger age group.

We give the estimated number of incident and prevalent cases for women screened for the first time at the age of 65 or older in the last column of Table 4.2. The joint confidence region (Figure 4.2(f)) for $1/\lambda$ and $exp(-\theta)$ does not show as high a correlation as in the other two groups. The estimated sojourn time is 14.28 (95% CI (8.33, 50.00)) years, which is much longer than the other groups. The estimate of $exp(-\theta)$ is 0.02 (95% CI:(0, 0.28)), which is much lower than the other groups. This implies that the false negative rate decreases quickly and stays low during most of the pre-clinical phase. The overall goodness of fit statistic is χ^2_{20}=69.00 (P-value<0.01) with the population incidence estimate equal to 2.63 cases per

1000 person-years.

From the goodness of fit statistics, the model fitting is poor, especially for the second model. We made reasonable assumptions but they are not being supported by the fit statistics. The reason might be there are not many cells but the sample size is large. Also some cells have zero frequency. Hence, there is a lot of variation but too few degrees of freedom to capture them. For model I, there are two parameters to estimate, while there are three parameters in model II. Therefore, the model fitting of model II is much poorer.

Goodness of fit statistics can be statistically significant with large samples, even though the proportional fit of frequencies is quite good (Tanaka, 1987). We found that model fitting had some discrepancies at the cell level. In Table 4.1, the expected numbers of prevalent cases from model I were very close to the observed numbers of prevalent cases. The observed numbers of prevalent cases after 0, 2, 4, and 6 years are 214, 77, 62, and 35, compared to the expected numbers of cases under the model being 215, 78, 67, 38. The estimates are very close to the observed numbers despite the poor fit statistics. The expected numbers of incident cases were either under or over estimated. In the second year after one negative screen, 40 cases of incident cancer were observed while only 22 cases were expected. In the fifth year, 5 cases were observed after two negative screens while 14 cases were expected. Similarly for model II, some expected frequencies were fairly close to the observed frequencies but others were quite different. For example, the observed number of incident cases in the second year after one negative screen is 40, which is almost the same as the model estimate. However, the observed number of prevalent cases was 214 compared to 255 cases being expected. Although the conclusion is that both models are not well supported by the data, we think our attempt is an encouragement for other researchers to further investigate the modeling of the disease process.

4.3.4 Expected benefits with different screening frequencies

After obtaining the distribution of the pre-clinical phase, we are able to predict the benefits expected with different screening frequencies and patterns. One could compute the expected prevalence and incidence in time, using formulae (4.1) and (4.3), and this would then give the expected proportion of cases during a given time period which would be detected with the proposed screening strategy, and the proportion of cases which would be missed by screens. Assuming an ideal situation where women follow the designated screening frequencies and everyone is observed for exactly the same length of time, one can calculate the expected percentages of screen-detected cancers with different screening frequencies (Day and Walter, 1984).

To simplify the calculation, we suppose that screening starts at age 50 and is carried out at regular intervals for 12 years. Table 4.4 shows the cumulative percentage of breast cancer cases that would be screen-detected over the 12-year period, based on the sojourn time and screen sensitivity from Table 4.5. The marginal gains decrease as the screening frequencies increase. For example, the increase in percentage of screen-detected cancer is 9% when being screened every year instead of once every two years, and it is 7% when being screened every two years compared with every three years. The excess benefit by screening every three years compared with every four years is relatively small. However, a substantial screening benefit is achieved with only one screen in 12 years. This suggests that the most important objective of the screening program should be to encourage first-time attendance for screening.

4.3.5 Application to clinical studies

We believe that the estimation procedures we presented can be used in clinical trials. We are able to obtain data from Health Insurance Plan (HIP) and Canadian National Breast Screening Study (CNBSS) from their publications (Shapiro, 1977; Shapiro *et al.*, 1988; Miller

Inter-screening interval(yrs)	Total screens over 12 years	Cumulative estimated prevalence	Cumulative estimated incidence	Percentage of screen -detected cancer
1	13	29.62	3.88	88
2	7	26.45	6.86	79
3	5	23.98	9.20	72
4	4	21.97	11.09	66
6	3	18.95	13.94	58
12	2	14.07	18.53	43

Table 4.4: Estimated percentage of screen-detected cancer with various frequencies of screening, over 12 years

et al., 2000).

The HIP study was carried out in the 1960s and was the first randomized breast cancer screening trial. Approximately 62,000 women, 40 to 64 years of age at entry, were randomly allocated to a control or screening group (Shapiro, 1977; Shapiro *et al.*, 1988). Women in the study group were offered four annual screening examinations. Two-view mammography and physical examinations were independently carried out on the screened group at each examination. Women in the control group followed their usual practices in obtaining medical care.

The CNBSS study is another randomized controlled trial, conducted in 15 urban centres in Canada (Miller *et al.*, 2000). Around 50000 women who enrolled from January 1980 through March 1986 were followed for an average of 8.5 years. We applied model I to those two studies to obtain the estimated sensitivity, sojourn time and incidence.

The estimation procedures we present can be used in both clinical trials and population-based observational studies. We have applied our first model to two clinical trials: we find that the estimated overall screening sensitivity and mean sojourn time in the HIP study are 0.94 for sensitivity and 2.32 for the mean sojourn time, which is very close to the estimates of Day and Walter (1984). Using the CNBSS data reported by Miller *et al.* (2000), we are able to produce the similar result as reported by Shen and Zelen (2001) using model I. Figure 4.2 (g) and (h) show the likelihood surfaces of these two studies using model I and the estimated

(a) Model I				
	Age Group	Sensitivity (95% CI)	Mean Sojourn Time(95% CI)	Incidence (per 1000 person-years)
OBSP	≤ 55	0.88(0.69,1)	3.94(2.78, 5.5)	2.16
	56-64	0.97(0.78, 1)	3.70(3.03, 4.76)	3.18
	≥ 65	0.99(0.64,1)	4.40(3.44, 9.09)	3.36
HIP	40-64	0.94(0.73,1)	2.32(1.75, 3.22)	2.10
CNBS	40- 49	0.84(0.62,1)	2.70(2.33, 4.55)	3.10
(b) Model II				
	Age Group	θ (95% CI)	λ (95% CI)	Incidence(per 1000 person-years)
OBSP	≤ 55	0.73(0.56, 1.03)	0.17(0.13, 0.22)	2.16
	56-64	1.46(0.99, 2.61)	0.10(0.07, 0.15)	3.18
	≥ 65	3.76(1.27, Inf)	0.07(0.02, 0.12)	3.36

Table 4.5: Parameter Estimates of OBSP, HIP and CNBSS

parameters are summarized in Table 4.5.

We find that sensitivity increases with age when examining the estimates across studies. For example, for women aged 40 - 49 in the CNBS, the estimated sensitivity is 0.84; while it is 0.99 for women aged 65 - 69 in OBSP. The mean sojourn time, on the contrary, appears to decrease with age, being 2.70 years for the group of women aged 40 - 49 and 4.4 years for those aged 65 and above.

4.4 Discussion

It is important to be able to evaluate a screening program without waiting for over 10 years for mortality results. The present work provides a simple way for estimating screening sensitivity and the distribution of preclinical duration for a breast cancer screening program. Recently, several authors (Auvinen *et al.*, 2002; Finne *et al.*, 2010; Pashayan *et al.*, 2009; Epstein *et al.*, 2001) have used the catch-up time method for estimating mean sojourn time under the assumption of independence between the time of clinical incidence and sojourn time. The catch-up method estimates the sojourn time as the time at which cumulative incidence in an unscreened population catches up with the accumulated number of cases in a screened group at baseline. In contrast, our models, like other Markov models, assume independence between sojourn time and the start of the pre-clinical phase. Draisma and Rosmalen (2013) compared the classic Markov model and the catch-up time model, and they found that the catch-up time method leads to different estimates of mean sojourn time when background incidence is increasing with time. In the case of breast cancer in women, sojourn time was known to be shorter and the background incidence is smaller and relatively stable, compared to prostate cancer in men. Therefore, we suggest that the classic Markov model might be more appropriate in this scenario.

Another assumption we made is that the sojourn time is exponentially distributed. Day and Walter (1984) compared the fit of three distributions: exponential, lognormal and a step function. They concluded that the exponential distribution fits the data fairly well and this also gives simpler expressions for the quantities of interest. For model II, we assume that the false-negative rate has the form of a negative exponential function, based on the idea that this false-negative rate will be high at the beginning of pre-clinical phase and drop quickly to nearly zero when one is close to the time of clinical incidence. In future work, it would be worthwhile to investigate how this model fits with other functions.

It is interesting to compare the confidence intervals of the estimates of mean sojourn time and sensitivity for different age groups. With model I, we have the estimated screening sensitivities to be 0.88 (95% CI: (0.69, 1)), 0.97 (95% CI:(0.78, 1)),0.99 (95% CI: (0.64,1)) for the group age 55 and younger, age 56 to 64 and age 65 and older respectively (Table 4.5). At the same time, the estimated mean sojourn times are 3.94 (95% CI:(2.78, 5.5)),3.70 (95% CI:(3.03, 4.76)) and 4.4 (95% CI,(3.44, 9.09)) years for these age groups. It seems that mean sojourn time is longer and the sensitivity is higher in older women. The magnitudes of the sojourn times correspond to approximately 35-40% of cases having a duration less than 2 years for all age groups, and 67-72% of cases having a duration less than 5 years. Recall that both the US and Canadian Taskforce on Preventative Health Care have recommended biennial screening mammography. Our results suggest that the preclinical time is more likely longer than 2 years and hence this provides additional evidence for the recommendations. We also find that the marginal gains by screening every two or three years are relatively small. This suggests that screening every three years could be an alternative screening strategy without losing much benefit but with much lower cost.

The results from model II are not as straightforward to explain as model I, because θ has a negative exponential function and so cannot be translated directly to a false negative rate. Though λ can be explained as the inverse of the sojourn time, its estimates differ from the estimates of the first model and the results from other literature. However, model II shows the same trend that sojourn time increases with age. We also found that the false negative rate decreases much faster for younger women when assuming that the false negative rate decreases exponentially. Even though the model fitting is much poorer compared with the first model, it is an interesting practice not to assume a constant false negative rate being. We hope our attempt will encourage other researchers to further investigate modeling of the disease process.

With electronic medical records and linked health databases becoming increasingly common, large population-based observational datasets will be increasingly available. To the best of my knowledge, the present study is the first to estimate breast cancer screening sensitivity and sojourn time for an observational study. The benefit of an observation study is that the sample size could be very large. One might assume that the more information we collect, the greater the reliability of the estimates. However, we found that the result became much different once we included irregular attendees (results not shown). One explanation could be the existence of strong selection bias. Women might be more likely to have some symptoms when they decide to return to the screening program after missing one or more scheduled exams. In this scenario, the screening test becomes diagnostic and the model assumptions are therefore violated.

Since there are in-situ breast carcinomas and not all cases found at screening would eventually progress to clinical phase, our analyses are limited to the invasive cancers; that is, inferences that we draw on the sojourn time and on sensitivity are based only on cases which become invasive. However, in-situ breast carcinoma does carry a raised risk for developing invasive breast cancer, and it is sometimes considered as pre-invasive disease or "early breast cancer". The exclusion of those cases might affect our estimates. Therefore, it is important to obtain complete follow-up information to properly identify invasive cases.

Another kind of selection bias is the self-selection on the part of those attending a screening program. This kind of bias would not only influence J, the underlying disease incidence, but might even influence the probability of returning for the second and subsequent screens. For example, voluntary attendees seem to be of higher socioeconomic status than those who do not attend screening, and this might result in a lower incidence of breast cancer (Borkhoff *et al.*, 2013; Tabar *et al.*, 1992). People living in high-immigration areas had the lowest screening participation rates (Tabar *et al.*, 1992). The effect of this self-selection bias is not easily evaluated; randomization perhaps provides the only possible way of estimating the magnitude of the bias.

Our model makes the simplifying assumption that only invasive cases would begin to manifest symptoms if no intervention takes place. This might be part of the reason why our estimated sensitivities are higher than in other investigations (Shen and Zelen, 2001). However, a subset of cancers diagnosed as invasive might be overdiagnoses and would never have developed into a clinically manifest tumor. Overdiagnosis in screening is defined as "a histological established diagnosis of invasive or intraductual breast cancer that would never have developed into a clinically manifest tumor during the patient's normal life expectancy if no screening examination had been carried out" (Kalager *et al.*, 2012). The probability of overdiagnosis in breast cancer screening has been discussed for clinical trials (Puliti *et al.*, 2012). Peeters *et al.* (1989) concluded that the rate of overdiagnosis in the Malmo mammographic screening trial was 10% of all screen-detected cancers for women aged 55-69 years. If a substantial proportion of cases detected as invasive in screening are actually not progressive and hence have unlimited sojourn time, the preclinical incidence and sojourn time will be over-estimated. Since there will be more apparent “true positive” cases than the actual situation, the sensitivity will likely be over-estimated as well. A more generalized model which incorporates the overdiagnosis rate might be helpful to better understand the process of the early diagnosis of breast cancer.

Chapter 5

Conclusion

In Chapter 2, we presented a model and methodology for making inferences on screening effectiveness and the factors affecting it, using population-level administrative databases. The model allows for explanatory variables, heterogeneity amongst examiners and unobserved cancers. The probability that an individual has an undetected cancer is calculated using the prevalence and test sensitivities related to their covariate information and health providers, avoiding the assumption that missed cancers occur completely at random with equal probabilities. The true unobserved cancer and detection status of screening participants are treated as latent variables, and we applied Bayesian inference using Markov Chain Monte Carlo (MCMC) and data augmentation to estimate the posterior distributions of the false and true positive rates. We showed how the Bayesian approach can be used to draw inferences about screening exam properties and disease prevalence while allowing for the possibility of conditional dependence between two exams. For a large population-based observational dataset, the individuals in the dataset are often heterogeneous and the outcome variables can be incompletely observed; however, the inclusion of continuous explanatory variables, random effects, and latent variables for disease status can accommodate this complexity when evaluating the performance of screening exams.

Using the model developed in Chapter 2, we examined the benefit of including clinical breast exam (CBE) in addition to mammography in the third chapter. It was a cohort study comprising women 50 to 69 years of age screened at the OBSP between January 1, 2002 and December 31, 2003 and followed for up to 12 months after their last screening examination. After adjusting for various provider and facility characteristics, this study found that although the addition of CBE to mammography improved the sensitivity for women receiving both exams, these women were also subject to increased risk of a false positive test. Our result suggests that CBE generally adds more to sensitivity and reduces more to specificity among women with high risk. Women with dense breast tissue benefit less than women with low risk or women currently using hormone therapy. This finding differs from the belief that CBE generally added incrementally more to sensitivity among women with dense breasts (Oestreicher *et al.*, 2005).

In Chapter 4, two estimation procedures were proposed to model the pre-clinical state duration, the false negative rate of screening exam and the underlying incidence rate in the screened population. We showed how to estimate those measures by using OBSP study data on the observed prevalence of disease at a series of screens and on the incidence of disease during intervals between those screens. We were able to predict the benefits expected with screening frequencies. We found that the marginal gains decrease as the screening frequencies increase. The excess benefit by screening every three years compared with every four years is relatively small. However, a substantial screening benefit is achieved with only one screen in 12 years. This suggests that the most important objective of the screening program should be to encourage first-time attendance for screening.

There are still questions to be answered. For example, we regard all screen detected cancer as having a favorable outcome. In other words, the possibility of over-diagnosis is ignored in the modeling. If a substantial proportion of cases are actually over-diagnosed, all sensitivity, specificity and sojourn time values will be over-estimated. A more rigorous solution would be to explicitly model cancer progression and use information on tumor size to estimate the proportion of cancer that would never have developed into a clinically manifest tumor during one's normal life expectancy.

Another question we have not yet tackled is the survival benefit of mass screening. Even though the purpose of this thesis is to evaluate a screening program without waiting for mortality results, the best demonstration of the benefit is a reduction in cancer-specific mortality in a screened group. To assess the survival benefit associated with early detection, the survival of screen-detected cancer cases needs to be adjusted by lead time. One can then compare the adjusted survival of screen-detected cases to symptom-detected cases. However, one of the challenges might be the identification of deaths from cancers due to the incompleteness or inaccuracy of medical records. It has recently been suggested that all-cause mortality is a more appropriate end point than disease-specific mortality in cancer screening trials, and that disease-specific mortality is biased in favour of screening. However, some researchers argue that breast cancer mortality is the targeted clinical outcome, and so it would be the more appropriate end point in a breast cancer screening trial. All-cause mortality was a poor and inefficient surrogate for breast cancer mortality (Gøtzsche and Olsen, 2000; Olsen and Gøtzsche, 2001; Black *et al.*, 2002; Duffy, 2001; Gøtzsche and Nielsen, 2009; Duffy *et al.*, 2002; Tabar *et al.*, 2002).

Randomized screening trials can only provide limited data to support recommendations on screening policy. To fill the gap, statistical modeling is now being increasingly used with observational studies. The models presented in this thesis allow policymakers, health care providers, and the public to ask and evaluate a range of questions regarding the benefit of a screening program, to project into the future. We hope our work provides insight into cancer natural history and about the evaluation of a screening service.

Appendix A

OBSP data linkage with the Ontario Cancer Registry

This document is provided by Analytics and Informatics, Cancer Care Ontario:

"The purpose of the linkage is to identify all Screen-detected and Post-Screen cancers and cancer stage among OBSP clients. This is done as part of the process of ensuring that the OBSP data are complete before the OBSP publicly reports in aggregate, non-identifiable format, on cancers diagnosed among women screened in the Program. The OBSP can then accurately report on cancer rates. The data are also used to exclude women who have been diagnosed with breast cancer from further screening in the OBSP. This is accomplished by indicating the client has been excluded from further participation due to provincial reasons. The data linkage process is initiated by the OBSP Provincial Office submitting an electronic list of clients to the Ontario Cancer Registry (OCR). The list includes the following OBSP patient variables:

- Client surname, first name, middle name, maiden name and alternate/other name

- Date of birth

- Ontario Health Insurance Number (HIN)

- Postal code, city and county information
- Location of first and last centre where client was screened.

There may be cases where an OBSP client has specifically withheld consent, meaning they have not authorized their primary care provider or other health care providers who perform more tests to release the results to the OBSP. These clients are excluded from the list provided to the OCR. The OCR provides the OBSP with an electronic file that matches those women identified as having been screened in the OBSP who were diagnosed with a cancer either screen-detected or between screening intervals (in other words, they did not have a cancer diagnosed when they were screened in the OBSP, and one was identified before they came back for their next screen or chose not to return to the Program). The following Personal Health Information (PHI) is on the spreadsheet:

- OCR Group ID
- OBSP client ID
- Case
- Diagnosis date
- Topography
- Laterality
- Histology code
- Hospital
- Last contact date
- Data source (whether Regional Cancer Centre (RCC) clinic record, day surgery, pathology or death certificate).

- TMN Stage and individual stage elements (clinical and pathological)

This data is then used by CCO for the Provincial Office to update the Client Management System (ICMS) with cancer stage and other cancer information. Once the updates are complete, the spreadsheet is retained for 1 year (electronically) until the next linkage."

Appendix B

Details of model specification for Chapter 2 Section 2.2

B.1 Odds ratio for actual versus observed true positives

First, note that $pr(D_i = 1|Y_i = 1) = 1$, so

$$\begin{aligned} pr(T_{i1} = 1|Y_i = 1) &= pr(T_{i1} = 1|D_i = 1, Y_i = 1)pr(D_i = 1|Y_i = 1) \\ &= pr(T_{i1} = 1|D_i = 1, Y_i = 1). \end{aligned} \tag{B.1}$$

Also, $pr(Y_i = 1|T_{i1} = 1, D_i = 1) = 1$ and

$$\begin{aligned} pr(T_{i1} = 1, Y_i = 1|D_i = 1) &= pr(Y_i = 1|T_{i1} = 1, D_i = 1)pr(T_{i1} = 1|D_i = 1) \\ &= pr(T_{i1} = 1|D_i = 1). \end{aligned} \tag{B.2}$$

Applying Bayes theorem

$$pr(T_{i1}|Y_i = 1, D_i = 1) = pr(T_{i1}, Y_i = 1|D_i = 1)/pr(Y_i = 1|D_i = 1).$$

yields an odds

$$\begin{aligned}\frac{pr(T_{i1}=1|Y=1)}{pr(T_{1i}=0|Y=1)} &= \frac{pr(T_{i1}=1, Y_i=1|D_i=1)/pr(Y_i=1|D_i=1)}{pr(T_{i1}=0, Y_i=1|D_i=1)/pr(Y_i=1|D_i=1)}\\ &= \frac{pr(T_{i1}=1|D_i=1)}{pr(T_{i1}=0, Y_i=1|D_i=1)}.\end{aligned}$$

Inserting this odds into the odds ratio gives

$$\begin{aligned}\frac{pr(T_{i1}=1|D_i=1)}{pr(T_{i1}=0|D_i=1)}\bigg/\frac{pr(T_{i1}=1|Y_i=1)}{pr(T_{i1}=0|Y_i=1)} =&\frac{pr(T_{i1}=0, Y_i=1|D_i=1)}{pr(T_{i1}=0|D_i=1)}\\ =&pr(Y_i=1|T_{i1}=0, D_i=1).\end{aligned}$$

with the last step resulting from

$$pr(T_{i1}=0, Y_i=1|D_i=1) = pr(Y_i=1|T_{i1}=0, D_i=1)pr(T_{i1}=0|D_i=1).$$

Considering the possible values of T_{i2}, the above formula can be written as

$$\begin{aligned}pr(Y_i=1|T_{i1}=0, D_i=1) =&pr(Y_i=1|T_{i2}=1, T_{i1}=0, D_i=1)pr(T_{i2}=1|T_{i1}=0, D_i=1)+\\ &pr(Y_i=1|T_{i2}=0, T_{i1}=0, D_i=1)pr(T_{i2}=0|T_{i1}=0, D_i=1)\\ =&pr(T_{i2}=1|T_{i1}=0, D_i=1)+\rho pr(T_{i2}=0|T_{i1}=0, D_i=1)\\ =&pr(T_{i2}=1|D_i=1)+\rho pr(T_{i2}=0|D_i=1)\\ =&p_{i2}+\rho(1-p_{i2}).\end{aligned}$$

B.2 Parametrization of true positive rates

As an alternative to modeling the true positive and false positive, consider the following:

- W_{ij} represents an examiner correctly identifying a cancerous lesion on an individual, modeled $W_{ij}|D_i \sim \text{Bernoulli}(r_{ij})$ when $D_i=1$ and $W_{ij}=0$ whenever $D_i=0$; and

- $V_{ij} \sim \text{Bernoulli}(q_{ij})$ represents an examiner declaring a test to be positive for any other reason, likely from concern over a feature which would not be confirmed as cancerous from further medical procedures.

An individual without cancer can only test positive with $V_{ij} = 1$. An individual with cancer can obtain a positive test in one of two ways: the examiner detects the cancer ($W_{ij} = 1$); or the examiner does not detect the cancer ($W_{ij} = 0$) but assigns a positive result because of some other issue ($V_{ij} = 1$). Assuming V_{ij} is independent of W_{ij}, the probability of a positive test of an individual with cancer is

$$\begin{aligned} pr(W_{ij} = 1 \text{ or } V_{ij} = 1) &= 1 - pr(W_{ij} = 0 \text{ and } V_{ij} = 0) \\ &= 1 - (1 - r_{ij})(1 - q_{ij}). \end{aligned}$$

The above is identical to the expression of p_{ij} in (2.3).

As a further motivation for this parametrization, consider an examiner with a tendency to give positive tests to a large proportion of cancer-free individuals with $q_{ij} = 0.9$. This examiner should have a true positive rate of at least 0.9, and a true positive rate of 0.91 would not indicate exceptional cancer-detecting abilities. In contrast, an examiner having $q_{ij} = 0.01$ and having positive tests for 91% of patients with cancer should be regarded as exceptional. Making inference on the covariates influencing r_{ij} as opposed to p_{ij} leads to coefficients being interpretable as “cancer detecting ability” independently of the covariate’s effect on the false positive rate.

B.3 The Algorithm

At each iteration m, the data augmentation step samples true cancer status $D_i^{(m)}$ and possible ‘accidental’ positives $V_{ij}^{(m)}$ (see B.2) conditional on the parameters and random effects.

Unobserved cancers are sampled from the distribution

$$pr(D_i = 1|T_{ij} = 0, Y_i = 0) = \frac{(1-\rho)(1-p_{i.})\psi_i}{(1-\rho)(1-p_{i.})\psi_i + (1-\psi_i)}$$

with $p_{i.} = pr(T_{i1} = 1 \text{ or } T_{i2} = 1|D_i = 1) = 1 - (1 - p_{i1})(1 - p_{i2})$. Accidental positives for screen-detected cancers sampled with

$$pr(V_{ij} = 1|T_{ij} = 1, D_i = 1) = \frac{q_{ij} - \psi p_{ij} q_{ij}}{\psi(1 - (1 - p_{ij})(1 - q_{ij})}.$$

The observation-level fixed effects parameters $\beta^{(m)}$ are block-updated conditional on the other parameters and latent variables using Random Walk Metropolis, with five blocks (β_{1j} and β_{2j} for $j = 1, 2$ and the cancer model β_1). New values are proposed with a multivariate Normal distribution with standard deviations ranging between 0.005 and 0.05, chosen by trial-and-error to give acceptance rates in the 0.5 to 0.8 range. At each iteration 10 updates of this step are performed, as the computational time for this step is relatively undemanding.

Examiner-level random effects $[\theta^{(m)}_{sje}, \eta^{(m)}_{sje}]$ undergo Random Walk Metropolis updating, again repeating 10 times. Proposal standard deviations are 0.8 for the true positive variables and 0.2 for the false positives. Screening site random effects $(\kappa^{(m)}_{sj}, \lambda^{(m)}_{sj})$ and fixed-effects parameters δ are sampled directly from Gaussian conditional distribution given the examiner-level random effects and variance matrices. Variance parameters $\sigma^{(m)}$, $\Sigma^{(m)}_{\cdot}$ and $\Gamma^{(m)}_{\cdot}$ are directly sampled from Wishart conditional distributions given the random effects. The distribution of ρ is Beta-distributed conditional on the other variables in the model and it is sampled directly.

Bibliography

Arends, L., Hamza, T., Van Houwelingen, J., Heijenbrok-Kal, M., Hunink, M., and Stijnen, T. (2008). Bivariate random effects meta-analysis of ROC curves. *Medical Decision Making*, **28**(5), 621–32.

Auvinen, A., Manen, L., Stenman, U.-H., Tammela, T., Rannikko, S., Aro, J., Juusela, H., and Hakama, M. (2002). Lead-time in prostate cancer screening. *Cancer Causes & Control*, **13**(3), 279–85.

Bancej, C., Decker, K., Chiarelli, A., Harrison, M., Turner, D., and Brisson, J. (2003). Original paper: Contribution of clinical breast examination to mammography screening in the early detection of breast cancer. *Journal of Medical Screening*, **10**(1), 16–21.

Banks, E. (2001). Hormone replacement therapy and the sensitivity and specificity of breast cancer screening: a review. *Journal of Medical Screening*, **8**(1), 29–35.

Barlow, W., Chi, C., Carney, P., Taplin, S., D'Orsi, C., Cutter, G., Hendrick, R., and Elmore, J. (2004). Accuracy of screening mammography interpretation by characteristics of radiologists. *Journal of the National Cancer Institute*, **96**(24), 1840–50.

Barton, M., Harris, R., and Fletcher, S. (1999). Does this patient have breast cancer?: The screening clinical breast examination: Should it be done? How? *Journal of the American Medical Association*, **282**(13), 1270–80.

Beam, C., Layde, P., and Sullivan, D. (1996). Variability in the interpretation of screening mammograms by us radiologists: findings from a national sample. *Archives of Internal Medicine*, **156**(2), 209–13.

Black, M. A. and Craig, B. A. (2002). Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine*, **21**(18), 2653–69.

Black, W. C., Haggstrom, D. A., and Welch, H. G. (2002). All-cause mortality in randomized trials of cancer screening. *Journal of the National Cancer Institute*, **94**(3), 167–73.

Bleyer, A. and Welch, H. G. (2012). Effect of three decades of screening mammography on breast cancer incidence. *New England Journal of Medicine*, **367**(21), 1998–05.

Bobo, J., Lee, N., and Thames, S. (2000). Findings from 752 081 clinical breast examinations reported to a national screening program from 1995 through 1998. *Journal of the National Cancer Institute*, **92**(12), 971–76.

Bobo, J. K., Shapiro, J. A., Schulman, J., and Wolters, C. L. (2004). On-schedule mammography rescreening in the national breast and cervical cancer early detection program. *Cancer Epidemiology Biomarkers & Prevention*, **13**(4), 620–30.

Borkhoff, C. M., Saskin, R., Rabeneck, L., Baxter, N. N., Liu, Y., Tinmouth, J., and Paszat, L. F. (2013). Disparities in receipt of screening tests for cancer, diabetes and high cholesterol in Ontario, Canada: A population-based study using area-based methods. *Canadian Journal of Public Health*, **104**(4).

Boyd, N., Guo, H., Martin, L., *et al.* (2007). Mammographic density and risk of breast cancer. *New England Journal of Medicine*, **356**, 227–36.

Brekelmans, C., van Gorp, J., Peeters, P., and Collette, H. (1996). Histopathology and growth rate of interval breast carcinoma: characterization of different subgroups. *Cancer*, **78**(6), 1220–28.

Brown, P. and Zhou, L. (2010). Mcmc for generalized linear mixed models with glmmbugs. *R Journal*, **2**, 13–17.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *Journal on Scientific Computing*, **16**(5), 1190–08.

Canadian Task Force (2011). Recommendations on screening for breast cancer in average-risk women aged 40–74 years. *Canadian Medical Association Journal*, **183**(17), 1991–01.

Carroll, R. (2006). *Measurement Error in Nonlinear Models: a Modern Perspective*. CRC Press, 2 edition.

Chiarelli, A., Mai, V., Moravan, V., Halapy, E., Majpruz, V., and Tatla, R. (2003). False-positive result and reattendance in the Ontario Breast Screening Program. *Journal of Medical Screening*, **10**(3), 129–33.

Chiarelli, A., Halapy, E., Nadalin, V., Shumak, R., O'Malley, F., and Mai, V. (2006). Performance measures from 10 years of breast screening in the Ontario Breast Screening Program, 1990/91 to 2000. *European Journal of Cancer Prevention*, **15**(1), 34–42.

Chiarelli, A., Majpruz, V., Brown, P., Thériault, M., Shumak, R., and Mai, V. (2009). The contribution of clinical breast examination to the accuracy of breast screening. *Journal of the National Cancer Institute*, **101**(18), 1236–43.

Chiarelli, A., Majpruz, V., Brown, P., Theriault, M., Edwards, S., Shumak, R., and Mai, V. (2010). Influence of nurses on compliance with breast screening recommendations in an organized breast screening program. *Cancer Epidemiology Biomarkers & Prevention*, **19**(3), 697–06.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, **49**(4), 327–35.

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

Dendukuri, N. and Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*, **57**(1), 158–67.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2013). *Analysis of Longitudinal Data*. Oxford University Press, 2 edition.

Draisma, G. and Rosmalen, J. (2013). A note on the catch-up time method for estimating lead or sojourn time in prostate cancer screening. *Statistics in Medicine*, **32**(19), 3332–41.

Duffy, S. (2001). Interpretation of the breast screening trials: a commentary on the recent paper by Gøtzsche and Olsen. *The Breast*, **10**(3), 209–12.

Duffy, S., Tabár, L., and Smith, R. A. (2002). The mammographic screening trials: commentary on the recent work by Olsen and Gøtzsche. *Journal of Surgical Oncology*, **81**(4), 159–62.

Epstein, J. I., Walsh, P. C., and Ballantine Carter, H. (2001). Dedifferentiation of prostate cancer grade with time in men followed expectantly for stage T1c disease. *The Journal of Urology*, **166**(5), 1688–91.

Espeland, M. and Handelman, S. (1989). Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics*, **45**(2), 587–599.

Finne, P., Fallah, M., Hakama, M., Ciatto, S., Hugosson, J., de Koning, H., Moss, S., Nelen, V., and Auvinen, A. (2010). Lead-time in the European Randomised Study of Screening for Prostate Cancer. *European Journal of Cancer*, **46**(17), 3102–08.

Gøtzsche, P. C. and Nielsen, M. (2009). Screening for breast cancer with mammography. *Cochrane Database Syst Rev*, **4**(1).

Gøtzsche, P. C. and Olsen, O. (2000). Is screening for breast cancer with mammography justifiable? *The Lancet*, **355**(9198), 129–34.

Green, B. and Taplin, S. (2003). Breast cancer screening controversies. *The Journal of the American Board of Family Practice*, **16**(3), 233–41.

Halapy, E., Chiarelli, A., Klar, N., and Knight, J. (2005). Accuracy of breast screening among women with and without a family history of breast and/or ovarian cancer. *Breast Cancer Research and Treatment*, **90**(3), 299–05.

Holowaty, E., Marrett, L., and Fehringer, G. (1995). Methods cancer incidence in Ontario: trends and regional variations in the 1980s. *Publications Ontario, Toronto, Ontario.*

Hui, S. and Walter, S. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, **36**(1), 167–71.

Humphrey, L. L., Helfand, M., Chan, B. K., and Woolf, S. H. (2002). Breast cancer screening: a summary of the evidence for the us preventive services task force. *Annals of Internal Medicine*, **137**(5_Part_1), 347–60.

Independent UK Panel (2012). The benefits and harms of breast cancer screening: an independent review. *Lancet*, **380**(9855), 1778–86.

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine*, **14**(5-7), 491–98.

Joseph, L., Gyorkos, T., and Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, **141**(3), 263–72.

Kalager, M., Adami, H.-O., Bretthauer, M., and Tamimi, R. M. (2012). Overdiagnosis of invasive breast cancer due to mammography screening: results from the Norwegian Screening Program. *Annals of Internal Medicine*, **156**(7), 491–99.

Kavanagh, A., Mitchell, H., and Giles, G. (2000). Hormone replacement therapy and accuracy of mammographic screening. *The Lancet*, **355**(9200), 270–74.

Kerlikowske, K., Grady, D., Rubin, S., Sandrock, C., and Ernster, V. (1995). Efficacy of screening mammography. *Journal of the American Medical Association*, **273**(2), 149–54.

Kerlikowske, K., Grady, D., Barclay, J., Sickles, E. A., and Ernster, V. (1996a). Effect of age, breast density, and family history on the sensitivity of first screening mammography. *Journal of the American Medical Association*, **276**(1), 33–38.

Kerlikowske, K., Grady, D., Barclay, J., Sickles, E. A., and Ernster, V. (1996b). Likelihood ratios for modern screening mammography: risk of breast cancer based on age and mammographic interpretation. *Journal of the American Medical Association*, **276**(1), 39–43.

Kuchenhoff, H., Mwalili, S., and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification simex. *Biometrics*, **62**(1), 85–96.

Lee, S. J. and Zelen, M. (1998). Scheduling periodic examinations for the early detection of disease: applications to breast cancer. *Journal of the American Statistical Association*, **93**(444), 1271–81.

Lees, A. W. (1996). *Breast Cancer Screening in Canada, Breast Cancer Advances in Biology and Therapeutics*. John Libbey Eurotext.

Litherland, J., Stallard, S., Hole, D., and Cordiner, C. (1999). The effect of hormone replacement therapy on the sensitivity of screening mammograms. *Clinical Radiology*, **54**(5), 285–88.

Macaskill, P. (2004). Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *Journal of Clinical Epidemiology*, **57**(9), 925–32.

Madigan, M., Ziegler, R., Benichou, J., Byrne, C., and Ho1, R. (1995). Proportion of breast cancer cases in the united states explained by well-established risk factors. *Journal of the National Cancer Institute*, **87**(22), 1681–05.

Marrett, L. (1990). *Geographic Distribution of Cancer in Ontario*. Ontario Cancer Treatment and Research Foundation.

McGlothlin, A., Stamey, J., and Seaman Jr, J. (2008). Binary regression with misclassified response and covariate subject to measurement error: A bayesian approach. *Biometrical Journal*, **50**(1), 123–34.

McPherson, K., Steel, C., and Dixon, J. (2000). Breast cancer-epidemiology, risk factors, and genetics. *British Medical Journal*, **321**(7261), 624–28.

Menten, J., Boelaert, M., and Lesaffre, E. (2008). Bayesian latent class models with conditionally dependent diagnostic tests: a case study. *Statistics in Medicine*, **27**(22), 4469–88.

Miller, A., Baines, C., and Turnbull, C. (1991). The role of the nurse-examiner in the National Breast Screening Study. *Canadian journal of public health. Revue canadienne de santé publique*, **82**(3), 162–70.

Miller, A. B., Baines, C. J., To, T., and Wall, C. (1992). Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. *Canadian Medical Association Journal*, **147**(10), 1459–67.

Miller, A. B., To, T., Baines, C. J., Wall, C., *et al.* (2000). Canadian National Breast Screening Study-2: 13-year results of a randomized trial in women aged 50–59 years. *Journal of the National Cancer Institute*, **92**(18), 1490–99.

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. *British Medical Journal*, **348-56**.

Neuhaus, J. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, **86**(4), 843–55.

Oestreicher, N., Lehman, C., Seger, D., Buist, D., and White, E. (2005). The incremental contribution of clinical breast examination to invasive cancer detection in a mammography screening program. *American Journal of Roentgenology*, **184**(2), 428–32.

Olsen, O. and Gøtzsche, P. C. (2001). Cochrane review on screening for breast cancer with mammography. *The Lancet*, **358**(9290), 1340–42.

Ontario Breast Screening Program (2013). Ontario Breast Screening Program 2011 report.

Pashayan, N., Duffy, S. W., Pharoah, P., Greenberg, D., Donovan, J., Martin, R. M., Hamdy, F., and Neal, D. E. (2009). Mean sojourn time, overdiagnosis, and reduction in advanced stage prostate cancer due to screening with PSA: implications of sojourn time on screening. *British Journal of Cancer*, **100**(7), 1198–04.

Peer, P. G., Verbeek, A. L., Straatman, H., Hendriks, J. H., and Holland, R. (1996). Age-specific sensitivities of mammographic screening for breast cancer. *Breast Cancer Research and Treatment*, **38**(2), 153–60.

Peeters, P. H., Verbeek, A., Straatman, H., Holland, R., Hendriks, J., Mravunac, M., Rothengatter, C., Van Dijk-Milatz, A., and Werre, J. (1989). Evaluation of overdiagnosis of breast cancer in screening with mammography: results of the Nijmegen programme. *International Journal of Epidemiology*, **18**(2), 295–99.

Puggioni, G., Gelfand, A., and Elmore, J. (2008). Joint modeling of sensitivity and specificity. *Statistics in Medicine*, **27**(10), 1745–61.

Puliti, D., Duffy, S. W., Miccinesi, G., De Koning, H., Lynge, E., Zappa, M., and Paci, E. (2012). Overdiagnosis in mammographic screening for breast cancer in Europe: a literature review. *Journal of Medical Screening*, **19**(suppl 1), 42–56.

Qu, Y., Tan, M., and Kutner, M. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, **52**(3), 797–10.

R Core Team (2014). *R: A Language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rindskopf, D. and Rindskopf, W. (1986). The value of latent class analysis in medical diagnosis. *Statistics in Medicine*, **5**(1), 21–27.

Rosychuk, R. and Islam, S. (2009). Parameter estimation in a model for misclassified Markov data–a Bayesian approach. *Computational Statistics & Data Analysis*, **53**(11), 3805–16.

Rosychuk, R. and Thompson, M. (2001). A semi-markov model for binary longitudinal responses subject to misclassification. *Canadian Journal of Statistics*, **29**(3), 395–04.

Roy, S., Banerjee, T., and Maiti, T. (2005). Measurement error model for misclassified binary responses. *Statistics in Medicine*, **24**(2), 269–83.

Rutter, C. and Gatsonis, C. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine*, **20**(19), 2865–84.

Schroen, A., Wobbes, T., and van der Sluis, R. (1996). Interval carcinomas of the breast: a group with intermediate outcome. *Journal of Surgical Oncology*, **63**(3), 141–44.

Séradour, B., Estève, J., Heid, P., and Jacquemier, J. (1999). Hormone replacement therapy and screening mammography: analysis of the results in the Bouches du Rhone programme. *Journal of Medical Screening*, **6**(2), 99–02.

Shapiro, S. (1977). Evidence on screening for breast cancer from a randomized trial. *Cancer*, **39**(6), 2772–82.

Shapiro, S., Venet, W., Strax, P., and Venet, L. (1988). *Periodic Screening for Breast Cancer: the Health Insurance Plan Project and its Sequelae, 1963-1986*, volume 3. Johns Hopkins University Press Baltimore.

Shen, Y. and Zelen, M. (1999). Parametric estimation procedures for screening programmes: stable and nonstable disease models for multimodality case finding. *Biometrika*, **86**(3), 503–15.

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**(15), 3490–99.

Shen, Y. and Zelen, M. (2005). Robust modeling in screening studies: estimation of sensitivity and preclinical sojourn time distribution. *Biostatistics*, **6**(4), 604–14.

Smith, R., Saslow, D., Andrews Sawyer, K., Burke, W., Costanza, M., Evans III, W., Foster Jr, R., Hendrick, E., Eyre, H., and Sener, S. (2003). American cancer society guidelines for breast cancer screening: update 2003. *CA: a Cancer Journal for Clinicians*, **53**(3), 141–52.

Straatman, H., Peer, P. G., and Verbeek, A. L. (1997). Estimating lead time and sensitivity in a screening program without estimating the incidence in the screened group. *Biometrics*, **53**(1), 217–29.

Tabar, L., Fagerberg, G., Duffy, S., Day, N., Gad, A., Gröntoft, O., *et al.* (1992). Update of the Swedish two-county program of mammographic screening for breast cancer. *Radiologic Clinics of North America*, **30**(1), 187–10.

Tabar, L., Fagerberg, G., Chen, H.-H., Duffy, S. W., Smart, C. R., Gad, A., and Smith, R. A. (1995). Efficacy of breast cancer screening by age. new results Swedish two-county trial. *Cancer*, **75**(10), 2507–17.

Tabár, L., Vitak, B., Chen, H.-H., Duffy, S. W., Yen, M.-F., Chiang, C.-F., Krusemo, U. B., Tot, T., and Smith, R. A. (2000). The Swedish two-county trial twenty years later: updated mortality results and new insights from long-term follow-up. *Radiologic Clinics of North America*, **38**(4), 625–51.

Tabar, L., Duffy, S., Yen, M., Warwick, J., Vitak, B., Chen, H., and Smith, R. (2002). All-cause mortality among breast cancer patients in a screening trial: support for breast cancer mortality as an end point. *Journal of Medical Screening*, **9**(4), 159–62.

Tanaka, J. S. (1987). " how big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, **58**(1), 134–46.

Uebersax, J. and Grove, W. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine*, **9**(5), 559–72.

Uk Trial Of Early Detection and Group (1988). First results on mortality reduction in the uk trial of early detection of breast cancer. *The Lancet*, **332**(8608), 411–16.

US Preventive Services Task Force and others (2009). Screening for breast cancer: Us preventive services task force recommendation statement. *Annals of Internal Medicine*, **151**(10), 716–26.

Walter, S. and Irwig, L. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology*, **41**(9), 923–37.

Walter, S., Macaskill, P., Lord, S., and Irwig, L. (2012). Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. *Statistics in Medicine*, **31**(11-12), 1129–1138.

Woodard, D., Gelfand, A., Barlow, W., and Elmore, J. (2007). Performance assessment for radiologists interpreting screening mammography. *Statistics in Medicine*, **26**(7), 1532–51.

Zelen, M. and Feinleib, M. (1969). On the theory of screening for chronic diseases. *Biometrika*, **56**(3), 601–14.